SIGNAL PROCESSING: ELEMENTS OF DETECTION AND ESTIMATION THEORY

ARMANDO VANNUCCI
Dipartimento di Ingegneria dell'Informazione Università degli Studi di Parma
Viale delle Scienze 181A, Parma, Italy
E-mail: vannucci@tlc.unipr.it

MARIA GABRIELLA DI BENEDETTO
Dipartimento INFOCOM Università degli Studi di Roma "La Sapienza"
Via Eudossiana 18, Roma, Italy
E-mail: gaby@acts.ing.uniroma1.it

The authors wish here to provide a basic introduction to the signal processing tools and techniques necessary for the purpose of detecting a signal corrupted by random noise. This is the core problem of *Detection Theory*. It is typically encountered, among various other scientifical and technological research areas, in the experimental study of Gravitational Waves (GW), where the signal is often of an electrical nature and is provided by a suitable transducer, which can be referred to, in general terms, as an antenna. Since the signal to be detected is rarely completely known, its shape can vary according to many unknown parameters, such as amplitude, frequency, direction of arrival etc. The evaluation of those parameters is thus essential for the correct determination of the GW to be detected. The theory and techniques for a reliable evaluation are referred to as *Estimation Theory*. The fundamental role of noise, as the main drawback in the process of detection and estimation, makes it necessary to introduce appropriate tools for its understanding and modelling: the theory of *Random Processes* provides the background for this purpose and for the successful exploitation of noise models for the detection and estimation problems.
This tutorial paper is organized in three sections: the first is a brief review of the theory of Stochastic Processes (a synonym for Random Processes), of which the reader is assumed to have already a basic knowledge; the second is an introduction to Detection Theory, discussing different optimization criteria for the detection problem, under binary or multiple hypotheses, and leading to the design method of a *receiver* operating in the common case of *Gaussian Noise*. The third section deals with *Estimation Theory*: a survey of the parameters estimation problem under different optimization criteria is given, based on the results of the previous section; some analytical tools for evaluating the reliability of the derived *estimators* are also introduced.
The prerequisites for a full understanding of the following material are: a good knowledge of the Theory of Probability and its main results and theorems, about which the reader is assumed to have taken university level courses, and a knowledge of the theory and techniques of *Fourier Analysis*, with which the reader is also assumed to be familiar.

1 Review of Stochastic Processes

When a physical quantity, such as an electrical field or a gravitational field, is measured for the purpose of gaining a certain amount of *information* on it, it is implicit that its value is, a priori, not known nor predictable. When the same quantity varies in time, the function describing time variations is called a *signal*: in order for a signal to bear some useful information it is thus necessary that it varies in an unpredictable way or, in other words, that it is a *random signal*. Those random variations may be caused by an appropriate device, as happens in *electrical digital communications* where an electrical field is varied by a *modulator* according to some data to be transmitted, or rather be intrinsic in the physical quantity to be measured, as is the case for *gravitational waves* where the evolution of the astrophysical source is responsible for its generations.

Stochastic Processes are the mathematical entity for modelling this kind of signals; one possible point of view is to think of a stochastic process as a whole family of functions of time, each function describing one among all of the possible evolutions of the physical quantity to be observed, thus each function being a *deterministic signal*. The observation is supposed to extend along the whole infinitely long time axis and thus, once the observation is performed, the shape of the unknown signal is completely determined and the process is said to have taken one of its *sample functions* or *realizations*. As depicted in Fig. 1, it is useful to remind the similarity of this approach with the classical description of a *Random Variable* (RV): prior to the execution of a random experiment all possible outcomes S_i of a RV, the so called *sample space* S, is needed to describe the variable itself, while after the observation of the actual outcome, the variable takes a certain value, or, in other words, a particular *determination*.

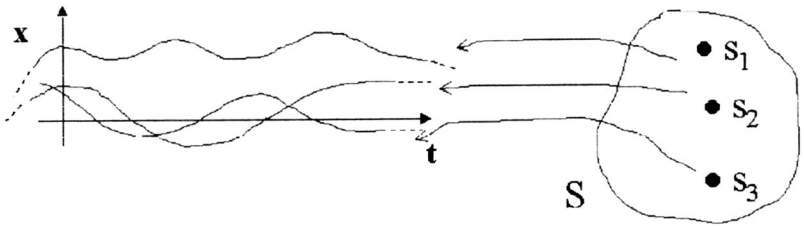

Figure 1. Correspondence between determinations of a random variable and realizations of a stochastic process.

Thus, a stochastic process is a family of signals (its realizations), each describing the same physical quantity (a voltage, a photon flux power ...), where it is not a priori known which signal will experimentally be observed. A process is often denoted by a capital symbol, like $X(t)$, to distinguish it from ordinary functions, or otherwise $X(t; s_i)$ to underline the analogy with RVs, as if the observed signal depended on the outcome s_i of an experiment defined on a probabilistic space S.

Another useful point of view for defining a process is gained by observing its value (i.e. the measured value of the physical quantity that the process represents) at a given time t_1. Prior to the observation, the value $X(t_1; s_i)$ that the process will take depends on which realization will be observed, and is thus a RV X_1 whose outcome depends on s_i. Similarly, by observing the process at an n-tuple of times (t_1, t_2,...,t_n), a random vector X=(X_1,X_2,...,X_n) is obtained, as shown in Fig. 2 for $n = 2$. A process can then be tought of as a collection of a noncountable infinity of random variables X_i, each extracted by observing the process at any possible time t_i.

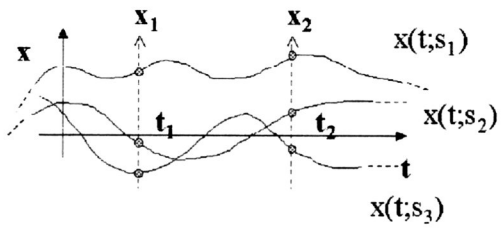

Figure 2. Random variables X_i extracted from the observation of a process; 3 sample functions are shown.

If, on the other hand, in the expression $X(t; s_i)$ it is the second variable s_i that is fixed, the process is reduced to a deterministic signal, i.e. one of its sample functions. If both variables are fixed, $X(t_1; s_i)$ is trivially a number. In the following we will usually omit the dependence on the probabilistic variable s_i, and denote a process by $X(t)$. The probabilistic approach relates any observation $X(t)$ of a process to a RV X: as a random variable, X can be described by the usual probabilistic functions, namely the *Probability Distribution Function* (PDF)

$$F_X(x;t) = P_r\{X(t) \leq x\}$$

where $P_r\{A\}$ denotes the probability of an event A, or equivalently its

first derivative, the *probability density function* (pdf)

$$f_X(x;t) = \frac{\partial F_X(x;t)}{\partial x}$$

These functions depend on two variables: the real number x and the observation time t, thus a knowledge of the pdf, or the PDF, for every possible t would be desirable; all these functions, for every t, constitute the first order statistics of the process. Despite the difficulty of practically obtaining all the first order statistics, they would not even be sufficient for a complete probabilistic description of the process $X(t)$: higher order statistics, such as those of second order

$$F_X(x_1, x_2; t_1, t_2) = P_r\{X(t_1) \leq x_1 \cap X(t_2) \leq x_2\}$$

$$f_X(x_1, x_2; t_1, t_2) = \frac{\partial^2 F_X(x_1, x_2; t_1, t_2)}{\partial x_1 \partial x_2}$$

or, in general, n-th order, $f_X(x_1, x_2, ..., x_n; t_1, t_2, ..., t_n)$, for n arbitrary large, would be needed. Given the statistical correlation of the RVs X_i (their outcomes all depend on the same realization $X(t; s_i)$ of the process), n-th order statistics cannot be derived from lower order statistics.

1.1 Ensemble averages

Usually, a complete statistical description of a process is not needed: it suffices to know some cumulative values yielding information about the *realizations in the average*. The process mean value, defined as

$$\eta_X t = E[X(t)] = \int_{-\infty}^{+\infty} x f_X(x;t) dx$$

where $E[\,]$ is an ensemble average, i.e. the value of the process X at time t is averaged over all possible realizations, is a deterministic function of time giving the shape of the mean realization. Even though $\eta_X(t)$ could not be one of the possible realizations of the process, it gives the a priori expected value of X at time t. Another fundamental function is the process **autocorrelation**, defined as

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

where, for a couple of observation times (t_1, t_2), R_X measures the correlation of the RVs $X_1 = X(t_1)$ and $X_2 = X(t_2)$.

The mean value can be computed if the first order statistics are known; similarly, the autocorrelation requires the second order statistics for its analytical computation: thus mean value and autocorrelation are first and second order ensemble averages, respectively. But, as happens for the average values of random variables, if these ensemble averages are, for some reason, known, it is anyway not possible to derive the first and second order statistics. Another useful second order ensemble average is the **autocovariance**

$$\phi_X(t_1,t_2) = E[(X(t_1)-\eta_X(t_1))(X(t_2)-\eta_X(t_2))] = R_X(t_1,t_2)-\eta_X(t_1)\eta_X(t_2)$$

that can be computed based on the previously introduced mean value and autocorrelation. As for the covariance in the theory of random variables, ϕ_X measure the statistical dependence between the two RVs $X(t_1)$ and $X(t_2)$. The last important average value for a process is the **mean square value**

$$P_X(t) = E\left[|X(t)|^2\right] = R_X(t,t)$$

measuring the average square magnitude of the process at time t. For the computation of all the four mentioned averages, only first and second order statistics would be needed.

The Harmonic process

As an example we shall consider a process $X(t)$ whose realizations are sine waves of fixed amplitude and frequency, having a random initial phase θ. When θ is uniformly distributed over the interval $[0,2\pi]$, the process is called *harmonic*: its general expression is

$$X(t) = a\cos(2\pi f_0 t + \theta)$$

with RV θ uniform in $[0,2\pi]$, and two examples of its realizations are shown in Fig. 3 together with the pdf of θ.

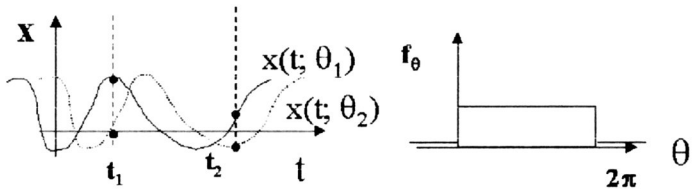

Figure 3. Sample functions of an harmonic process, observed at times t_1 and t_2.

The harmonic process thus represents a sine wave starting at a completely random time and can be practically thought of as the output of a stable oscillator that is turned on at an unknown time. It is one of a wide class of processes, called *parametric processes*, whose peculiarity is the explicit analytic dependence of the process expression on one or more random parameters.

