Detection and Discrimination

Introduction

The remaining sessions in the course find us, at long last, addressing the central issues:

> How do we distinguish between radar returns generated in the presence and absence of a target?

and

> Given a detection procedure, how do we characterise its effectiveness quantitatively?

In posing these questions more precisely we will draw heavily on the probabilistic concepts introduced in the last session; the mathematical toolkit established in the earlier sessions then helps us work out their answers. Today we will concentrate on issues of principle, while paying the price of our adopting simple, and occasionally unrealistic, models of the radar returns. Subsequent sessions will remedy this defect, when we discuss the K distribution model and its applications in some detail.

Some of the material we will cover today will probably be familiar to you already; other bits come fairly close to topics of active research. If we can identify the common principles underlying both the everyday and the erudite and use each to illuminate the other, then we will be getting somewhere. We also discuss Kalman filtering, whose fundamental principles are very similar to those of detection and estimation and provide a useful introduction to adaptive filtering. Some exercises are provided, that fill in details glossed over in the session and suggest useful extensions to the material we cover today.

Statistical models for probabilities of detection and false alarm

A radar system presents us with a signal **x**, displayed perhaps as a function of time or spatial co-ordinates. If this signal changes noticeably as a result of the presence of a target it should be possible to detect that target. Usually there is a significant increase in the signal in the presence of the target; this implies that we might well perform detections by setting a threshold and ascribing any signal in excess of our threshold to a detection. This procedure need not be fool-proof; mis-attributions of large values of **x** derived from the radar returns from the background (i.e. clutter) will give rise to false alarms. Obviously, if we set the threshold sufficiently high we will tend to avoid such false alarms, but only at the cost of missing some 'real' detections. Detection performance calculations attempt to quantify this trade off between detections and false alarms.

To make significant progress we must first characterise the signal **x** in the presence and absence of the target. A probabilistic description in terms of the pdf (probability density function) $P(x)$ of its value x is convenient and, given the complexity and uncertain nature of the processes that contribute to the radar return, is as much as we can justifiably hope for. The probability that **x** takes values x between x_1 and x_2 is given by

$$\int_{x_1}^{x_2} dx\, P(x) \qquad (1)$$

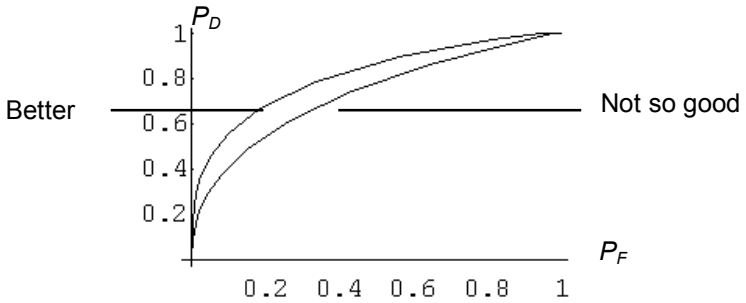
To be consistent with the usual properties of a probability this pdf must be positive and satisfy the normalisation condition

$$\int_{-\infty}^{\infty} dx P(x) = 1. \qquad (2)$$

We let P_A denote the pdf of the signal derived from the return from the ambient background; P_T will denote the pdf of the signal in the presence of the target. If we now take a threshold X we can express the probabilities of detection and false alarm as follows

$$P_D = \int_X^{\infty} dx P_T(x); \quad P_F = \int_X^{\infty} dx P_A(x) \qquad (3)$$

A plot of probability of detection vs. probability of false alarm, mapped out as the threshold X varies, is known as a Receiver Operation Characteristic (ROC) curve; ideally we would like a high probability of detection to be achieved at the expense of a small probability of false alarm. A couple of typical ROC curves are shown:



The ROC curve corresponding to an inability to distinguish between target and background is just a straight line of unit slope. (In many cases it is more convenient to plot ROC curves against log axes; this distorts their shape somewhat but the basic idea remains the same.)

Thus we see that the problem of detection can be regarded as that of deciding whether the pdf of x is better described by P_A or P_T, given a value of x. We have already argued that a simple thresholding on x will allow us to make this distinction with some measure of success; the question remains as to whether we can do better than this. We will now see that there are circumstances in which we can.

Likelihood ratios and optimal detection.

Let us consider the case where we wish to differentiate, on the basis of measurements of a random variable $\mathbf{z}$, between the two possible pdfs of its values z $P_{\mathbf{z},A}(z)$ and $P_{\mathbf{z},B}(z)$. Given a measurement z we assign this value, on the basis of some test, as yet unspecified, to the set of values Z_A corresponding to the former distribution or Z_B, corresponding to the latter. Between them Z_A and Z_B contain all possible values of z. On the basis of this classification we define a probability of detection (correct assignment to distribution A) and a probability of false alarm (incorrect assignment to distribution A)

as

$$P_{\mathrm{d}} = \int_{Z_A} dz P_{\mathbf{z},A}(z) \text{ and } P_{\mathrm{f}} = \int_{Z_A} dz P_{\mathbf{z},B}(z) \qquad (4)$$

We now define the optimum decision rule as that which maximises the probability of detection whilst maintaining the false alarm rate at a constant value α. This specification of optimality is known as the Neyman-Pearson criterion. To determine the test that satisfies this criterion we now consider the quantity

$$F = P_{\mathrm{d}} + \lambda(\alpha - P_f) \tag{5}$$

As the value of the false alarm rate is maintained at α we see that the maximisation of *F* corresponds to an equivalent maximisation of the probability of detection. We now introduce our explicit expressions for the probabilities of detection and false alarm to give

$$F = \lambda\alpha + \int_{Z_A} dz(P_{\mathbf{z},A}(z) - \lambda P_{\mathbf{z},B}(z)) \tag{6}$$

Thus we see that *F* will be maximised if the integral in this expression is carried out over the region where the integrand takes all its positive values. This allows us to identify the optimal decision rule as that in which an observation of *z* is identified as coming from $P_{\mathbf{z},A}(z)$ if, and only if,

$$\Lambda(z) = \frac{P_{\mathbf{z},A}(z)}{P_{\mathbf{z},B}(z)} > \lambda\,. \tag{7}$$

i.e. that the decision rule is based on the likelihood ratio $\Lambda(z)$. In practise it is frequently more convenient to carry out this thresholding on a monotonic function of the likelihood ratio such as its logarithm. The Lagrange multiplier λ can now be identified as the threshold on the likelihood ratio that establishes the given false alarm rate α. Thus, if $P(\Lambda)_{\Lambda,B}$ is the pdf of the likelihood ratio derived from a measurement of *z* drawn from distribution *B,* λ is defined implicitly by

$$\alpha = \int_{\lambda}^{\infty} d\Lambda\, P(\Lambda)_{\Lambda,B} \tag{8}$$

Here we have introduced the likelihood ratio test on the basis of the Neyman-Pearson criterion; exactly the same test emerges from a consideration of the so-called Bayes risk analysis. In this an intuitively reasonable cost function is constructed in terms of probabilities of detection and false alarm and the optimum test procedure is identified as that which minimises this quantity. Detailed discussions of this approach, which is algebraically more complex but is more readily extendible to the analysis of the testing of multiple, rather than binary, hypotheses, are given in the standard textbooks by van Trees (*'Detection, Estimation and Modulation Theory, Part 1'*, John Wiley, New York, 1968) and Middleton ('*An Introduction to Statistical Communication Theory'*, McGraw-Hill, New York, 1960). Here we merely stress that the two approaches lead to the same, optimal, test procedure .