We shall now proceed with the calculation of the previously introduced ensemble averages: the mean value is, by the theorem of expectation, equal to zero

$$\eta_X t = E[X(t)] = \int_0^{2\pi} \frac{1}{2\pi} a\cos(2\pi f_0 t + \theta)d\theta = 0$$

while the autocorrelation, through some trigonometric simplifications, takes the following expression

$$R_X(t_1,t_2) = E[X(t_1)X(t_2)] = \int_0^{2\pi} \frac{1}{2\pi} a\cos(2\pi f_0 t_1 + \theta)\cos(2\pi f_0 t_2 + \theta)d\theta =$$
$$= \frac{a^2}{2}\cos(2\pi f_0(t_1 - t_2))$$

that only depends on the difference between the chosen observation times. The process autocovariance and mean square value are:

$$\phi_X(t_1,t_2) = R_X(t_1,t_2) \qquad P_X(t) = R_X(t,t) = \frac{a^2}{2}$$

1.2 Stationarity

We have just shown that, for the harmonic process, the mean value is independent of the observation time t (it is constant and equal to zero) and the autocorrelation only depends on $(t_1 - t_2)$. These particular properties shown for the ensemble averages of this process are common to a wider, and extremely important, class of processes: any process with the two mentioned properties is called **Wide Sense Stationary (WSS)**. For these, the mean value η_X is thus constant and the autocorrelation $R_X(\tau)$ is expressed with only one time variable $\tau = t_1 - t_2$.

Another definition is that of **Strict Sense Stationary (SSS)** processes: a process is said to be SSS if its statistical properties are invariant to a shift of the time origin. In other words, it can be said that the sample functions of a SSS process have the same probability of being observed with any time displacement. Therefore, given $X(t)$, its statistics are the same of the process $X(t-t_0)$, whose realizations are a time displaced version of those of $X(t)$, and are equally likely to happen. Having the same statistics implies, for the first order statistics, $f_X(x;t) = f_X(x;t-t_0)$, for any t_0; thus this first order pdf must be independent of the observation time and only depends on the variable

x. Resorting to the definition of mean value, it is easy to show that η_X must be constant and independent of time. A similar reasoning can be applied to the second order statistics too: if the process is SSS, $f_X(x_1, x_2; t_1, t_2) = f_X(x_1, x_2; t_1 - t_0, t_2 - t_0)$ must hold, and then these second order pdf must only depend on the difference of the two time arguments. Through the definition of autocorrelation, also $R_X(\tau)$ will depend only on the difference τ between the observation times.

For any SSS process it is then demonstrated that it obeys the two properties defining the class of WSS processes and thus that the first class includes the second as a subset. SSS is indeed a much more strict definition since it is based on the complete statistics of the process, while WSS only requires some average values to respect some properties. It is then evident that the implication SSS⇒WSS cannot be reversed; anyway it must be noted that the definition of WSS is much easier to verify in practice, as has been done for the harmonic process, than that of SSS. From now on we will concentrate mainly on the important class of WSS processes, showing the meaning and properties of their autocorrelation.

The following equality

$$R_X(-\tau) = E[X(t_1)X(t_1 + \tau)] = E[X(t_2 - \tau)X(t_2)] = R_X(\tau)$$

shows that R_X must be an even function of τ; moreover, since $R_X(0) = E[X^2(t)]$, the autocorrelation must have a positive value on the origin of the τ axis. A third property, that we will not demonstrate, states that the maximum absolute value for R_X must be reached on the origin, analytically $|R_X(\tau)| \leq R_X(0)$. Fig. 4 shows pictorially the implications of these properties on the graphic of the autocorrelation for WSS processes: the second function shown is not allowed since it is not even, the third is even but has a negative value on the origin and the fourth respects the first two properties but has its absolute maximum on the two sidelobes rather than on the origin.

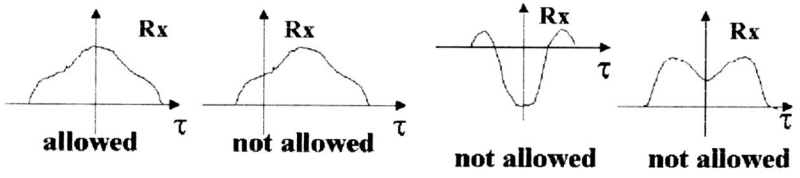

Figure 4. Some graphical properties for the autocorrelation of WSS processes.

As previously outlined, $R_X(\tau)$ is the correlation between the two RVs

$X(t)$ and $X(t + \tau)$ and correlation is, for RVs, a measure of their statistical dependence: if, for instance, RVs $X(t)$ and $X(t + \tau)$ are independent for a certain value of τ, their correlation equals the product of their mean values and thus $R_X(\tau) = \eta_X(t)\eta_X(t+\tau)$, while the autocovariance $\phi_X(\tau)$will be equal to zero. For WSS processes, the value of autocorrelation and autocovariance is the same for any time displacement of the observation times, provided their difference is fixed; then, if the autocovariance has a finite duration on the τ axis, being zero outside a finite interval [-M,M], M can be regarded as the *memory of the process*. The reason for this is that any two observations of the process at times differing by more than M yield two uncorrelated random variables.

Another useful interpretation for the autocorrelation of WSS processes is gained by defining their *power*. Thinking of the process as a random signal conveying energy, if, for instance, $X(t)$ is a voltage or current signal, it can be applied to a resistive load of 1Ω. This way, the instantaneous power dissipated on the load at time t is $|X(t)|^2$, which is a RV whose expectation equals the WSS process mean square value P_X, independent of t. For a single realization, the time average of the dissipated power is

$$P_X(s_i) = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} |X(t; s_i)|^2 dt$$

which is still a random variable, as underlined by the dependence on s_i. The ensemble average over all realizations of this RV is defined as the **power of the process**. If the process is WSS, this average is constant and equals the previously mentioned average of istantaneous powers P_X:

$$P_X = E[P_X(s_i)] = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} E[|X(t; s_i)|^2] dt$$

the expectation in the last integral being constant and equal to $R_X(0)$. Thus the power of a WSS process is constant and equal to the value of its autocorrelation for $\tau = 0$. We will further exploit the relationship between a process power and its autocorrelation, after introducing the topic of process filtering.

1.3 Filtering

A filter is defined as a *Linear Time-Invariant* (LTI) system; as usual in systems theory, such systems are characterized by their response to a *Dirac delta* pulse $\delta(t)$ applied at their input: the output is referred to as the impulse response $h(t)$ of the system. The general relationship between the input and

the output of a filter is given by the *convolution* of the input signal with the system's impulse response, mathematically defined as

$$y(t) = \int_{-\infty}^{+\infty} x(\tau)h(t-\tau)d\tau = x(t) * h(t)$$

where $x(t)$ is the input signal and $y(t)$ is the output. Signal filtering can as well be performed in the frequency domain, resorting to the properties of *Fourier Transforms*: by the *theorem of convolution*, the output signal's transform is simply given by multiplying the transform of the input signal with the system *transfer function* $H(f)$, which is defined as the Fourier transform of the impulse response, according to the simple equation $Y(f) = X(f)H(f)$.

When the input to a LTI system is a stochastic process instead of a deterministic signal, the output is also a stochastic process whose realizations are the filtered versions of the realizations of the input process, being the system deterministic. It is then useful to compute some ensemble averages for the output process, once this averages are given for the input, in order to explicit how the filter acts on the process as a whole. If $x(t)$ is one possible realization of the input process, then performing convolution with $h(t)$ and averaging with respect to all the realizations gives mean value of the output process.

$$\eta_Y(t) = E\left[\int_{-\infty}^{+\infty} x(\tau)h(t-\tau)d\tau\right] = \int_{-\infty}^{+\infty} \eta_x(\tau)h(t-\tau)d\tau = \eta_x(t) * h(t)$$

Computing the autocorrelation of the output process, similarly yields

$$\begin{aligned} R_y(t_1,t_2) &= E\left[\int_{-\infty}^{+\infty} x(t_1-\tau)h(\tau)d\tau \int_{-\infty}^{+\infty} x(t_2-u)h(u)du\right] \\ &= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} R_x(t_1-\tau, t_2-u)h(\tau)h(u)dud\tau \end{aligned}$$

The two above equations can be further simplified when the input process $X(t)$ is WSS

$$\eta_Y = \eta_X \int_{-\infty}^{+\infty} h(\tau)d\tau \qquad R_Y(\tau) = R_X(\tau) * h(\tau) * h(-\tau)$$

which shows that also the output process obeys the two properties defining wide sense stationarity. It can then be stated that linear filtering preserves WSS; it could be also demonstrated that also strict sense stationarity is preserved. This is a nontrivial result, that does not hold for other kinds of systems, such as nonlinear memoryless which preserve the WSS but not the SSS properties.

1.4 *Power Spectral Density*

The last equation, relating the autocorrelations of the input and output processes of a filter, is more easily expressed in the frequency domain, where the double convolution is tranformed into a product: $F[R_Y(\tau)] = F[R_X(\tau)]H(f)H(-f)$, where $F[.]$ denotes Fourier transform and $H(-f) = H^*(f)$ for real systems, i.e. for systems having a real impulse response. It is not clear, however, at this stage what such transform could possibly mean. If, however, we introduce the symbol $S_X(f) = F[R_X(\tau)]$, and recalling the properties of Fourier transforms, the mean power of the WSS process $X(t)$ can be obtained through the integration of $S_X(f)$. Moreover, suppose to apply the process $X(t)$ at the input of an ideal passband filter: the situation is depicted in fig.5. Since the filter is real, a multiplication of $S_X(f)$ by the square magnitude of the transfer function $H(f)$ gives $S_Y(f)$

$$S_Y(f) = S_X(f)|H(f)|^2$$

and since $H(f)$ is equal to one in its passband and zero elsewhere, $S_Y(f)$ is nothing but a *slice* of the function $S_X(f)$.

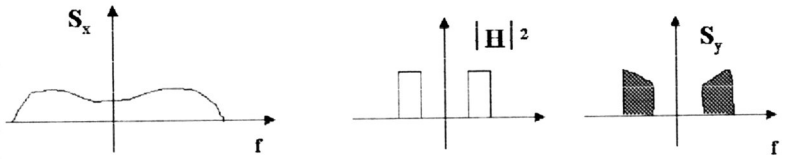

Figure 5. Power spectral density of an ideally bandpass filtered process.

We know that the mean power of the output process is

$$P_Y = R_Y(0) = \int_{-\infty}^{+\infty} S_Y(f)df \geq 0$$

which is the area of the shaded region in Fig. 5. Supposing the filter $H(f)$ has a narrow band centered around the frequency f_0, it suppresses almost all the frequency components of the input process' realizations but those in the neighbourhood of f_0: then it can be said that the mean power of the output process coincides with the power of the input process pertaining those frequencies only. Since, for an extremely narrowband filter, this power is proportional to $S_X(f_0)$, the conclusion is that the introduced function $S_X(f)$ is a description of how the mean power of the input process is distributed on the frequency axis. This conclusion is the celebrated **Wiener-Kintchine**

theorem: $S_X(f)$ is called the **Power Spectral Density (PSD)** of the WSS process $X(t)$ and, from the above reasoning, it is implied that it must be a nonnegative function whose integral on the whole f axis yields the mean power P_X. A similar relationship holds for the so called **covariance spectrum** $\Phi_Y(f)$, which is defined as the Fourier transform of the autocovariance of the process $Y(t)$

$$\Phi_Y(f) = F[\phi_Y(\tau)] = \Phi_X(f)|H(f)|^2$$

Since we stated that the duration of the autocovariance is related to the correlation time, or memory, of the process, processes with short memory, for which their value at time t has little or no statistical dependence on their value at a later time $t + \tau$, with properly large τ, tend to have a covariance spectrum with a wide band.

1.5 Gaussian Processes

Often, the value of a process at a certain time is determined by the superposition of many independent random phoenomena. Since the celebrated *Central Limit Theorem* states that the sum of n independent RVs with finite mean and variance approaches a Gaussian distribution for increasing n, it is intuitively understood how the observation of a process can be a Gaussian RV.

Strictly speaking, a process is called Gaussian if, for any n-tuple of times $\underline{t} = [t_1, t_2, ..., t_n]$, its observed values form a vector $\underline{x} = [x_1, x_2, ..., x_n]$ of jointly Gaussian RVs with joint pdf

$$f_X(\underline{x}; \underline{t}) = \frac{1}{\sqrt{(2\pi)^n |C|}} e^{-\frac{1}{2}[(\underline{x}-\underline{\eta})^T C^{-1}(\underline{x}-\underline{\eta})]}$$

where $\eta_i = E[x_i] = E[X(t_i)] = \eta_X(t_i)$ are the elements of the *mean value vector* $\underline{\eta}$ and the *covariance matrix* C has elements

$$C_{ij} = Cov[x_i, x_j] = E[(X(t_i) - \eta_X(t_i))(X(t_j) - \eta_X(t_j)] = \phi_X(t_i; t_j)$$

Since these elements only depend on the mean value and autocovariance functions of the process, statistics of any order can be computed once $\eta_X(t)$ and $\phi_X(t_1, t_2)$ are given. Moreover, if the Gaussian process is WSS, it can be shown, using η_X and $\phi_X(\tau)$, that the statistics of any order are invariant to a shift of the time origin and thus $X(t)$ is also SSS.

These properties make Gaussian processes a very special kind of processes; another fundamental property is that Gaussianity is preserved by linear filtering: writing convolution in a limit form

$$y(t) = \lim_{\Delta t \to \infty} \left[\sum_{i=-\infty}^{+\infty} x(i\Delta t) h(t - i\Delta t) \Delta t \right]$$

the output process at any time t is shown to be a linear combination of the Gaussian RVs $x(i\Delta t)$ and thus Gaussian. If $X(t)$ is Gaussian and WSS then $Y(t)$ is WSS, as stated in the last section, and also gaussian, then both are SSS.

White Gaussian noise

Noise is often the sum of many independent random effects: for example, the thermal agitation of electrons in a conductor at absolute temperature T produces a random voltage signal at the terminals of the conductor; this stochastic process is referred to as **Johnson noise** or **thermal noise.**

Given a complex impedance Z(f), the celebrated **Nyquist theorem** states that thermal noise, i.e. the voltage N(t) measured at the terminals of the impedance, is a stationary Gaussian process N(t) with mean value $\eta_N = 0$ and covariance spectrum $\Phi_n(f) = 2KT\text{Re}\{\text{Z(f)}\}$, where Re{·} stands for the real part and K is the Boltzmann constant.

If Z(f)=R is a pure resistance, then the covariance spectrum and the PSD are constant and the process N(t) is called **white noise** in analogy with pure white light, which contains all frequency components in the same amount. It is evident that white noise is only an abstraction, otherwise its mean power P_N computed integrating the PSD would be infinite, but for the frequencies of interest (about up to 10^{12} Hz) thermal noise can usually be thought of as white; for higher frequencies the actual spectrum will be slowly decreasing. Also, it is common to deal with filtered versions of thermal noise: since filtering alters the PSD, this is referred to as *coloured noise*.

Inverse-transforming the covariance spectrum of white noise yields for the autocorrelation $\phi_N(\tau) = 2KTR\delta(\tau)$, a Dirac delta with zero duration. Since we interpreted this duration as the memory of the process, we can state that a white noise process has the most rapid, istantaneous, random variations since its value at time t is statistically independent of its value at immediately preceding or subsequent times.

We will often use the properties of gaussian processes and white noise in the following chapters, where noise plays a major role in the detection and estimation problems.

2 Elements of Detection Theory

The problem of detecting an information bearing signal arises in many scientifical and technological fields: communication theory, radar theory and experimental study of gravitational waves, just to name a few. The main difficulty is that the information bearing signal may suffer from many impairments while it transits from the *source* that generated it to the destination equipment that is able to detect it: *weakening*, due to dispersion of its energy, *distortion*, due to physical means that supports its transmission, and *noise*, which is always present in any measuring equipment, are the main impairments. The solution to the problem is to design a suitable **receiver** that is able to detect signals which are weakened, distorted and corrupted by random noise. The object of detection theory is to devise the design methodology of such receivers and the assessment of their performance.

There may be different schemes, depending on the application area, in which a detection problem is at hand: in the GW case, for instance, the main question is to decide, at a certain time, if a useful GW signal in the midst of noise has been detected by the measuring equipment or else the measured data consists of noise alone; we talk of **binary hypotheses testing,** in this case. Another scheme is that of *digital communications*, in which it is costumary to transmit one among many, say M, different signals which is in a one-to-one correspondence with a certain *symbol* to be received at destination; we talk of **multiple hypotheses testing**, in this case.

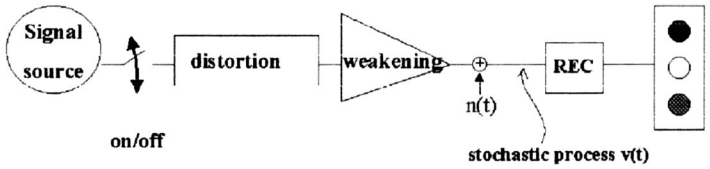

Figure 6. Binary hypotheses testing scheme.

In Fig. 6 a binary hypotheses testing scheme is depicted: the receiver must make a decision whether the source is on or off, and thus turn the traffic light green or red, based on the observation of its input signal $v(t)$. This signal is typically the voltage across the terminals of an antenna and, since the receiver ignores what signal will show up at its input, $v(t)$ is a stochastic process, its randomness being influenced by several factors: in particular, the state of the source (on/off) and the insertion of random noise $n(t)$. The aim of the receiver is, of course, to make mistakes as seldom as possible: this is often a though task since, particularly when dealing with very weak signals, a useful signal plus a strong noise can resemble the shape of noise alone.

2.1 The mathematical approach

Theory of Probability represents a useful mathematical tool for modelling a detection problem: in the general case of multiple hypotheses testing, we can associate the M hypotheses, each representing the detection of a certain signal, to M different *events* H_j. Suppose the receiver observes the signal $v(t)$ in an interval of length T, measuring n data points that are collected in a vector $v = (v_1, v_2, ..., v_n)$: v is a vector of real *samples* extracted from $v(t)$, that lies in the R^n space, R being the set of real numbers. We will discuss later of how the sampling procedure shall be performed; for the moment, the problem can be stated as: choose j, and thus the hypothesis M_j, based on the outcome of the random vector $\underline{v} \in \mathrm{R}^n$. The key to the solution of this problem is the knowledge of the joint conditional pdfs $p_j(v_1, v_2, ..., v_n)$ (we will denote a pdf with the letter p in this chapter), each subscript j denoting conditioning on the event H_j. The problem can thus be restated as: choose which of the M different $p_j(...)$ is believed to characterize the observed data. The receiver must then univocally associate an H_j to any point of the data space, thus dividing R^n in M disjoint regions R_j: any such decomposition is called a **strategy**. Given the strategy, it is determined the rule according to which the receiver decides for a certain hypothesis, having observed the measured data.

2.2 Decision criteria: the Bayes's strategy

To decide which strategy to employ we must seek a decision criterion: it is reasonable to choose that strategy maximizing the probability of correct decision P_c, or equivalently the one minimizing the probability of error $P_e = 1 - P_c$. Assume we know the **prior probabilities** $\Pr(H_j)$ with which the j-th hypothesis is verified, since the receiver chooses H_j whenever $\underline{v} \in \mathrm{R}_j$, we can write, according to the *total probability theorem*

$$P_c = \sum_{j=1}^{M} P_r(H_j) P_r(\underline{v} \in \mathrm{R}_j / H_j) = \sum_{j=1}^{M} P_r(H_j) \int_{R_j} p_j(\underline{v}) d^n v$$

and also the **posterior probabilty** that H_j is true when $\underline{v}$ is observed can be expressed through the Bayes's theorem

$$P_r(H_j/\underline{v}) = \frac{P_r(H_j) p_j(\underline{v})}{p(\underline{v})}$$

with $p(\underline{v}) = \sum_{j=1}^{M} P_r(H_j)p_j(\underline{v})$ the unconditional pdf of the data vector. We can now insert the last equation in the expression of P_c, obtaining

$$P_c = \sum_{j=1}^{M} \int_{R_j} p(\underline{v}) P_r(H_j/\underline{v}) d^n v \qquad (p(\underline{v}) \geq 0)$$

Since we wish to maximize P_c, the best strategy, according to the last expression, is to choose the hypothesis with the greatest posterior probability, given the observed data $\underline{v}$. This conclusion was first enunciated by reverend Thomas Bayes in 1763 in his "*An essay on the doctrine of chances*" and is referred to as the **Bayes's rule.**

In the simple case of binary hypotheses, for the detection problem, the receiver can only choose between the presence of noise only in the observed data, which is referred to as the **null hypothesis** and labelled as H_0, and the presence of a useful signal besides noise, labelled H_1 and called the **alternative hypothesis.** Following Bayes's rule, the receiver must evaluate and compare the posterior probabilities: neglecting the unconditional pdf $p(\underline{v})$, which is irrelevant for the comparison,

$$P_r(H_1)p_1(\underline{v}) \underset{H_0}{\overset{H_1}{\gtrless}} P_r(H_0)p_0(\underline{v}) \qquad \Lambda(\underline{v}) \stackrel{\Delta}{=} \frac{p_1(\underline{v})}{p_0(\underline{v})} \qquad \Lambda_0 \stackrel{\Delta}{=} \frac{\Pr(H_0)}{\Pr(H_1)} \qquad \Lambda(\underline{v}) \underset{H_0}{\overset{H_1}{\gtrless}} \Lambda_0$$

the first above expression must be evaluated, choosing the appropriate H_j. Equivalently, introducing the **likelihood ratio** $\Lambda(\underline{v})$, as above, it can be compared with a fixed decision threshold, determined through the prior probabilities.