Thus we have identified the optimum statistic with which to distinguish between signals drawn from two separate distributions. This procedure requires a detailed knowledge of the pdfs of these two distributions, which may not always be available in practice. Other, sub-optimal, test statistics are frequently used, chosen on the basis of convenience and general applicability.

Gaussian statistical models

The Gaussian distribution is a much used statistical model pdf that has the advantages of relative tractability and widespread validity. The former derives in part from the latter; a good model will be studied with sufficient vigour to ensure that it becomes tractable (e.g. through the study of the error function, which characterises the probabilities of detection and false alarm derived from the

Gaussian pdf). The wide applicability of the Gaussian model is a consequence of the central limit theorem, which shows that, subject to various conditions, the sum of a 'large' number of random variables has a Gaussian distribution. Thus the Maxwell distribution of velocities and the Green's function of the diffusion equation both have a characteristic Gaussian form; you might like to check these out for yourself. We will now use various Gaussian distributions to illustrate our discussion.

The simplest example is that of a single, 1-dimensional Gaussian random variable **x**, whose values *x* have a pdf

$$P(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-(x-m)^2/(2\sigma^2)\right). \tag{9}$$

We recall that this is characterised by two parameters, the mean *m* and the standard deviation σ, that the characteristic function of this distribution is

$$C(k) = \left\langle \exp(ikx) \right\rangle = \exp(ikm)\exp\left(-k^2\sigma^2/2\right) \tag{10}$$

that its moments have the factorisation property

$$\left\langle (x-m)^{2n} \right\rangle = \frac{(2n)!}{n!2^n} \left\langle (x-m)^2 \right\rangle^n \tag{11}$$

and that the probability of x exceeding a threshold *X* is given in terms of the error function:

$$\frac{1}{\sqrt{2\pi\sigma^2}} \int_X^\infty dx \exp\left(-(x-m)^2/(2\sigma^2)\right) = \frac{1}{2}\left(1 - \mathrm{erf}\left((X-m)/\sigma\sqrt{2}\right)\right) \tag{12}$$

where

$$\mathrm{erf}(z) = \frac{2}{\sqrt{\pi}} \int_0^z dt \exp(-t^2). \tag{13}$$

The *n*-dimensional generalisation of the Gaussian distribution, that includes the effects of correlation between its components, is fairly straightforward. In particular the characteristic function and moment results go through as you would expect; the analogue of the error function isn't so obvious, though we will consider things like this in the analysis of multi-channel observations later in the session. One particularly useful multivariate Gaussian process is the circular 'complex' Gaussian with two independent components of zero mean and equal variance. The I and Q components of a speckle/thermal noise signal provide an example of this Thus we have

$$P(E_I, E_Q) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{E_I^{\,2} + E_Q^{\,2}}{2\sigma^2}\right) \tag{14}$$

This can be expressed in terms of amplitude and phase variables

$$E = \sqrt{E_I^{\,2} + E_Q^{\,2}} \quad \theta = \tan^{-1}\left(\frac{E_Q}{E_I}\right) \tag{15}$$

so that

$$P(E,\theta) = \frac{E}{2\pi\sigma^2}\exp\left(-E^2/(2\sigma^2)\right). \tag{16}$$

The marginal pdfs of the phase and amplitude are

$$P(E) = \frac{E}{\sigma^2}\exp\left(-E^2/(2\sigma^2)\right); \quad P(\theta) = \frac{1}{2\pi}. \tag{17}$$

The intensity $i = E^2$ has a particularly simple pdf:

$$P(I) = \frac{1}{\langle I\rangle}\exp\left(-I/\langle I\rangle\right). \tag{18}$$

A simple performance calculation - a Swerling 2 target in thermal noise

Now let's do our first performance calculation. We model the target and background returns by complex Gaussians with different mean intensities:

$$P_a(I) = \frac{1}{\langle I_a\rangle}\exp\left(-I/\langle I_a\rangle\right) \quad P_t(I) = \frac{1}{\langle I_t\rangle}\exp\left(-I/\langle I_t\rangle\right) \tag{19}$$

(It seems sensible to assume that $\langle I_t\rangle > \langle I_a\rangle$; we are in fact modelling the clutter and target returns as complex Gaussian processes of different powers, which we add together to give a resultant complex Gaussian process. Here we are working with intensities and have 'thrown away' any phase-borne information in the signals.)

The log likelihood ratio takes the simple form

$$\Lambda = \log\left(\frac{\langle I_a\rangle}{\langle I_t\rangle}\right) + I\left(\frac{1}{\langle I_a\rangle} - \frac{1}{\langle I_t\rangle}\right) \tag{20}$$

In this case we see that thresholding on the intensity is the optimal detection procedure, as well as a convenient one. The expressions for the probabilities of detection and false alarm for a given threshold I_T are:

$$P_D = \exp\left(-I_T/\langle I_t\rangle\right) \quad P_F = \exp\left(-I_T/\langle I_a\rangle\right) \tag{21}$$

and are particularly simple in this case. If we have *N* independent measurements of the intensity the appropriate pdfs are

$$P_a(I) = \frac{1}{\langle I_a\rangle^N}\exp\left(-\sum_{k=1}^{N} I_k \Big/ \langle I_a\rangle\right) \quad P_t(I) = \frac{1}{\langle I_t\rangle^N}\exp\left(-\sum_{k=1}^{N} I_k \Big/ \langle I_t\rangle\right) \tag{22}$$

In this case we see that thresholding on the sum of the available intensities gives us the optimum detection procedure. To generate ROC curves we must now calculate the pdfs of such sums of intensities drawn from the target and background distributions. To do this efficiently we proceed

via the characteristic functions of the distributions of the single intensity measurements (we use the Laplace form as I is necessarily positive)

$$C(s) = \langle \exp(-Is) \rangle = \frac{1}{1 + s\langle I \rangle} \tag{23}$$

Thus the corresponding characteristic function of the pdf of the sum of N independent intensities is given by (why?)

$$C_N(s) = \langle \exp(-Is) \rangle^N = \frac{1}{\left(1 + s\langle I \rangle\right)^N} \tag{24}$$

this leads to following pdf for the sum of intensities

$$\begin{aligned} P(I) &= \frac{1}{2\pi i} \int_{\gamma - j\infty}^{\gamma + j\infty} ds \frac{\exp(sI)}{\left(1 + s\langle I \rangle\right)^N} \\ &= \frac{I^{N-1}}{(N-1)!\langle I \rangle^N} \exp\left(-I/\langle I \rangle\right) \end{aligned} \tag{25}$$

You can use this result to calculate ROC curves in example 1.

The foregoing example is not entirely trivial; it provides a model for the detection of a rapidly fluctuating target in thermal noise and as such is quite useful or the assessment of performance of the small target detection.

Swerling 0 and the Rice distribution.