In binary decisions, two kinds of errors can be made: an error of the first kind happens when choosing for the alternative hypothesis when no signal is present: this is called **a false alarm** and denoted by $(\rightarrow H_1|H_0)$. Errors of the second kind happen whenever choosing the null hypothesis while H_1 is at hand: it is called **a false dismissal** and denoted by $(\rightarrow H_0|H_1)$. In practical situations, false alarms and false dismissals may have different costs of operation and even correct decisions, like $(\rightarrow H_1|H_1)$, which is called a **detection**, may have their costs. Following Bayes's rule, the average cost of operation is

$$\bar{C} = \int_{R_0} C(\rightarrow H_0|\underline{v}) p(\underline{v}) d^n v + \int_{R_1} C(\rightarrow H_1|\underline{v}) p(\underline{v}) d^n v$$

and, through some calculations, it is seen that it can be minimized again by the same decision rule introduced before, comparing now the likelihood ratio against a modified threshold.

$$\Lambda'_0 = \frac{P_r(H_0)[C(\rightarrow H_1/H_0) - C(\rightarrow H_0/H_0)]}{P_r(H_1)[C(\rightarrow H_0/H_1) - C(\rightarrow H_1/H_1)]}$$

This refinement of the Bayes's rule is referred to as the **Bayes's strategy.**

The core of binary hypotheses testing is, as for the general M-ary case, to find the hypersurface D that splits the data space R^n in two disjoint regions, R_0 and R_1, related with the two hypotheses. The probabilities Q_0 of a false alarm and Q_d of correct detection will then read

$$Q_0 \stackrel{\Delta}{=} P_r(\rightarrow H_1/H_0) = \int_{R_1} p_0(\underline{v}) d^n v \qquad Q_d \stackrel{\Delta}{=} P_r(\rightarrow H_1/H_1) = \int_{R_1} p_1(\underline{v}) d^n v$$

and are referred to as the **size** and the **power** of the test, respectively. Through these, it is easy to calculate the probability of error $P_e = Pr(H_0)Q_0 + Pr(H_1)(1 - Q_1)$ or the probability of a false dismissal, which is $(1\text{-}Q_d)$.

2.3 The Neyman-Pearson criterion

In many practical applications, it is difficult, if not impossible, to establish an a-priori value for the probabilities of H_0 and H_1: in the detection of gravitational waves it could even be meaningless to state the probability with which a GW signal is present in the observed data. But, in those situations where H_1 happens "rarely" it can be seen, from its last expression, that the probability of error is dominated by false alarms.

Therefore, an alternative reasonable decision criterion, first established by Neyman and Pearson in 1933, would be "for a given affordable Q_0, seek a strategy maximizing Q_D", or, otherwise stated, **given the size, maximize the power of the test.** We shall now see that the optimal strategy under this criterion again calls for the likelihood ratio, requiring the comparison with a certain threshold λ:

$$\Lambda(\underline{v}) \stackrel{\Delta}{=} \frac{p_1(\underline{v})}{p_0(\underline{v})} \underset{H_0}{\overset{H_1}{\gtrless}} \lambda \qquad (1)$$

In order to demonstrate it, note that, according to the last expression, the hypersurface D contains data points for which $\Lambda(\underline{v}) = \lambda$, while on the R_0 and R_1 sides of D the likelihood ratio is smaller or larger than λ, respectively.

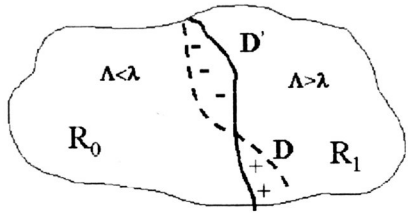

Figure 7. An intuitive demonstration for the Neyman-Pearson decision strategy.

Assume that the strategy is changed by modifying D into another hypersurface D' while keeping Q_0 constant, as sketched in fig.7. Since Q_0 is the integral over R_1 of the conditional pdf $p_0(\underline{v})$, this modification can be accomplished in an infinity of ways, but, whichever may be the new D' (solid line in fig.7), the power

$$Q_d = \int_{R_1} p_1(\underline{v}) d^n v = \int_{R_1} \Lambda(\underline{v}) p_0(\underline{v}) d^n v$$

is easily seen to have decreased. Intuitively, we can see that by modifying D we gave away from R_1 those points, labelled with "-" in the figure, while gaining the points labelled "+" which are of "less value" if measured by $\Lambda(\underline{v})$, as in the above integral.

Since the optimal test is that expressed in (1), how does one set the threshold λ? The first thing to note is that $\Lambda(\underline{v})$ is itself a random variable, depending on the random data, with some conditional pdfs $P_j(\Lambda)$ that can be calculated, under each H_j. The size and the power of the test can be then rewritten as

$$Q_0 = P_r(\Lambda(\underline{v}) > \lambda / H_0) = \int_\lambda^\infty P_0(\Lambda) d\Lambda; Q_d = P_r(\Lambda(\underline{v}) > \lambda / H_1) = \int_\lambda^\infty P_1(\Lambda) d\Lambda$$

and λ be determined from the first integral once the size is fixed, according to the Neyman-Pearson criterion. The same criterion ensures that the second integral yields the maximum possible value. Note that, under this new criterion the threshold is independent of the prior probabilities $\Pr(H_j)$, which can be unknown.

2.4 Sufficient statistics

As seen in the previous sections, both under Bayes's and Neyman-Pearson's criteria, the likelihood ratio $\Lambda(\underline{v})$ embodies all the relevant information, extracted from the data, for performing an optimal decision. Any function of the data points with the latter property is called a **sufficient statistics** for the detection test: besides Λ, other sufficient statistics exist, such as, for instance, any monotone increasing function $G(\Lambda)$ of the likelihood ratio. This should of course be compared with a different threshold G_0, to be determined: in order to set a proper G_0, the conditional pdfs of the new random variable G must be calculated. Once $P_0^G G$ and $P_1^G G$ and are known, one of the two following equations can be used

$$\frac{P_1^G(G_0)}{P_0^G(G_0)} = \Lambda_0 \qquad Q_0 = \int_{G_0}^\infty P_0^G(G) dG$$

Bayes *Neyman – Pearson*

to determine G_0, depending on which criterion is adopted. The most common among these other sufficient statistics is the **log-likelihood ratio**

$$G(\Lambda) = ln\left[\frac{p_1(\underline{v})}{p_0(\underline{v})}\right]$$

of which the usefulness will be evident from the following sections.

2.5 Detection in Gaussian noise

Let us now switch back to the problem of obtaining the data samples $\underline{v}$ from the observed signal $v(t)$. The most celebrated problem in detection theory is to recognize if, observing a noisy signal $v(t)$ in the time interval $0 < t < T$, there is a useful signal $s(t)$ "immersed" in gaussian noise $n(t)$ or else $v(t)$ consists of purely noise. Therefore, this is a binary hypothesis test, for which the alternative hypothesis H_1 is expressed by the equation $v(t) = s(t) + n(t)$, while the null hypothesis H_0 can be written as $v(t) = n(t)$; it is assumed here

that the signal $s(t)$ is completely determined while noise is described by its ensemble averages. It is important to recognize that, under both hypotheses, $v(t)$ is a gaussian process: then it can be described by its ensemble averages, computed on the samples of $v(t)$. Instead of using time samples, like $v(t_i)$, assume to use the following original sampling method.

A set of function $\{f_k(t)\}$ is said to be orthonormal on $t \in [0,T]$ if they obey the following equalities

$$\int_0^T f_k(t) f_j(t) dt = 0 \qquad \int_0^T f_k^2(t) dt = 1$$

Some examples of such sets are the Fourier harmonics and the Legendre polynomials. Using an orthonormal set, any function $v(t)$ can be represented on [0,T] by a Fourier-like series using its coefficients v_k, according to the following equations

$$v(t) = \sum_{k=1}^{\infty} v_k f_k(t) \qquad v_k = \int_0^T v(t) f_k(t) dt \qquad \int_0^T v^2(t) dt = \sum_{k=1}^{\infty} v_k^2$$

and the last equation is a Parseval-like theorem expressing the energy of $v(t)$ on the interval [0,T]. The coefficients v_k are easily seen to be Gaussian RVs, since $v(t)$ is Gaussian: their mean value and correlations, under the null hypothesis H_0, can be computed as follows, using the linearity of expectation E[.] and the definition of autocovariance of the noise process $\phi(t,s) = E[n(t)n(s)]$.

$$E[v_k / H_0] = \int_0^T E[n(t)] f_k(t) dt = 0$$

$$\varphi_{jk} = E[v_j v_k / H_0] = \int_0^T \int_0^T f_j(t) E[n(t)n(s)] f_k(s) dt\, ds = \int_0^T \int_0^T f_j(t) \varphi(t,s) f_k(s) dt\, ds$$

One of the reasons for using the samples v_k and not time samples is that if $n(t)$ is white noise the variance of time samples would be infinite.

When setting up the likelihood ratio for the data samples v_k, it is extremely useful to have uncorrelated samples: this task can be accomplished by choosing a particular set of orthonormal functions. Suppose we select those functions which are the eigenfunctions of the following **homogeneous Fredholm integral equation**

$$\lambda_k f_k(t) = \int_0^T \phi(t,s) f_k(s) ds$$

Here the kernel $\phi(t,s)$ acts as a linear operator that "stretches and rotates" functions in an Hilbert space of infinite dimension: the eigenfunctions being those that are stretched but not rotated. It is possible to show that if the kernel is real and symmetrical (or complex but hermitian) the equation has real eigenvalues and orthogonal eigenfunctions, as demonstrated by the following equations

$$\lambda_k \int_0^T f_j(t) f_k(t) dt = \int_0^T \int_0^T f_j(t) \varphi(t,s) f_k(s) ds\, dt$$

$$\lambda_j \int_0^T f_k(t) f_j(t) dt = \int_0^T \int_0^T f_k(t) \varphi(t,s) f_j(s) ds\, dt$$

$$\textit{subtracting} \Rightarrow (\lambda_k - \lambda_j) \int_0^T f_k(t) f_j(t) dt = 0$$

Once demonstrated orthogonality, orthonormality can be accomodated by a multiplying factor for each function. These particular functions constitute the **Karhunen-Loève** basis: expanding $v(t)$ with this basis implies uncorrelated samples v_k, each with variance λ_k

$$\phi_{jk} = \lambda_k \int_0^T f_j(t) f_k(t) dt = \begin{cases} \lambda_k & j = k \\ 0 & j \neq k \end{cases}$$

and, since the samples are uncorrelated and Gaussian, they are also independent.

It is now simple to calculate the likelihood ratio. Suppose, for the sake of simplicity, that we base our decision on the first n samples only: since v_k are independent Gaussian variables with variance λ_k, under both hypotheses, zero mean under H_0 and mean value

$$E[v_k | H_1] = \int_0^T s(t) f_k(t) dt = s_k$$

under H_1, where s_k is the k-th coefficient of $s(t)$ expanded with Karhunen-Loève basis, the joint conditional pdfs of the samples will be

$$p_0(v_1, v_2 \ldots v_n) = \prod_{k=1}^{n} \frac{1}{\sqrt{2\pi\lambda_k}} e^{-\frac{(v_k)^2}{2\lambda_k}} \qquad p_1(v_1, v_2 \ldots v_n) = \prod_{k=1}^{n} \frac{1}{\sqrt{2\pi\lambda_k}} e^{-\frac{(v_k - s_k)^2}{2\lambda_k}}$$

from which it is easy to form the likelihood ratio $\Lambda_n(\underline{v})$. Letting $n \to \infty$, $\Lambda_n(\underline{v})$ becomes a sufficient statistics, the linear and log-likelihood ratios then

become

$$\Lambda(\underline{v}) = e^{\sum_{k=1}^{\infty}\left[\frac{s_k v_k}{\lambda_k} - \frac{s_k^2}{2\lambda_k}\right]} \qquad G = \sum_{k=1}^{\infty} \frac{s_k v_k}{\lambda_k}$$

The log-likelihood ratio G is a Gaussian random variable, that can be practically computed avoiding the calculation of the coefficients v_k, s_k, λ_k. Resorting to the definition of v_k

$$G = \sum_{k=1}^{\infty} \frac{s_k}{\lambda_k} \int_0^T f_k(t) v(t) dt = \int_0^T \left[\sum_{k=1}^{\infty} \frac{s_k}{\lambda_k} f_k(t)\right] v(t) dt \qquad (2)$$

and the term in square brackets is the expansion, on the Karhunen-Loève basis, of a function $q(t)$ that is seen to solve the **inhomogeneous Fredholm equation**

$$\int_0^T \phi(t,u) q(u) du = \sum_{k=1}^{\infty} \frac{s_k}{\lambda_k} \lambda_k f_k(t) = s(t)$$

in which the useful signal $s(t)$ appears on the right hand side. Therefore, G can be computed by integrating the product $q(t)v(t)$ on [0,T]. Its statistics, conditional mean and variance, are

$$Var[G] = \int_0^T \int_0^T q(t) E[n(t)n(u)] q(u) du dt = \int_0^T q(t) s(t) dt \overset{\Delta}{=} d^2$$

$$E[G/H_0] = \int_0^T q(t) E[n(t)] dt = 0 \qquad E[G/H_1] = \int_0^T q(t) E[s(t) + n(t)] dt = d^2$$

where the symbol d^2 has been introduced: we will further exploit the meaning of this parameter. Through the Gaussian conditional pdfs of G, its likelihood ratio, which is referred to as the **likelihood functional**, can be computed

$$\Lambda(G) = \frac{P_1^G(G)}{P_o^G(G)} = \exp\left[\int_0^T q(t) v(t) dt - \frac{1}{2}\int_0^T s(t) q(t) dt\right]$$

where the first integral coincides with G and the second with the definition of d^2. The sufficient statistics G is suitable for the detection problem under both decision criteria: with Bayes detection it must be compared with the threshold G_0

$$G \underset{H_0}{\overset{H_1}{\gtrless}} G_0 = \frac{1}{2} d^2 + \ln \Lambda_0$$

while with Neyman-Pearson detection G_0 must be calculated from

$$Q_0 = \frac{1}{\sqrt{2\pi d^2}} \int_{G_0}^{\infty} e^{-\frac{G^2}{2d^2}} dG = erfc\left(\frac{G_0}{d}\right)$$

$$Q_d = \frac{1}{\sqrt{2\pi d^2}} \int_{G_0}^{\infty} e^{-\frac{(G-d^2)^2}{2d^2}} dG = erfc\left(\frac{G_0}{d} - d\right)$$

Using the latter expressions, it is possible, for a given Q_0, to evaluate G_0 for any value of d and then compute Q_d, deriving the plot of Fig.8, which is reprinted from [1]. This plot evaluates the performance of the receiver, under Neyman-Pearson criterion, giving the attainable power Q_d once the size Q_0 is chosen: the parameter d depends, by definition, only on the signal $s(t)$ to be detected and the noise autocovariance $\phi(t,u)$.

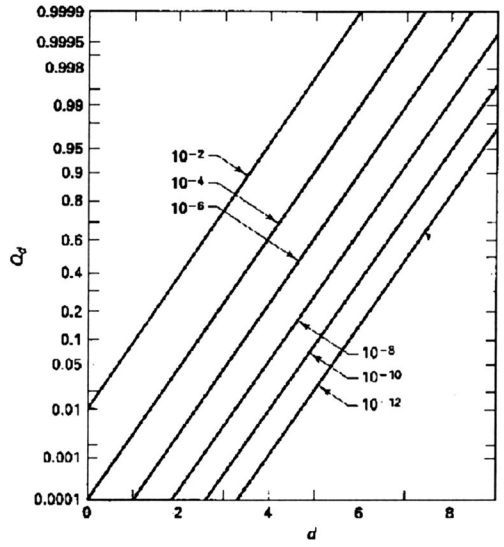

Figure 8. Probability of detection: curves indexed by the probability of false alarm Q_0.

Detection in stationary white noise

In the simple case of stationary white noise, assume the covariance spectrum is constant and equal to $\Phi(f) = N/2$, the autocovariance function is then $\phi(t,u) = (N/2)\delta(t-u)$ and the solution to the inhomogeneous Fredholm equation is simply

$$\int_0^T \frac{N}{2}\delta(t-u)q(u)du = s(t) \Rightarrow q(t) = \frac{2}{N}s(t)$$

provided that the signal $s(t)$ vanishes outside the observation interval [0,T]. In this case, the statistics G can be found through the integral

$$G = \int_0^T s(t)v(t)dt$$

referred to as the **correlation** between the useful signal and the observed process $v(t)$; the factor 2/N can be absorbed into the decision threshold G_0. The likelihood functional is simplified as well, and the term d^2 represents the signal energy to power spectral density ratio, briefly referred to as the **signal-to-noise ratio**:

$$d^2 = \frac{2}{N}\int_0^T s^2(t)dt = \frac{2E_s}{N}$$

Note that since d grows with T, performance also grows with the observation time, which is an intuitive result.

The optimality of the log-likelihood ratio G was first formalized by Granader in 1950, stating that the statistic G is optimum for the detection problem if $q(t)$ is square integrable on [0,T]; these results were further extended in 1967 by Kadota demonstrating that it suffices that d^2 is finite for G being optimum. If d is infinite, it is seen that Q_0=0 and Q_D=1: in this case a singular detection is said to occur.

The matched filter

In the general case of Gaussian noise, the correlation between the observed $v(t)$ and $q(t)$, defining G in accordance with (2), can be practically realized by constructing a filter with impulse response $k(t) = q(T-t)$, for $0 < t < T$, and zero elsewhere. This filter, first introduced in 1943, is called the **matched filter** since its impulse response is the "mirrored and shifted" version of $q(t)$. The output of the matched filter is found by convolution between $k(t)$ and the observed $v(t)$

$$y(t) = \int_0^t v(\tau)k(t-\tau)d\tau = \int_{t-T}^t v(\tau)q(T-t+\tau)d\tau$$

and can be sampled at time $t = T$ to yield the seeked value of G. Fig. 9 shows

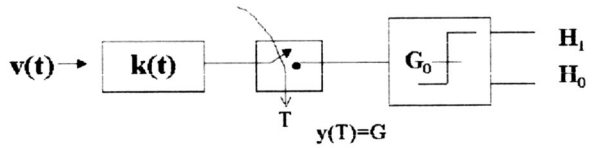

Figure 9. Optimal receiver scheme for binary detection in Gaussian noise.

the scheme of an optimal detector in this case: it is made up by a matched filter, followed by a sampler and a binary quantizer with the proper threshold G_0. It could be also demonstrated that the use of a matched filter minimizes the effect of noise on the random value of G that exits the sampler.

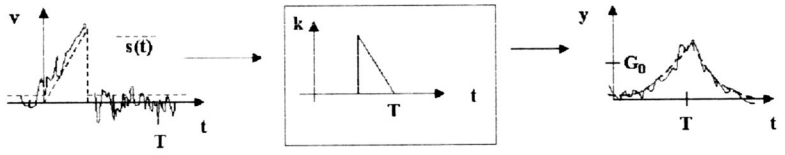

Figure 10. An example of the matched filter action, in the case of white noise.

In the simple case of white noise $q(t) = s(t)$, and thus the impulse response of the filter $k(t)$ will be matched to the useful signal $s(t)$. Fig. 10 depicts how the noisy signal $v(t)$, which is the sum of $s(t)$ (dashed line) and white noise, is filtered producing an output whose peak value would be on the sampling time T in the absence of noise (dashed line): random fluctuations, due to noise, may bring this value below the threshold thus causing a false dismissal. The use of a matched filter minimizes these fluctuations. The matched filter, previously introduced in communication theory, was first applied to a detection problem by Peterson, Birdsall and Fox in 1954. It must be noted that the signal to be detected has to be completely known in all its parameters, including the arrival time.

3 Elements of estimation theory

So far, we have been dealing with the problem of recognizing the presence of a finite number, one or more, of specified signals in the midst of noise; this problem is referred to as the **simple hypotheses testing.** Since these signals

have been assumed completely known, and so is the statistical description of noise, the conditional pdfs $p_j(\underline{v})$ of the observed samples are also completely known. But signals to be detected are rarely univocally specified: more often it happens that these signals are specified only partially, since they depend on one or more unknown parameters (their amplitude, time of arrival, carrier frequency...). Thus the problem becomes to recognize if any among an entire class of signals has been detected (H_1) or only noise is present (H_0), the class $s(t;\underline{\theta})$ being specified by a vector $\underline{\theta} = (\theta_1, \theta_2, ..., \theta_m)$ of unknown parameters: in this case we talk of **composite hypotheses testing**. Even worse, it may be necessary, in the hypothesis H_1 that a useful signal is present, to evaluate those parameters, based on the observation of $v(t)$: this is the **parameter estimation** problem.

To introduce some terminology, the m unknown parameters are called the **estimanda**: these are, in general, random variables whose prior joint pdf $z(\theta_1, \theta_2, ..., \theta_m)$ may be known or not. Under hypothesis H_1, the observed signal is $v(t) = s(t;\underline{\theta}) + n(t)$ and the joint conditional pdf of data samples becomes dependent on the unknown parameters: it can be expressed as $p(\underline{v};\underline{\theta}) = p_0(\{v_k - s_k(\underline{\theta})\})$, where $p_0(\{n_k\})$ is the joint pdf of noise samples alone and $s_k(\underline{\theta})$ are samples extracted from the useful signal. This way, the estimanda become parameters of the distribution of the observations: our problem is then to seek those values of θ_i for which $p(\underline{v};\underline{\theta})$ best describes the observed data, in some sense. We must then seek a strategy that associates m numbers $\hat{\theta}_i$, called the estimates, to the estimanda, based on the data. This strategy will determine a vector of m functions of the data that performs this association: such a vector $\underline{\hat{\theta}} = (\hat{\theta}_1(\underline{v}), \hat{\theta}_2(\underline{v}), ..., \hat{\theta}_m(\underline{v}))$ is called the **estimator**.

We can think of the estimanda and the estimates to lie in a parameters space Θ: since the estimation is performed in the presence of noise, the estimation procedure repeated in two different experiments will produce two different values for the estimates, even if the estimanda are the same. Thus, the estimates will be always in error: estimates and estimanda are then, in general, represented by two different points in Θ. The aim of **estimation theory** is the design of receivers that minimize the errors due to noise and, possibly, predict how large the irreducible errors will be, on the average.