We can also carry out performance calculations in which the target signal does not fluctuate (Swerling 0 model). Thus we can represent the vector of I and Q components of the received signal as $\mathbf{E} = \mathbf{A} + \mathbf{n}$ where $\mathbf{n}$ is the thermal noise process. As the noise process is isotropic we can chose the signal vector $\mathbf{A}$ to define the Q direction in the co-ordinate system in which we perform our integrations. Thus we have:

$$\begin{aligned} C(s) &= \left\langle \exp(-sE^2 \right\rangle \\ &= \frac{1}{\pi\langle I \rangle} \int_{-\infty}^{\infty} dE_I \int_{-\infty}^{\infty} dE_Q \exp\left(-s({E_I}^2 + {E_Q}^2)\right) \exp\left(-\frac{{E_I}^2}{\langle I \rangle}\right) \exp\left(-\frac{(E_Q - A)^2}{\langle I \rangle}\right) \\ &= \frac{1}{(1 + s\langle I \rangle)} \exp\left(-\frac{A^2 s}{(1 + s\langle I \rangle)}\right) \end{aligned} \tag{26}$$

(Prove it if you want.)

To find the pdf of the intensity of the process we Laplace invert this expression

$$P(I) = \frac{1}{2\pi i} \int_{\gamma - j\infty}^{\gamma + j\infty} ds \frac{\exp(sI)}{(1 + s\langle I \rangle)} \exp\left(-\frac{sA^2}{1 + s\langle I \rangle)}\right) \tag{27}$$

a little bit of algebraic manipulation allows us to recast this as:

$$P(I) = \frac{1}{\langle I \rangle} \exp\left(-\frac{I + A^2}{\langle I \rangle} \right) \frac{1}{2\pi i} \int_{\gamma - j\infty}^{\gamma + j\infty} \frac{dp}{p} \exp\left(p + \frac{A^2 I}{\langle I \rangle^2 p} \right) \tag{28}$$

We now expand $\exp\left(A^2 I / (\langle I \rangle^2 p) \right)$ in a series and invert the Laplace transform term by term, noting that

$$\frac{1}{2\pi i} \int_{\gamma - j\infty}^{\gamma + j\infty} \frac{dp}{p^{n+1}} \exp(p) = \frac{1}{n!} \tag{29}$$

This leads us to

$$\begin{aligned} P(I) &= \frac{1}{\langle I \rangle} \exp\left(-\frac{I + A^2}{\langle I \rangle} \right) \sum_{n=0}^{\infty} \left(\frac{A^2 I}{2\langle I \rangle} \right)^n \frac{1}{(n!)^2} \\ &= \frac{1}{\langle I \rangle} \exp\left(-\frac{I + A^2}{\langle I \rangle} \right) I_0\left(\frac{2A\sqrt{I}}{\langle I \rangle} \right) \end{aligned} \tag{30}$$

The modified Bessel function I has been identified from its series representation. To get the characteristic function of the pdf of N incoherently summed intensities we merely raise our earlier result to the appropriate power; the pdf is again obtained by Laplace inversion:

$$\begin{aligned} P(I)_N &= \frac{1}{2\pi i} \int_{\gamma - j\infty}^{\gamma + j\infty} ds \, \frac{1}{(1 + s\langle I \rangle)^N} \exp\left(sI - \frac{NA^2 s}{1 + s\langle I \rangle} \right) \\ &= \frac{\exp\left(-\frac{I + NA^2}{\langle I \rangle} \right)}{\langle I \rangle} \left[\frac{I}{NA^2} \right]^{\frac{N-1}{2}} I_{N-1}\left(\frac{2A\sqrt{NI}}{\langle I \rangle} \right) \end{aligned} \tag{31}$$

This result shows how a steady signal is ‘brought out’ of clutter by incoherent averaging. Filling in the details is part of exercise 1.

To carry out the optimum detection of a steady signal in thermal noise we would have to use the likelihood ratio derived from the exponential and Rice distributions:

$$\Lambda = \frac{\exp\left(-\left(I + A^2 \right) / \langle I \rangle \right) I_0\left(2A\sqrt{I} / \langle I \rangle \right)}{\exp\left(-I / \langle I \rangle \right)} = \exp\left(-A^2 / \langle I \rangle \right) I_0\left(2A\sqrt{I} / \langle I \rangle \right) \tag{32}$$

In principle, one could use this exact form for the likelihood ratio; the presence of the modified Bessel function makes this computationally inconvenient in practice. The modified Bessel function can be approximated in the limits where the signal is very large and very small compared with

$\langle I \rangle$; our labours in the previous sessions tell us how to do this. Thus in the large signal limit we have

$$\log \Lambda \approx -\frac{A^2}{\langle I \rangle} + \frac{2A\sqrt{I}}{\langle I \rangle} \tag{33}$$

so that thresholding on the received voltage (i.e. $\sqrt{I}$) approximates to the optimum procedure; such an instrument would be referred to as a linear detector. In the small signal limit

$$\Lambda \approx 1 + \left(I - \langle I \rangle\right) \frac{A^2}{\langle I \rangle^2} \tag{34}$$

and thresholding on the intensity *I* (or the square of the received voltage) is the thing to do. In the early days of radar much was made of this distinction though the performance of the two detectors over the whole range of signal to noise ratios does not differ at all appreciably.

Generalised likelihood ratios

So far we have been able to deduce useful detection procedures from model clutter and target plus clutter pdfs that contain parameters (e.g. *A*, $\langle I \rangle$), without specifying what these are. In more complicated cases we have to be more careful. Thus we might write our likelihood ratio as

$$\Lambda = \frac{P_t(x|\{b_1\})}{P_a(x|\{b_0\})}. \tag{35}$$

Here P_t and P_a are the pdfs of the 'target and clutter' and 'clutter' signals respectively; $\{b_1\}$ and $\{b_0\}$ are sets of parameters that characterise P_t and P_a. If we have full prior knowledge of $\{b_1\}$ and $\{b_0\}$ we can make our decision on the basis of a single measurement of the signal by forming the likelihood ratio Λ and comparing it with a threshold *T*. If Λ exceeds this threshold we ascribe the signal to 'target and clutter', otherwise it is ascribed to 'clutter' alone; the size of *T* chosen for the test determines the probability of false alarm for the decision process. If, however, we do not know $\{b_1\}$ and $\{b_0\}$ *a priori* we must first estimate these parameters from the received signals and then, using these estimates, form the appropriate likelihood ratio. We can now base our detection decision on this quantity. These estimates can be derived from a set of signals $\{x_i\}$ on the basis of the likelihood maximisation criterion, which is in turn made credible by Bayes theorem. Thus, for that given set of signals, we find the values of the parameters $\{b_1\}$ and $\{b_0\}$ that maximise the values of the 'target and clutter' and 'clutter' multivariate pdfs respectively. From Bayes theorem these correspond the most likely model parameters, given the set of signals $\{x_i\}$. If we have *N* independent signals $\{x_i\}$, *i* = 1, ..., *N*, the appropriate multivariate pdfs are

$$\begin{aligned} P_t^{(N)}\left(\{x_i\}|\{b_1\}\right) &= \prod_{i=1}^{N} P_t\left(x_i|\{b_1\}\right) \\ P_a^{(N)}\left(\{x_i\}|\{b_0\}\right) &= \prod_{i=1}^{N} P_a\left(x_i|\{b_0\}\right) \end{aligned} \tag{36}$$

We now estimate the parameters $\{b_1\}$ and $\{b_0\}$, of which there are M_1 and M_0 respectively, from the equations

$$\begin{aligned}\frac{\partial \log\left(P_t^{(N)}\left(\{x_i\}|\{b_1\}\right)\right)}{\partial b_{1,k}} &= 0; \quad k = 1 \cdots M_1 \\ \frac{\partial \log\left(P_a^{(N)}\left(\{x_i\}|\{b_0\}\right)\right)}{\partial b_{0,k}} &= 0; \quad k = 1 \cdots M_0\end{aligned} \tag{37}$$

These estimates, which we denote by $\left\{\hat{b}_1\right\}$ and $\left\{\hat{b}_0\right\}$, are employed to construct the N independent signal likelihood ratio

$$\hat{L} = \frac{\prod_{i=1}^{N} P_t\left(x_i|\left\{\hat{b}_1\right\}\right)}{\prod_{i=1}^{N} P_a\left(x_i|\left\{\hat{b}_0\right\}\right)}. \tag{38}$$

We can then use this quantity, or a suitable approximation to it, as the basis of a detection procedure. This approach has proved to be very effective; most of its current applications are classified.

A simple example

To illustrate these principles we will consider a very simple detection problem, essentially that of distinguishing between independent Gaussian random variables drawn from a distribution with a zero mean and given variance, and from a distribution with the same variance, but having a non-zero mean. This elementary problem nonetheless highlights many of the principles exploited in currently used small maritime target detection algorithms.

The joint pdf of *N* independent samples drawn form the zero mean Gaussian distribution is

$$P_1\left(\{x_k\}\right) = \frac{1}{\left(2\pi\sigma^2\right)^{\frac{N}{2}}} \exp\left(-\frac{1}{2\sigma^2}\sum_{k=1}^{N} x_k^2\right) \tag{39}$$

the corresponding pdf for samples drawn from the non-zero mean distribution is

$$P_2\left(\{x_k\}\right) = \frac{1}{\left(2\pi\sigma^2\right)^{\frac{N}{2}}} \exp\left(-\frac{1}{2\sigma^2}\sum_{k=1}^{N} \left(x_k - m\right)^2\right) \tag{40}$$

If we know values of the parameters defining these distributions we can form the likelihood ratio

$$\frac{P_1\left(\{x_k\}\right)}{P_2\left(\{x_k\}\right)} \tag{41}$$

and identify the sufficient statistic

$$\lambda = \sum_{k=1}^{N} x_k \tag{42}$$

as the optimum discriminant in this case. This is a sum of Gaussian random variables and so is itself a Gaussian random variable. Its pdfs, when constructed from zero mean and non-zero mean Gaussians, are

$$\begin{aligned} P(\lambda) &= \frac{1}{\sqrt{2N\pi\sigma^2}} \exp\left(-\frac{\lambda^2}{2N\sigma^2} \right) \\ P(\lambda) &= \frac{1}{\sqrt{2N\pi\sigma^2}} \exp\left(-\frac{(\lambda - mN)^2}{2N\sigma^2} \right) \end{aligned} \tag{43}$$

Thus, for a given threshold Λ we have the probabilities of detection and false alarm (correct assignment to the non-zero mean class, incorrect assignment to the zero mean class) given by

$$\begin{aligned} P_D(\Lambda) &= \frac{1}{2} \operatorname{erfc}\left(\frac{\Lambda - Nm}{\sqrt{2N\sigma^2}} \right) \\ P_{FA}(\Lambda) &= \frac{1}{2} \operatorname{erfc}\left(\frac{\Lambda}{\sqrt{2N\sigma^2}} \right) \end{aligned} \tag{44}$$

Using these results it is possible to trace out the ROC curves (probability of detection vs. probability of false alarm) characterisng the performance of this simple 'detector'.

So far we have assumed that we know the parameters σ, m. If however we assume that one set of samples is drawn from a zero mean distribution and the other from a non-zero mean distribution, neither of whose variances we know, we cannot carry through the foregoing analysis. Instead we have to adopt the so-called generalised likelihood ratio approach, in which the data provide us with estimates to be incorporated into the discriminant. In the zero mean case we have a likelihood of the form

$$P_1(\{x_k\}) = \frac{1}{\left(2\pi\,\sigma_1^2\right)^{\frac{N}{2}}} \exp\left(-\frac{1}{2\sigma_1^2} \sum_{k=1}^{N} x_k^2 \right) \tag{45}$$

Given the data $\{x_k\}$ we can we estimate σ_1^2 as that value which maximises this likelihood; thus we find that

$$\hat{\sigma}_1^2 = \frac{1}{N} \sum_{k=1}^{N} x_k^2 \equiv \left\langle x^2 \right\rangle \tag{46}$$

If we assume that the data are drawn from the non-zero mean distribution with the pdf

$$P_2(\{x_k\}) = \frac{1}{\left(2\pi\sigma_2^2\right)^{\frac{N}{2}}} \exp\left(-\frac{1}{2\sigma_2^2} \sum_{k=1}^{N} (x_k - m)^2 \right) \tag{47}$$

The mean and variance parameters can then be estimated by likelihood maximisation as

$$\begin{aligned} \hat{m} &= \frac{1}{N} \sum_{k=1}^{N} x_k \equiv \langle x \rangle \\ \hat{\sigma}_2^2 &= \frac{1}{N} \sum_{k=1}^{N} (x_k - \hat{m})^2 = \langle x^2 \rangle - \langle x \rangle^2 \end{aligned} \tag{48}$$

If these parameters are now introduced into the likelihood ratio (41) we find that the quantity

$$\chi = \frac{\langle x^2 \rangle}{\langle x^2 \rangle - \langle x \rangle^2} \tag{49}$$

emerges as a discriminant with which we can distinguish between the zero and non-zero mean distributions (speaking loosely, χ will tend to take larger values in the latter case, especially when the mean is significantly bigger than the variance). To investigate the behaviour of χ more fully we first note that

$$\begin{aligned} \chi &= 1 + \frac{\bar{x}^2}{s^2} \\ \bar{x} &= \langle x \rangle, \quad s^2 = \langle x^2 \rangle - \langle x \rangle^2 \end{aligned} \tag{50}$$

where $\bar{x}$, s^2 provide estimators of the mean and variance of the distribution form which the $\{x_k\}$ are drawn. These estimators are themselves random variables; it can be shown that their joint pdf takes the form

$$P(\bar{x}, s) = 2\sqrt{\frac{1}{\pi}} \left(\frac{n}{2\sigma^2} \right)^{\frac{n}{2}} \frac{1}{\Gamma((n-1)/2)} \exp\left(-\frac{1}{2\sigma^2} \left(ns^2 + n(\bar{x} - m)^2 \right) \right) s^{n-2}; \quad -\infty < \bar{x} < \infty, \quad 0 \le s < \infty \tag{51}$$

Here σ, m are the root variance and mean of the underlying Gaussian distribution; $\Gamma(z)$ is the gamma function. We now consider the random variable

$$t = \frac{\bar{x}}{s}, \tag{52}$$

in terms of which our discriminant takes the form

$$\chi = 1 + t^2 \tag{53}$$

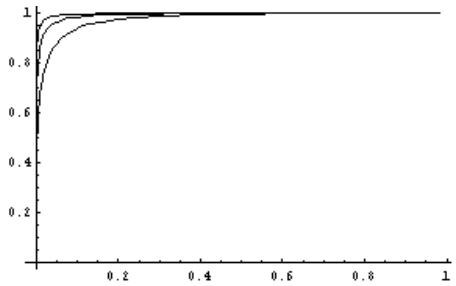
A simple change in variables yields the joint pdf of *s* and *t*:

$$P(t,s)=\frac{2}{\Gamma((n-1)/2)\Gamma(1/2)}\left(\frac{n}{2\sigma^2}\right)^{\frac{n}{2}}s^{n-1}\exp\left(-\frac{n}{2\sigma^2}\left(\left(1+t^2\right)s^2+m^2-2mst\right)\right);\quad -\infty<t<\infty,\ 0\le s<\infty \tag{54}$$