The key to the solution of the estimation problem is the knowledge of the joint conditional pdf of the data $p(\underline{v}|\underline{\theta})$, given the estimanda. In order to derive an estimation strategy, let us define some quantities that give indications about the goodness of a particular estimator: it is intuitive that a good estimator $\hat{\theta}_k$ is required to be close to the true value θ_k (the estimandum), on the average.

$$\overline{\hat{\theta}}_k \triangleq E[\hat{\theta}_k(\underline{v})\,|\,\underline{\theta}] \qquad\qquad E[\hat{\theta}_k(\underline{v}) - \theta_k)\,|\,\underline{\theta}] = \overline{\hat{\theta}}_k - \theta_k$$

The two above expressions define the estimator expected value and the mean deviation from the true value, also called the **bias**, respectively; if the bias is equal to zero, the estimator is said to be unbiased. Both expectations are conditioned on the true values of the estimanda; expected values are then evaluated with respect to the statistical variations of noise only. Another fundamental figure of merit for an estimator is the **Mean Square Error** (MSE)

$$\overline{E}_k \triangleq E[(\hat{\theta}_k(\underline{v}) - \theta_k)^2 | \underline{\theta}] \qquad \Rightarrow \quad \overline{E}_k = Var[\hat{\theta}_k(\underline{v})] + (\overline{\hat{\theta}}_k - \theta_k)^2$$

that can also be computed through the variance of the estimator and the bias, as shown above. It is evident that a good estimator should have a low bias and a low variance but, in practice, this is often a compromise.

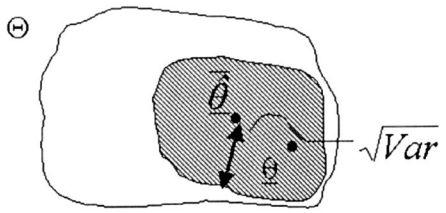

Figure 11. Estimator expected value and estimate, in the parameters space Θ.

Figure 11 shows the shaded region, whose *radius* is the estimator standard deviation, around the estimator mean value in which the actual estimator value will more often fall; the bias is the distance between the two points in the area. The calculation of bias and MSE requires the knowledge of the conditional pdf $q(\hat{\theta}_k|\underline{\theta})$ of every estimate: these are obtainable from the conditional pdf of data $p(\underline{v}|\underline{\theta})$, since the estimator elements $\hat{\theta}_k$ are deterministic functions of the data.

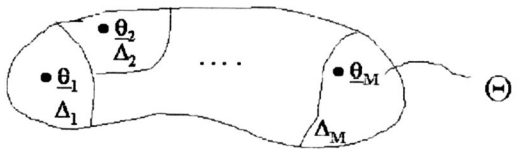

Figure 12. Partition of the parameters space: every region is associated with an hypothesis H_j.

3.1 Maximum a posteriori (MAP) estimators

Suppose to partition the parameters space in M disjoint regions, each called Δ_j and having a vector $\underline{\theta}_j$ in its "center", as depicted in Fig. 12. If we introduce M events H_j, each stating that "the vector of estimanda $\underline{\theta}$ lies in region Δ_j", we can choose one of these events, based on the observation of the data $\underline{v}$, and then associate the data to the i-th central point $\underline{\theta}_j$, so defining the estimator function: $\hat{\underline{\theta}}(\underline{v}) = \underline{\theta}_j$. This way, the estimation problem is reduced to a multiple hypotheses testing problem, whose solution can be found by applying the strategies devised in the previous chapter. In particular, following the Bayes's rule, one should choose the i-th hypothesis for which the posterior probability $\Pr(H_i|\underline{v})$ is largest. In the present case, these probabilities can be computed through the expression

$$P_r(H_i|\underline{v}) = \int_{\Delta_i} p(\underline{\theta}|\underline{v}) d^m\theta$$

which in turn requires the knowledge of the conditional pdf of parameters, given the data. This function is in general not given but it can be computed through the application of the Bayes's theorem in its continuous form

$$p(\underline{\theta}|\underline{v}) = \frac{z(\underline{\theta})p(\underline{v}|\underline{\theta})}{p(\underline{v})}$$

since we assumed to know the conditional probability of the data, given the estimates, appearing in the numerator.

It is evident that an estimator associating the data to the M points $\underline{\theta}_i$ is not the optimal choice but, passing to the limit for M$\rightarrow \infty$, the regions Δ_j become smaller and smaller thus making $\Pr(H_i|\underline{v}) \propto p(\underline{\theta}_i|\underline{v})$. The optimal estimator, according to the MAP strategy, implementing the Bayes's rule, is then the one selecting the estimates that maximize the conditional pdf of the estimates, given the data. Stated mathematically, the strategy is to choose

that satisfies

$$p(\hat{\underline{\theta}}(\underline{v})|\underline{v}) \geq p(\underline{\theta}|\underline{v}) \quad \forall \underline{\theta} \in \Theta$$

To gain a deeper insight on the MAP strategy, let us denote with $\underline{\theta}'$ one of the many possible estimates (not necessarily the selected one): the conditional pdf of the estimate $q(\underline{\theta}'|\underline{\theta})$, given the true estimanda, for $\underline{\theta}'$ to be a good estimate should be sharply peaked around the true value $\underline{\theta}'=\underline{\theta}$. While

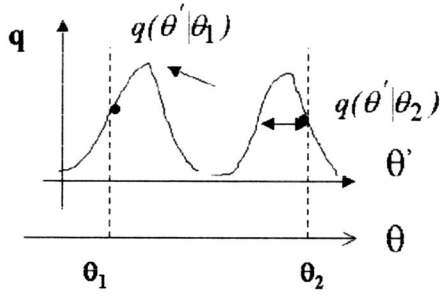

Figure 13. Distribution of the estimate θ', given two possible values for the estimandum.

in the estimation problem the estimanda are random by definition, also the selected estimates are random, because of noise, prior to the observation of data. In the pictorial representation of Fig. 13, where the estimandum is assumed a scalar θ for simplicity, even when the true estimandum is given, for example $\theta = \theta_2$, we do not know which $\underline{\theta}'$ will be the actual estimate before we observe the noisy data: $\underline{\theta}'$ a priori spans an entire interval, labelled $\leftrightarrow$ in the figure. The more $q(\underline{\theta}'|\underline{\theta})$ is peaked around $\theta = \theta_2$, the more accurate the estimate will be, on the average. Since we do not know what value the actual estimandum will have, it can be introduced a measure of the average peakedness for the pdf $q(\underline{\theta}'|\underline{\theta})$, according to the following expression

$$\bar{Q} = \int_{\Theta} z(\underline{\theta}) q(\underline{\theta}|\underline{\theta}) d^m\theta$$

that averages $q(.)$ with respect to the prior distribution of the estimanda. The meaning of $q(\underline{\theta}'|\underline{\theta})$ is the probability that the chosen estimate is $\underline{\theta}'$: it can be computed by summing the probability that data takes values on which the estimator $\hat{\underline{\theta}}(\underline{v})$ is equal to $\underline{\theta}'$, according to the following equation

$$q(\underline{\theta}'|\underline{\theta}) = \int_{R^n} \delta(\underline{\theta}' - \hat{\underline{\theta}}(\underline{v})) p(\underline{v}|\underline{\theta}) d^n v$$

Putting the latter expression into the previous integral gives a new expression for $\bar{Q}$, that can be further simplified using Bayes's theorem

$$\bar{Q} = \int_{\Theta} z(\underline{\theta}) \int_{R_n} \delta(\underline{\theta} - \underline{\hat{\theta}}(\underline{v})) p(\underline{v}|\underline{\theta}) d^n v d^m \theta \quad \Rightarrow \quad \bar{Q} = \int_{R_n} p(\underline{v}) p(\underline{\hat{\theta}}(\underline{v})|\underline{v}) d^n v$$

Since the MAP strategy chooses the estimator maximing $p(\underline{\hat{\theta}}(\underline{v})|\underline{v})$, it also maximizes the last integral ($p(\underline{v}) \geq 0$) and thus $\bar{Q}$.

3.2 Maximum Likelihood (ML) estimates

Through Bayes's theorem it can also be said that MAP strategy seeks the maximization of the product $z(\underline{\theta})p(\underline{v}|\underline{\theta})$. Unfortunately, the first factor, $z(\underline{\theta})$, is often unknown or even with little meaning (for instance when the parameters are deterministic but unknown), thus making the applicability of MAP strategy impossible, since it requires the knowledge of the prior distribution of the parameters. In all those situations in which there is great uncertainty on the parameters, the prior distribution of these can be thought to be much broader than the conditional distribution of data samples $p(\underline{v}|\underline{\theta})$, which is only due to noise. Referring to Fig. 14, where scalars are employed for simplicity, suppose that the scalar datum v_0 is observed and that the conditional distribution of data $p(v|\theta)$ has a maximum on v_0 for $\theta = \theta_1$. Since the prior pdf of the parameter, $z(\theta)$, is increasing for $\theta > \theta_1$, we can expect that the optimal MAP estimate lies somewhere between θ_1 and θ_2, the latter point identifying the maximum $z(\theta_2)$ for the distribution of the parameter. If, as assumed, the function $z(\theta)$ is relatively flat, we can reasonably think that the MAP estimate, i.e. the one maximizing $z(\underline{\theta})p(\underline{v}|\underline{\theta})$, lies close to θ_1, which maximizes $p(\underline{v}|\underline{\theta})$ only.

The choice of an estimate for which the conditional distribution of data $p(\underline{v}|\underline{\theta})$ is maximum on the observed data $\underline{v}$ is called the **Maximum Likelihood** (ML) strategy: the estimator $\underline{\hat{\theta}}(\underline{v})$ is chosen as to satisfy the following property

$$p(\underline{v}|\underline{\hat{\theta}}(\underline{v})) \geq p(\underline{v}|\underline{\theta}) \quad \forall \underline{\theta} \in \Theta$$

Thinking again of the product $z(\underline{\theta})p(\underline{v}|\underline{\theta})$, it is recognized that when the pdf of $z(\underline{\theta})$ parameters is very peaked, almost all the information on the estimanda is contained in it and, knowing $z(\underline{\theta})$, one can reliably guess an estimate without observing the data. If, on the other hand, the function $z(\underline{\theta})$ is really broad, it can be neglected in the comparison between different estimates, thus reducing the MAP strategy to the ML strategy. When $z(\underline{\theta})$ is not known, there is no choice: it must be neglected and the ML strategy is

to be applied. A useful property of ML estimates is that, if in one problem it is easier to estimate not the parameter but a monotone function of it (e.g. $f(\theta) = \theta^3$) we can estimate $\hat{\theta^3}$ and then set $\hat{\theta} = \sqrt[3]{\hat{\theta^3}}$, using the inverse function of $f(.)$. This useful procedure cannot be applied with MAP estimates since a different weight is assigned, through $z(\theta)$, to corresponding ranges of θ and $f(\theta)$, in the transformation performed by $f(.)$.