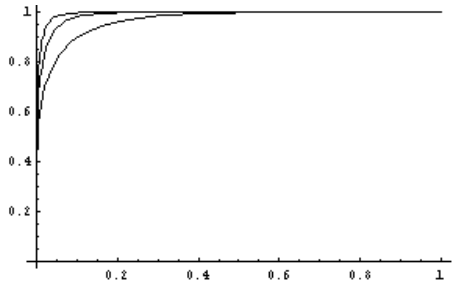
from which the marginal pdf and cumulative distribution of t can be obtained by integration. Thus we can write the probability that the discriminant exceeds a threshold

$$\text{Prob}(\chi>\Lambda)=\frac{1}{\Gamma\left(\frac{n-1}{2}\right)}\left(\frac{n}{2\sigma^2}\right)^{\frac{n-1}{2}}\int_0^\infty dss^{n-2}\exp\left(-\frac{ns^2}{2\sigma^2}\right)\left(2-\text{erf}\left(\sqrt{\frac{n}{2\sigma^2}}\left(s\sqrt{(\Lambda-1)}-m\right)\right)-\text{erf}\left(\sqrt{\frac{n}{2\sigma^2}}\left(s\sqrt{(\Lambda-1)}+m\right)\right)\right) \tag{55}$$

This then allows us to plot out the ROC curves for detections based on the likelihood and generalised likelihood ratios; the former is seen to work rather better than the latter, as we might expect. In fact we need to process 10 or 15 samples in the latter to get performance comparable with that obtained from 3 or 4 samples in the former.



ROC curves for likelihood ratio detection, based on 2,3 and 4 samples. $\sigma=1.0,\quad m=0,2.0$



ROC curves for generalised likelihood ratio detection, based on 5,10 and 15 samples.
$\sigma=1.0,\quad m=0,2.0$

Principles of Kalman filtering

In many situations we are presented with an incoming stream of measurements, from which we wish to discern some underlying process. As our knowledge of this process is at best likely to be statistical we are faced with the problem of estimating values of a random process from a series of measurements. Detailed knowledge of the measured process is required if an optimum solution to the problem is to be achieved; standard methods are predicated on a knowledge of the modeled system's correlation function or power spectrum. Furthermore it is frequently assumed that the process is stationary and that an arbitrarily long record is available; the standard Weiner filter is derived on the basis of these assumptions. In many circumstances these conditions are not satisfied; we must undertake the sequential analysis of a stream of data whose correlation properties are not known and need not be stationary. Fortunately it is possible to undertake such an analysis, using the so-called Kalman filter. This assumes a Langevin type model for the underlying process similar to those discussed in the previous section, and takes explicit account of the changes in our prior knowledge that occur as we process a sequence of data. If the underlying process is stationary and Gaussian, simple likelihood maximization estimation leads us to a recursive filtering algorithm; this can be extended to non-Gaussian and non-stationary processes for which it satisfies a minimum variance criterion, if not that of absolute optimality. We will now develop the principles of Kalman filtering, first from the standpoint of a stationary Gaussian model, then from that of least squares fitting and the associated innovations sequence. Details of some potentially unfamiliar algebraic manipulations are included; it is hoped that these will render the matrix inversion lemma and its use in the generation of formal identities involving various covariance matrices less mysterious than might otherwise be the case. (We have already looked at these manipulations briefly in the session on vectors and matrices.) While the density of equations does increase rather alarmingly from now on, the basic ideas are relatively simple.

In the previous session we saw how a statistically varying quantity can be modeled by a differential equation, driven by a white Gaussian noise process; when cast in the form of a set of coupled first order differential equations this can be solved, formally and numerically, to generate a sequence of values taken by the process at discrete time intervals. Measurements of the statistical process can be represented in much the same way; values of the process $\mathbf{x}(t)$ yielding measurements $\mathbf{y}(t)$ through

$$\mathbf{y}(t) = \mathbf{C}(t) \cdot \mathbf{x}(t) + \mathbf{n}(t)$$
$$\mathbf{n}(t)\mathbf{n}(t')^T = \delta(t - t')\mathbf{Q}(t)$$

The vectors **x,y** need not be of the same dimension (and the matrix **C** need not be square) while the noise values **n** are assumed to be uncorrelated. Taken in conjunction with a statistical model of the underlying physical process, (56) provides us with a model of the measured output of the system that is suitable for use in the construction of a Kalman filter.

Before we consider Kalman filtering in any detail we look at a simpler problem that shares many of its salient features. This is the problem of the sequential estimation of a quantity **a** from a sequence of measurements **r**. We know that **a** is a random variable with a known Gaussian distribution; beyond that we have no prior knowledge of its value. A sequence of values of **r** is available, which we use to successively refine out estimate of **a** (which remains constant throughout the whole process) producing an ever-narrower Gaussian distribution that captures our current state of knowledge of **a.** Thus we have our prior knowledge expressed through

$$P(\mathbf{a}) = \frac{1}{(2\pi)^{n/2}\left|\det \mathbf{\Lambda_a}\right|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{a} - \mathbf{m}_0)^T \cdot \mathbf{\Lambda}_\mathbf{a}^{-1} \cdot (\mathbf{a} - \mathbf{m}_0)\right) \qquad (57)$$

Our estimate of **a** based on this would be $\mathbf{m}_0$ and maximizes the likelihood (57); $\mathbf{\Lambda_a}$ would give us an indication of the accuracy of this estimate. We are now provided with a measurement

$$\mathbf{r}_1 = \mathbf{C}\cdot\mathbf{a} + \mathbf{w}_1 \tag{58}$$

The noise process **w** is a zero mean Gaussian vector process with a covariance matrix $\mathbf{\Lambda_w}$. Thus we can identify the conditional probability

$$P(\mathbf{r}_1 \mid \mathbf{a}) = \frac{1}{(2\pi)^{m/2}\left|\mathbf{\Lambda_w}\right|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{r}_1 - \mathbf{C}\cdot\mathbf{a})^T \cdot \mathbf{\Lambda_w^{-1}} \cdot (\mathbf{r}_1 - \mathbf{C}\cdot\mathbf{a})\right) \tag{59}$$

Bayes' theorem now allows us to express the likelihood of the value taken by **a**, conditioned on this first observation

$$P(\mathbf{a} \mid \mathbf{r}_1) = \frac{P(\mathbf{r}_1 \mid \mathbf{a})P(\mathbf{a})}{P(\mathbf{r}_1)} \tag{60}$$