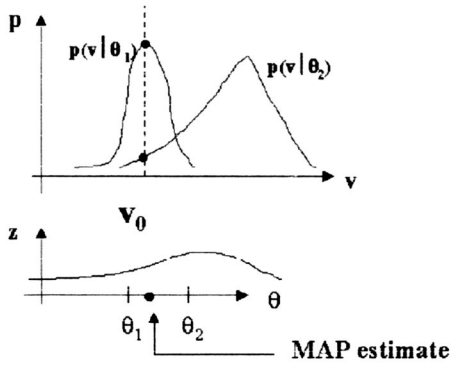

Figure 14. Correspondence between MAP estimate and ML estimate, for scalar parameter and data.

3.3 The likelihood functional and the Greater Likelihood Ratio Test

The strategies just described both require the comparison of the conditional pdfs of data $p(\underline{v}|\underline{\theta})$, for different values of $\underline{\theta}$, for deriving an estimator. To overcome the difficulty of having to deal with an infinity of samples, when the dimension n of the vector $\underline{v}$ grows, the latter pdf can be divided by the joint pdf $p_0(\underline{v})$ of pure noise samples, which does not alter the results of comparisons. Provided that $\underline{\theta}$ appears only in the expression of the useful signal $s(t;\underline{\theta})$, we can then resort to the function

$$\Lambda[v(t)|\underline{\theta}] = \lim_{n\to\infty} \frac{p(\underline{v}|\underline{\theta})}{p_0(\underline{v})}$$

which is referred to as the **likelihood functional**. Thus, the ML strategy requires finding the estimator $\underline{\hat{\theta}}[v(t)]$ for which $\Lambda[v(t)|\underline{\theta}]$ is the largest. In the case of white Gaussian noise, it is easy to see that:

$$\Lambda[v(t)|\underline{\theta}] = \exp\left[\frac{2}{N}\int_0^T v(t)s(t;\underline{\theta})dt - \frac{1}{2}\frac{2}{N}\int_0^T [s(t;\underline{\theta})]^2 dt\right] \quad (3)$$

where the last integral coincides with the energy d^2 of the signal s with parameters $\underline{\theta}$.

How can we practically search for an estimate maximizing the likelihood functional: in the case of white Gaussian noise, a solution is to build a **bank of matched filters**, with impulse responses $k(t;\underline{\theta}_j)$, each filter being matched to a signal $s(t;\underline{\theta}_j)$ identified by a different set of parameters, as depicted in Fig. 15. As outlined in the figure, this approach corresponds to a *sampling of*

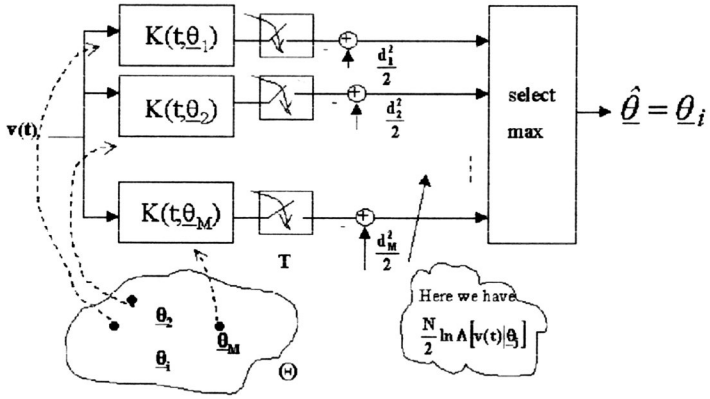

Figure 15. Receiver with a bank of matched filters.

the parameters space Θ: at the output of the j-th filter we have the result of the first integral appearing in (3), for $\underline{\theta} = \underline{\theta}_j$; through the subtraction of the corresponding energy d_j^2, a monotonic function of the likelihood functional can be maximized for choosing the estimate. Ideally, a continuum of filters would be needed, which is impossible, and thus the chosen estimate is always an approximation of the true estimanda. Some countermeasures to enhance the reliability of the estimate would be to provide parameter sets $\underline{\theta}_j$ that are densely spaced over the region of Θ where the true estimanda are expected to lie. Since it is not implicit that the observed signal $v(t)$ contains some useful signal in it (it can be only noise, in the hypothesis H_0) it is possible, after the estimation procedure, to switch back to the detection problem, deciding for a H_0 or H_1 based on the comparison of the selected likelihood functional

with a threshold that can be chosen according to the guidelines of the previous chapter. This approach is called the **Greater Likelihood Ratio Test** (GLRT) and is formalized by the following decision rule

$$\max_{\underline{\theta}} \Lambda\left[v(t)|\underline{\theta}\right] = \Lambda\left[v(t)|\underline{\hat{\theta}}\right] \underset{H_0}{\overset{H_1}{\gtrless}} \Lambda_0$$

Estimation of arrival time

In general, the designer of the receiver should avoid building a bank of matched filters: however, this is possible only in some simple cases. To help understanding, suppose the problem is to determine if a certain signal $s(t)$ of finite duration T' is detected in the presence of white Gaussian noise. The signal can arrive at the receiver with a random time displacement θ, to be estimated, of which we don't know the prior distribution. The observed signal is thus $v(t) = s(t-\theta_0) + n(t)$, where θ_0 is the actual arrival time, as shown in Fig. 16. We assume that the observation extends in [0,T].

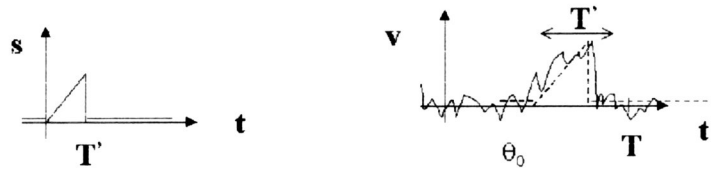

Figure 16. Signal with random arrival time corrupted by white Gaussian noise.

According to the ML estimation strategy, the receiver must maximize the likelihood functional or else its logarithm, the log-likelihood functional.

$$\max_{\theta} \Lambda\left[v(t)|\theta\right] \Rightarrow \max_{\theta}\left[\int_0^T v(t)s(t-\theta)dt\right]$$

This would require building a bank of filters, each with an impulse response $k(t;\underline{\theta}_j) = s(T-(t-\underline{\theta}_j))$. Practically, only one filter with $k(t) = s(-t)$ is

needed: the output of the filter at time θ

$$G(\theta) = \int_0^T v(t)s(t-\theta)dt \approx \int_{-\infty}^{\infty} v(\tau)s(-(\theta-\tau))d\tau$$

coincides with the log-likelihood functional and thus the receiver will choose, as the estimate, the time t_m when the output reaches its maximum.

3.4 Performance of ML estimators

In this section we wish to derive an expression for the bias and the variance of a ML estimator, these parameters representing the figures for evaluating the performance of an estimator. For the sake of simplicity, we will develop a general approach for the case of a single parameter θ, with true value θ_0, using the natural logarithm of the likelihood functional, called **log-likelihood ratio** $g(v|\theta)$, where v briefly represents the observed signal. Since $p_0(v)$, appearing in the definition of $\Lambda[v|\theta]$, does not depend on θ, it will be omitted for the moment without loss of generality: the definition of the log-likelihood ratio is thus $g(v|\theta) = \ln p(v|\theta)$. In parallel, the approach will be applied to the practical case of estimation of the arrival time, introduced in the last subsection, using the complete definition for $g(v|\theta)$, that reads

$$g(v|\theta) = \ln\frac{p(v|\theta)}{p_0(v)} = \frac{2}{N}\int_0^T v(t)s(t-\theta)dt - \frac{1}{2}\frac{2}{N}\int_0^T [s(t-\theta)]^2 dt \quad (4)$$

where the observed noisy signal is $v(t) = s(t-\theta_0) + n(t)$, in this case.

Since, in general the selected estimator $\hat{\theta}$ is the one maximizing g, the derivative of g with respect to θ must have a zero for $\theta = \hat{\theta}$ such derivative is denoted by g'_θ and is given by

$$g'_\theta(v|\theta) \stackrel{\Delta}{=} \frac{\partial g(v|\theta)}{\partial\theta} = \frac{p'_\theta(v|\theta)}{p(v|\theta)}$$

The value of g'_θ on the estimator value can be expressed through a Taylor series expansion in the neighbourhood of the true estimandum θ_0: assuming the estimator is sufficiently close to the estimandum, the terms above the first can be neglected

$$g'_\theta(v|\hat{\theta}) = g'_\theta(v|\theta_0) + (\hat{\theta}-\theta_0)g''_\theta(v|\theta_0) + \ldots = 0 \quad (5)$$

In the expression (4), for the arrival time, the second integral is constant with θ, being the energy of $s(.)$, when the observation time T is sufficiently large; the first and second derivatives then read

$$g'_\theta(v|\theta) = -\frac{2}{N}\int_0^T v(t)s'(t-\theta)dt \qquad g''_\theta(v|\theta) = \frac{2}{N}\int_0^T v(t)s''(t-\theta)dt$$

Based on (5) we can then approximate the deviation of the estimator from the estimandum: in the general case and for the arrival time case its expressions is

$$\hat{\theta}-\theta_0 \approx -\frac{g'_\theta(v|\theta_0)}{g''_\theta(v|\theta_0)} \qquad \hat{\theta}-\theta_0 \approx -\frac{-\frac{2}{N}\int_0^T v(t)s'(t-\theta_0)dt}{\frac{2}{N}\int_0^T v(t)s''(t-\theta_0)dt} \quad (6)$$

respectively. The expected value of the numerator of the above formulas can be shown to be zero. For the general case, the expectation with respect to noise (the only stochastic entity appearing in the expression), denoted by $E_n[.]$, is

$$E_n[g'_\theta(v|\theta_0)] = \int_{\mathbf{R}^n} g'_\theta(v|\theta_0)p(v|\theta_0)d^n v = \frac{\partial}{\partial_\theta}\left[\int_{\mathbf{R}^n} p(v|\theta)d^n v\right]_{\theta=\theta_0} = 0 \quad (7)$$

which follows from $p'_\theta(v|\theta) = p(v|\theta)g'_\theta(v|\theta)$ (see the definition of g'_θ) and the *normalization property* of a pdf. The same result is of course obtained for the specialized case of the arrival time, where the last equality in the following equation is again justified when T is long enough to contain the whole duration of $s(.)$.

$$E_n\left[\frac{-2}{N}\int_0^T [s(t-\theta_0)+n(t)]s'(t-\theta_0)dt\right] =$$
$$= \frac{-2}{N}\int_0^T s(t-\theta_0)s'(t-\theta_0)dt = \frac{-1}{N}[s^2(t-\theta_0)]_0^T = 0$$