If we substitute (57) and (59) into this we can determine our improved estimate of **a** as that which maximizes the likelihood $P(\mathbf{a} \mid \mathbf{r}_1)$. This calculation is easy to carry through in this case, as all the likelihoods involved are Gaussian; the estimate and its associated variance follow from a simple completion of the square in the exponent. Thus the width of the Gaussian distribution of the values of **a**, taking account of the single measurement $\mathbf{r}_1$, is given by

$$\mathbf{\Lambda}_{\varepsilon_1}^{-1} = \mathbf{\Lambda_a^{-1}} + \mathbf{C}^T \cdot \mathbf{\Lambda_w^{-1}} \cdot \mathbf{C} \tag{61}$$

and can be seen to decrease, while its mean, which corresponds to our improved estimate of **a,** is

$$\begin{aligned} \hat{\mathbf{a}}_1 &= \mathbf{\Lambda}_{\varepsilon_1} \cdot \left(\mathbf{C}\cdot\mathbf{\Lambda_w^{-1}}\cdot\mathbf{r}_1 + \mathbf{\Lambda_a^{-1}}\cdot\mathbf{m}_0\right) \\ &= \mathbf{m}_0 + \mathbf{\Lambda}_{\varepsilon_1}\cdot\mathbf{C}^T\cdot\mathbf{\Lambda_w^{-1}}\cdot(\mathbf{r}_1 - \mathbf{C}\cdot\mathbf{m}_0) \end{aligned} \tag{62}$$

If we now receive another measurement $\mathbf{r}_2$ this can be used to refine the estimate further, $P(\mathbf{a})$ in the forgoing analysis now being replaced by $P(\mathbf{a} \mid \mathbf{r}_1)$. This process can be repeated, using each successive measurement, resulting in an updated estimate of **a** and the variance characterising the quality of that estimate. In an obvious notation we have

$$\mathbf{\Lambda}_{\varepsilon_n}^{-1} = \mathbf{\Lambda}_{\varepsilon_{n-1}}^{-1} + \mathbf{C}^T \cdot \mathbf{\Lambda_w^{-1}} \cdot \mathbf{C} \tag{63}$$

$$\hat{\mathbf{a}}_n = \hat{\mathbf{a}}_{n-1} + \mathbf{\Lambda}_{\varepsilon_n} \cdot \mathbf{C}^T \cdot \mathbf{\Lambda_w^{-1}} \cdot \left(\mathbf{r}_n - \mathbf{C}\cdot\hat{\mathbf{a}}_{n-1}\right) \tag{64}$$

We see the error in the estimate becoming progressively smaller and the estimate being modified with each successive piece of data as it is made available. Should there be a large measurement noise, described by $\mathbf{\Lambda_w}$, then relatively little 'attention' is paid to the incoming data; furthermore later arrivals have less effect on the estimate as $\mathbf{\Lambda}_{\varepsilon_n}$ gets progressively smaller with increasing *n.* Both these observations are quite sensible, and in accord with our intuition. The quantity $\left(\mathbf{r}_n - \mathbf{C}\cdot\hat{\mathbf{a}}_{n-1}\right)$ is an innovation, i.e. the difference between the current measurement and that based on the current estimate, and isolates the 'new' information in the current measurement.

This is multiplied by a 'gain' that embodies our confidence in the veracity of this innovation, and added to the fruits of our previous labours. The whole thing displays a characteristic 'predictor-corrector' form.

Thus far we have used a sequence of measurements to improve our estimate of a single, constant quantity. To make further progress we now let the measured quantity evolve in time, in parallel with the measurement process. Initially we assume that this time evolution is characterized by a matrix $\mathbf{\Phi}$ that is constant. The measurement matrix is also taken to be unvarying. (We will see how to relax these constraints when we come to formulate the filtering process in terms of least squares fitting and innovations sequences. For the present we assume that all the processes involved are Gaussian so that the estimation equations can be sorted out almost by inspection.)

$$\begin{aligned}\mathbf{x}(n+1) &= \mathbf{\Phi}\cdot\mathbf{x}(n)+\mathbf{v}(n)\\ \mathbf{v}(n)\mathbf{v}(n')^T &= \delta_{n,n'}\mathbf{Q}_1\end{aligned} \tag{65}$$

$$\begin{aligned}\mathbf{y}(n) &= \mathbf{C}\cdot\mathbf{x}(n)+\mathbf{w}(n)\\ \mathbf{w}(n)\mathbf{w}(n')^T &= \delta_{n,n'}\mathbf{Q}_2\end{aligned} \tag{66}$$

Our assumed prior knowledge is that quantity **x** has a normal distribution, characterized by a mean $\mathbf{x}(0)$ and a variance $\mathbf{K}(0)$. We now receive our first measurement $\mathbf{y}(1)$, from which we estimate the corresponding $\mathbf{x}(1)$. Arguing just as for the first step of the sequential estimation procedure we form the estimator $\hat{\mathbf{x}}(1)$, which is the mean of a normal conditional distribution $P(\mathbf{x}(1)\,|\,\mathbf{y}(1))$ whose covariance is $\mathbf{K}(1)$

$$\begin{aligned}\hat{\mathbf{x}}(1) &= \mathbf{x}(0)+\mathbf{K}(1)\cdot\mathbf{C}^T\cdot\mathbf{Q}_2^{-1}(\mathbf{y}(1)-\mathbf{C}\cdot\mathbf{x}(0))\\ \mathbf{K}(1)^{-1} &= \mathbf{K}(0)^{-1}+\mathbf{C}^T\cdot\mathbf{Q}_2^{-1}\cdot\mathbf{C}\end{aligned} \tag{67}$$

This familiar procedure must be modified when we receive the next measurement $\mathbf{y}(2)$. We wish to use this to estimate the corresponding value of the underlying process $\mathbf{x}(2)$, recognizing that this is different from $\mathbf{x}(1)$. In the sequential estimation problem we used Bayes' theorem to combine our prior knowledge of the measurement process and the previous estimate to obtain a new estimate. Here we have to accommodate our understanding of the measurement process and the evolution of the underlying **x** process. To this end we write

$$P(\mathbf{x}(2),\mathbf{y}(1),\mathbf{y}(2)) = P(\mathbf{x}(2)\,|\,\mathbf{y}(1),\mathbf{y}(2))P(\mathbf{y}(1),\mathbf{y}(2)) = P(\mathbf{x}(2)\,|\,\mathbf{y}(1),\mathbf{y}(2))P(\mathbf{y}(2)\,|\,\mathbf{y}(1))P(\mathbf{y}(1)) \tag{68}$$

and

$$P(\mathbf{x}(2),\mathbf{y}(1),\mathbf{y}(2)) = P(\mathbf{y}(2)\,|\,\mathbf{x}(2),\mathbf{y}(1))P(\mathbf{y}(1),\mathbf{x}(2)) = P(\mathbf{y}(2)\,|\,\mathbf{x}(2),\mathbf{y}(1))P(\mathbf{x}(2)\,|\,\mathbf{y}(1))P(\mathbf{y}(1)) \tag{69}$$

If we equate these two expressions we find that the required conditional probability density of the values $\mathbf{x}(2)$, given the values of $\mathbf{y}(1)$ and $\mathbf{y}(2)$, is given by

$$P(\mathbf{x}(2)\,|\,\mathbf{y}(1),\mathbf{y}(2)) = \frac{P(\mathbf{y}(2)\,|\,\mathbf{x}(2),\mathbf{y}(1))P(\mathbf{x}(2)\,|\,\mathbf{y}(1))}{P(\mathbf{y}(2)\,|\,\mathbf{y}(1))} \tag{70}$$

The model of the measurement process tells us that

$$P(\mathbf{y}(2) \mid \mathbf{x}(2), \mathbf{y}(1)) = \frac{1}{(2\pi)^{m/2}\sqrt{|\det \mathbf{Q}_2|}} \exp\left(-\frac{1}{2}(\mathbf{y}(2) - \mathbf{C}\cdot\mathbf{x}(2))^T \cdot \mathbf{Q}_2^{-1} \cdot (\mathbf{y}(2) - \mathbf{C}\cdot\mathbf{x}(2))\right), \tag{71}$$

which shows no explicit dependence on $\mathbf{y}(1)$ and can be written as $P(\mathbf{y}(2) \mid \mathbf{x}(2))$. Finally we need to construct $P(\mathbf{x}(2) \mid \mathbf{y}(1))$; it is at this point that we introduce our model (65) for the underlying process $\mathbf{x}$. We see from the linearity of the model that $\mathbf{x}(2)$ is a normal process, whose distribution is determined solely by its mean and variance. Of these the former is given by $\mathbf{\Phi}\cdot\hat{\mathbf{x}}(1)$; the latter is a sum of two terms, one derived from the measurement noise and the other from $\mathbf{K}(1)$, the variance of our estimate of $\mathbf{x}(1)$. This can be written as

$$\mathbf{K}(2 \mid 1) = \mathbf{Q}_1 + \mathbf{\Phi}\cdot\mathbf{K}(1)\cdot\mathbf{\Phi}^T \tag{72}$$

so that

$$P(\mathbf{x}(2) \mid \mathbf{y}(1)) = \frac{1}{(2\pi)^{n/2}\sqrt{|\det \mathbf{K}(2 \mid 1)|}} \exp\left(-\frac{1}{2}(\mathbf{x}(2) - \mathbf{\Phi}\cdot\hat{\mathbf{x}}(1))^T \cdot \mathbf{K}(2 \mid 1)^{-1} \cdot (\mathbf{x}(2) - \mathbf{\Phi}\cdot\hat{\mathbf{x}}(1))\right) \tag{73}$$

These results can be combined to give

$$P(\mathbf{x}(2) \mid \mathbf{y}(1), \mathbf{y}(2)) \propto \exp\left(-\frac{1}{2}(\mathbf{x}(2) - \mathbf{\Phi}\cdot\hat{\mathbf{x}}(1))^T \cdot \mathbf{K}(2 \mid 1)^{-1} \cdot (\mathbf{x}(2) - \mathbf{\Phi}\cdot\hat{\mathbf{x}}(1)) - \frac{1}{2}(\mathbf{y}(2) - \mathbf{C}\cdot\mathbf{x}(2))^T \cdot \mathbf{Q}_2^{-1} \cdot (\mathbf{y}(2) - \mathbf{C}\cdot\mathbf{x}(2))\right) \tag{74}$$

Using this result we can now identify the optimum estimate of $\mathbf{x}(2)$ and the variance $\mathbf{K}(2)$ of the associated distribution; completing the square is sufficient in this Gaussian case. Thus we find that

$$\begin{aligned} \hat{\mathbf{x}}(2) &= \mathbf{\Phi}\cdot\hat{\mathbf{x}}(1) + \mathbf{K}(2)\cdot\mathbf{C}^T \cdot \mathbf{Q}_2^{-1} \cdot (\mathbf{y}(2) - \mathbf{C}\cdot\mathbf{\Phi}\cdot\hat{\mathbf{x}}(1)) \\ \mathbf{K}(2)^{-1} &= \mathbf{K}(2 \mid 1)^{-1} + \mathbf{C}^T \cdot \mathbf{Q}_2^{-1} \cdot \mathbf{C} \end{aligned} \tag{75}$$

Once again we see that the optimum estimate takes the 'predictor-corrector' form. The Kalman gain takes the form of the 'quotient' of the variances of the current estimate and the measurement noise and weights the innovation formed from the incoming measurement, the previous estimate and our knowledge of the underlying process **x**.

The expressions we have just derived can be transformed using the so-called matrix inversion lemma, which crops up in many situations where we might wish to manipulate covariance matrices and their inverses. Thus we note that, for two co-dimensional square matrices **A** and **B**,

$$\mathbf{A} = (\mathbf{A} + \mathbf{B}) - \mathbf{B};$$

pre and post multiplication by $\mathbf{A}^{-1}$ and $(\mathbf{A} + \mathbf{B})^{-1}$ (or *vice versa*) then gives us

$$\begin{aligned}(\mathbf{A}+\mathbf{B})^{-1} &= \mathbf{A}^{-1} - \mathbf{A}^{-1}\cdot\mathbf{B}\cdot(\mathbf{A}+\mathbf{B})^{-1} \\ &= \mathbf{A}^{-1} - (\mathbf{A}+\mathbf{B})^{-1}\cdot\mathbf{B}\cdot\mathbf{A}^{-1}\end{aligned} \tag{76}$$

The first of these identities can be iterated to yield

$$(\mathbf{A}+\mathbf{B})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\cdot\mathbf{B}\cdot\mathbf{A}^{-1} + \mathbf{A}^{-1}\cdot\mathbf{B}\cdot\mathbf{A}^{-1}\cdot\mathbf{B}\cdot\mathbf{A}^{-1}\cdots; \tag{77}$$

frequently this expansion can be re-arranged and then re-summed to yield useful matrix identities. To illustrate this procedure we 'invert' the relationship (c.f. (75))

$$\mathbf{K}(2)^{-1} = \mathbf{K}(2\,|\,1)^{-1} + \mathbf{C}^T\cdot\mathbf{Q}_2^{-1}\cdot\mathbf{C}$$

to give

$$\begin{aligned}\mathbf{K}(2) &= \mathbf{K}(2\,|\,1) - \mathbf{K}(2\,|\,1)\cdot\mathbf{C}^T\cdot\mathbf{Q}_2^{-1}\cdot\mathbf{C}\cdot\mathbf{K}(2\,|\,1) + \mathbf{K}(2\,|\,1)\cdot\mathbf{C}^T\cdot\mathbf{Q}_2^{-1}\cdot\mathbf{C}\cdot\mathbf{K}(2\,|\,1)\mathbf{C}^T\cdot\mathbf{Q}_2^{-1}\cdot\mathbf{C}\cdot\mathbf{K}(2\,|\,1)\cdots \\ &= \mathbf{K}(2\,|\,1) - \mathbf{K}(2\,|\,1)\cdot\mathbf{C}^T\cdot\left(\mathbf{Q}_2^{-1} - \mathbf{Q}_2^{-1}\cdot\mathbf{C}\cdot\mathbf{K}(2\,|\,1)\mathbf{C}^T\cdot\mathbf{Q}_2^{-1}\cdots\right)\cdot\mathbf{C}\cdot\mathbf{K}(2\,|\,1) \\ &= \mathbf{K}(2\,|\,1) - \mathbf{K}(2\,|\,1)\cdot\mathbf{C}^T\left(\mathbf{Q}_2 + \mathbf{C}\cdot\mathbf{K}(2\,|\,1)\mathbf{C}^T\right)^{-1}\cdot\mathbf{C}\cdot\mathbf{K}(2\,|\,1)\end{aligned} \tag{78}$$