An important formula can be derived for the expected value of the denominators in (6), by carrying on the calculations of the expectation in (7): through some algebra, we have, in the general case

$$\frac{\partial}{\partial_\theta}\left[\int_{\mathbf{R}^n} g'_\theta(v|\theta)p(v|\theta)d^n v\right]_{\theta=\theta_0} =$$

$$\left[\int_{\mathbf{R}^n} g''_\theta(v|\theta)p(v|\theta)d^n v + \int_{\mathbf{R}^n} g'_\theta(v|\theta)p'(v|\theta)d^n v\right]_{\theta=\theta_0} = \qquad (9)$$

$$\left[\int_{\mathbf{R}^n} g''_\theta(v|\theta)p(v|\theta)d^n v + \int_{\mathbf{R}^n} [g'_\theta(v|\theta)]^2 p(v|\theta)d^n v\right]_{\theta=\theta_0} = 0$$

$$\Rightarrow \quad E_n[g''_\theta(v|\theta_0)] = -E_n[[g'_\theta(v|\theta_0)]^2]$$

with the last equation relating the first and second derivatives of the log-likelihood ratio. Applying the last result to the arrival time problem, we get, for the expectation of the denominator in (6):

$$E_n\left[\frac{2}{N}\int_0^T [s(t-\theta_0)+n(t)]s''(t-\theta_0)dt\right] = \frac{2}{N}\int_0^T s(t-\theta_0)s''(t-\theta_0)dt =$$

$$= -E_n\left[\left(\frac{2}{N}\right)^2\int_0^T\int_0^T [s(t-\theta_0)+n(t)]s'(t-\theta_0)[s(u-\theta_0)+n(u)]s'(u-\theta_0)dudt\right] =$$

$$= \frac{-2}{N}\int_0^T [s'(t-\theta_0)]^2 dt \qquad (10)$$

where the autocovariance of white noise has been used and the result of the expected value in (8) has been used to simplify some terms. Rewriting the expression of the estimation error for the arrival time in a different form

$$\hat{\theta}-\theta_0 \approx -\frac{-\frac{2}{N}\int_0^T n(t)s'(t-\theta_0)dt}{-\frac{2}{N}\int_0^T [s'(t-\theta_0)]^2 dt + \frac{2}{N}\int_0^T n(t)s''(t-\theta_0)dt}$$

it could be demonstrated that, for high signal to-noise ratios (SNR), the second term of the denominator is negligible, compared to the first; the first term representing the denominator expected value (see (10)), and the second representing its statistical fluctuations. The latter is a general result: in the expression of the estimator error, for the general estimation problem, the numerator always contains noise terms, while the denominator consists of a nonzero expected value plus noise terms that can be neglected, for high SNR. With this result, that is not evident from (6), we can rewrite the general expression for the estimator error in (6), considering, in the denominator, only its expected value:

$$\hat{\theta}-\theta_0 \approx -\frac{g'_\theta(v|\theta_0)}{E_n[g''_\theta(v|\theta_0)]} \qquad (11)$$

now, computing the bias, only the numerator expectation must be evaluated: since the latter has been shown to be zero, the estimator is unbaiased, under

our approximations. A more important result is obtained computing the estimator variance: using the expression (11), the result in (9) and some simplifications:

$$Var_n[\hat{\theta}(v)] = Var_n\left[-\frac{g'_\theta(v|\theta_0)}{E_n[g''_\theta(v|\theta_0)]}\right] = \frac{1}{E_n[[g'_\theta(v|\theta_0)]^2]}$$

The last expression, known as **Fisher's formula**, then relates the mean square error of an estimator to the variance of the first derivative of the log-likelihood ratio with respect to noise, this variance being computed on the true value of the estimandum.

Fisher's formula can be applied to the arrival time problem: having already computed $g'_\theta(v|\theta)$ and using (10) for its mean square value, we can write concisely

$$Var\left[\hat{\theta}(v)\right] = \frac{N}{2E\beta^2} \qquad d^2 \triangleq \frac{2E}{N} \qquad \beta^2 \triangleq \frac{\int_{-\infty}^{+\infty}[s'(t)]^2 dt}{\int_{-\infty}^{+\infty} s(t)^2 dt} = \frac{\int (2\pi f)^2 |S(f)|^2 df}{\int |S(f)|^2 df}$$

where T has been assumed sufficiently long, as to extend the integrals on the whole time axis, E is the energy of $s(t)$, d^2 is the signal-to-noise ratio, and the factor β^2 can be computed using well known properties of Fourier transforms. This latter factor is referred to as the **mean square bandwidth** of the signal $s(t)$: this name is justified by the last expression which gives an indication on the *dispersion* of the signal spectrum S(f) on the frequency axis.

As stated at the beginning of this section, we have used the conditional pdf of data $p(v|\theta)$ for our derivations; in general, the previously defined likelihood functional $\Lambda[v(t)|\theta]$ is appropriate, and all the results hold with $g[v(t)|\theta] = \ln\Lambda[v(t)|\theta]$ being the log-likelihood functional, computed on the observed signal.

The same approach followed here can be applied to the general case of multiple parameters estimation: we will omit the demonstrations but it can be shown that, for large signal-to-noise ratios, the ML estimates are unbiased and their covariance matrix can be computed, again, using the log-likelihood functional, according to

$$Cov\left[\hat{\theta}_j,\hat{\theta}_k\right]\overset{\Delta}{=}E\left[(\hat{\theta}_j-\theta_{j0})(\hat{\theta}_k-\theta_{k0})\right]=\left(\Gamma^{-1}\right)_{jk} \quad (12)$$

where the elements of the matrix Γ to be inverted are given by

$$\Gamma_{il}\overset{\Delta}{=}E_n\left[g'_{\theta_i}(v|\underline{\theta}_0)g'_{\theta_l}(v|\underline{\theta}_0)\right]$$

The matrix Γ is called the Fisher information matrix; we will further exploit its meaning for the case of white Gaussian noise, in the following section.

3.5 ML estimation in Gaussian noise: the ambiguity function

Recall that the likelihood functional, in the case of Gaussian noise, has the following expression

$$\Lambda[v(t)|\underline{\theta}]=\exp\left[\int_0^T v(t)q(t;\underline{\theta})dt-\frac{1}{2}\int_0^T s(t;\underline{\theta})q(t;\underline{\theta})dt\right]$$

where $q(t;\underline{\theta})$ is the function solving the Fredholm inhomogeneous equation

$$\int_0^T \phi(t;u)q(u;\underline{\theta})du=s(t;\underline{\theta})$$

Deriving the natural logarithm of Λ with respect to the i-th parameter, gives

$$g'_{\theta_i}[v(t)|\underline{\theta}]=\int_0^T n(t)\frac{\partial q(t;\underline{\theta})}{\partial\theta_i}dt+\text{constant terms}$$

where the constant terms are those not subject to stochastic variations. Thus, the covariance of the latter derivatives only depends on the noise process $n(t)$, and the elements of the Fisher information matrix can be computed by applying the definition of noise autocovariance function and the definition of q(.):

In the simple case of white noise, it has been seen that the function q(.) is proportional to the signal to be detected: for any value of the parameters set, $q(t;\underline{\theta})=(2/N)s(t;\underline{\theta})$ The Fisher information matrix elements can then be written as

$$\Gamma_{il}=\frac{\partial^2 H(\underline{\theta}_1,\underline{\theta}_2)}{\partial\theta_{1i}\partial\theta_{2l}}|_{\underline{\theta}_1=\underline{\theta}_2=\underline{\theta}_0} \quad (13)$$

$$\Gamma_{il}\overset{\Delta}{=}Cov\left[g'_{\theta_i}(v|\underline{\theta}_0)g'_{\theta_l}(v|\underline{\theta}_0)\right]=E_n\left[\int_0^T\int_0^T n(t)n(u)\frac{\partial q(u;\underline{\theta}_0)}{\partial\theta_i}\frac{\partial q(t;\underline{\theta}_0)}{\partial\theta_l}dudt\right]=$$

$$=\int_0^T\frac{\partial s(t;\underline{\theta}_0)}{\partial\theta_i}\frac{\partial q(t;\underline{\theta}_0)}{\partial\theta_l}dt$$

where the following function of the parameters, called the **ambiguity function**,

$$H(\underline{\theta}_1,\underline{\theta}_2)=\frac{2}{N}\int_0^T s(t-\underline{\theta}_1)s(t-\underline{\theta}_2)dt$$

has been introduced. The ambiguity function takes, besides the multiplying factor (2/N), the form of a correlation integral, over the interval [0,T], between the useful signal s(t), considered for two different values of the parameters set. Resorting to the interpretation of this integral, introduced in signal analysis, as a measure of the similitude between different signals, the ambiguity function gives us an indication on how much the signals $s(t;\theta_1)$ and $s(t;\theta_2)$ are similar on [0,T].

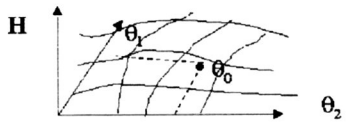

Figure 17. An example of the ambiguity function for a scalar parameter θ.

This interpretation of $H(\underline{\theta}_1,\ \underline{\theta}_2)$ has some important implications: referring to fig. 17, that represents a possible ambiguity function in the case of a scalar parameter θ, for the sake of visualization, and considering the function in the neighbourhood of $\underline{\theta}_1=\underline{\theta}_2=\underline{\theta}_0$, the more the function is flat around this point the less reliable the ML estimator will be. To demonstrate this result mathematically, consider that a flat function implies small partial derivatives, thus, according to (13), the Fisher information matrix elements are small; the inverse of this matrix has then large diagonal elements and, since, referring to (12), these represent the variance of the estimator elements, large variances result in a less reliable estimator, that is more subject to the fluctuations of its value induced by noise.

This important result could also be justified intuitively, again referring to the meaning of the correlation integral defining $H(\underline{\theta}_1,\ \underline{\theta}_2)$; if this function

is relatively flat around the estimanda value $\underline{\theta}_0$, then significant deviations of the parameters from the true estimanda value do not change the useful signal shape significantly, since the value of the correlation integral remains nearly the same. The receiver has thus difficulty in estimating the true $\underline{\theta}_0$ because the useful received signal $s(t;\underline{\theta}_0)$ and another signal $s(t;\underline{\theta}_1)$, with $\underline{\theta}_1 \neq \underline{\theta}_0$ but close to it, are relatively similar, thus mistakeable, having a large correlation. If, on the other hand, the ambiguity function is sharply peaked around $\underline{\theta}_1 = \underline{\theta}_2 = \underline{\theta}_0$, the ML estimates are reliable, but, referring to fig. 15, a bank of many matched filters is needed. In other words, since the signal $s(t;\underline{\theta})$ is very sensitive to small variations of the parameters, a dense sampling of the parameters space (implemented by the matched filter bank) is required in order to have at least one filter whose impulse response is nearly matched to the true useful signal $s(t;\underline{\theta}_0)$ that is to be detected.

References

1. C. W. Helstrom, Elements of Signal Detection & Estimation (chaps. 1,2,6), Prentice-Hall, 1995.
2. C. W. Helstrom, Statistical Theory of Signal Detection, Pergamon Press, 1968 (2^{nd} ed.).
3. H. L. Van Trees, Detection, Estimation and Modulation Theory, vol.1 – Malabar (Florida): Krieger; New York: Wiley, 1968 - 1971.