If this expression for the Kalman gain is now substituted into the predictor corrector equation (75) we obtain an alternative expression for the optimum estimate. To this end we write

$$\begin{aligned}\mathbf{K}(2)\cdot\mathbf{C}^T\cdot\mathbf{Q}_2^{-1} &= \mathbf{K}(2\,|\,1)\cdot\mathbf{C}^T\cdot\left(\mathbf{Q}_2^{-1} - \left(\mathbf{Q}_2 + \mathbf{C}\cdot\mathbf{K}(2\,|\,1)\mathbf{C}^T\right)^{-1}\cdot\mathbf{C}\cdot\mathbf{K}(2\,|\,1)\mathbf{C}^T\cdot\mathbf{Q}_2^{-1}\right) \\ &= \mathbf{K}(2\,|\,1)\cdot\mathbf{C}^T\cdot\left(\mathbf{Q}_2 + \mathbf{C}\cdot\mathbf{K}(2\,|\,1)\mathbf{C}^T\right)^{-1}\end{aligned} \tag{79}$$

so that

$$\hat{\mathbf{x}}(2) = \mathbf{\Phi}\cdot\hat{\mathbf{x}}(1) + \mathbf{K}(2\,|\,1)\cdot\mathbf{C}^T\cdot\left(\mathbf{Q}_2 + \mathbf{C}\cdot\mathbf{K}(2\,|\,1)\mathbf{C}^T\right)^{-1}\cdot\left(\mathbf{y}(2) - \mathbf{C}\cdot\Phi\cdot\hat{\mathbf{x}}(1)\right) \tag{80}$$

While this expression is formally slightly more complicated than that in (75), it requires less work to be done inverting matrices in its evaluation.

Much the same analysis then applies to the processing of the third and subsequent measurements: the optimum running estimate of the process **x** is generated recursively as follows

$$\begin{aligned}&\mathbf{K}(n\,|\,n-1) = \mathbf{\Phi}\cdot\mathbf{K}(n-1)\cdot\mathbf{\Phi}^T + \mathbf{Q}_1 \\ &\mathbf{K}(n) = \mathbf{K}(n\,|\,n-1) - \mathbf{K}(n\,|\,n-1)\cdot\mathbf{C}^T\cdot\left(\mathbf{Q}_2 + \mathbf{C}^T\cdot\mathbf{K}(n\,|\,n-1)\cdot\mathbf{C}\right)^{-1}\cdot\mathbf{C}\cdot\mathbf{K}(n\,|\,n-1) \\ &\hat{\mathbf{x}}(n) = \mathbf{\Phi}\cdot\hat{\mathbf{x}}(n-1) + \mathbf{K}(n)\cdot\mathbf{C}^T\cdot\mathbf{Q}_2^{-1}\cdot\left(\mathbf{y}(n) - \mathbf{C}\cdot\mathbf{\Phi}\cdot\hat{\mathbf{x}}(n-1)\right) \\ &\mathbf{K}(n)\cdot\mathbf{C}^T\cdot\mathbf{Q}_2^{-1} = \mathbf{K}(n\,|\,n-1)\cdot\mathbf{C}^T\cdot\left(\mathbf{Q}_2 + \mathbf{C}\cdot\mathbf{K}(n\,|\,n-1)\mathbf{C}^T\right)^{-1}\end{aligned} \tag{81}$$

It is reassuring to note that, when we set $\mathbf{\Phi} = \mathbf{1}$, $\mathbf{Q}_1 = \mathbf{0}$, these equations reduce to those of sequential estimation (63), (64). This approach, based on a Gaussian model, was developed by

Swerling (who perhaps felt a little aggrieved that Kalman got all the credit); Gauss himself also discussed the basic ideas, back in the early years of the nineteenth century.

So far we have taken stationary Gaussian models for both the underlying process **x** and the measurements **y**; this has allowed us to calculate estimates, derived by likelihood maximization, in a straightforward fashion. The resulting scheme has a simple predictor-corrector form, which highlights the role of the innovations process. However the Kalman filtering technique is of much greater generality than this simple presentation might suggest. In particular it can be applied directly to non-stationary, non-Gaussian processes, for which the **Φ, C, Q** matrices vary with time. It is also possible to modify the filtering technique, which estimates $\mathbf{x}(n)$ from the series of measurements $\mathbf{y}(1),\cdots\mathbf{y}(n)$, to estimate $\mathbf{x}(n+s)$. When *s* is greater than zero we have a prediction process; negative *s* likewise gives us a smoothing process. The algorithms appropriate to these more general cases are derived in a rather different manner, based on a least squares criterion and focusing attention even more closely on the innovations sequence. Without the motivation provided by the foregoing likelihood maximization analysis, this approach might appear a little contrived; now we are in a position to better appreciate what it achieves. Bearing this in mind we will review the innovations based derivation, which was first presented in the seminal work of Kalman, '*A new approach to linear filtering and prediction problems*', J. Basic Engineering, **82**,35-45, 1960

Central to Kalman's development is the concept of a least squares estimate, expanded in an incomplete orthogonal set of basis functions. Thus we might write

$$\begin{aligned} f(x) &\approx f_n(x) = \sum_{k=1}^{n} a_k \phi_k(x) \\ \langle \phi_k(x) \phi_l(x) \rangle &= \delta_{k,l} \end{aligned} \tag{82}$$

We now wish to minimize the mean square error in this representation: thus we have

$$\begin{aligned} \Sigma &= \left\langle (f(x) - f_n(x))^2 \right\rangle = \left\langle f(x)^2 - 2f(x)\sum_{k=1}^{n} a_k \phi_k(x) + \sum_{k,l=1}^{n} a_l a_k \phi_k(x)\phi_l(x) \right\rangle \\ \frac{\partial \Sigma}{\partial a_k} &= 0 \Rightarrow \langle f(x)\phi_k(x) \rangle = a_k; k = 1,\cdots n \end{aligned} \tag{83}$$

The least square error condition identifies the coefficients in the expansion as the projections of the function *f* onto the basis functions; these coefficients do not depend on *n*. (You may remember seeing things like this before, in session 4.) We now note that the minimization condition can also be written as

$$\left\langle \left(f(x) - \sum_{k=1}^{n} a_k \phi_k(x) \right) \phi_m(x) \right\rangle = 0; \quad m = 1,\cdots n \tag{84}$$

This in turn shows that the error in the incomplete expansion, chosen to minimize this error, is itself orthogonal to the basis functions used in the expansion

$$\langle (f(x) - f_n(x)) \phi_m(x) \rangle = 0; \quad m = 1,\cdots n\,. \tag{85}$$

This argument can be generalized to vector processes and be stated formally as

$$\mathbf{f} = \sum_{n=1}^{M} \mathbf{A}(n) \cdot \mathbf{u}(n)$$

$$\left(\mathbf{f} - \sum_{k=1}^{M} \mathbf{A}(k) \cdot \mathbf{u}(k) \right) \left(\mathbf{f} - \sum_{k=1}^{M} \mathbf{A}(k) \cdot \mathbf{u}(k) \right)^{T} = \Sigma$$

$$\frac{\partial \Sigma}{\partial \mathbf{A}(m)^{T}} = 0 = -\mathbf{f}\mathbf{u}(m)^{T} + \sum_{k=1}^{M} \mathbf{A}(k) \cdot \mathbf{u}(k)\mathbf{u}(m)^{T} \tag{86}$$

$$\mathbf{e} = \mathbf{f} - \sum_{k=1}^{M} \mathbf{A}(k) \cdot \mathbf{u}(k)$$

$$\therefore \left\langle \mathbf{e}\mathbf{u}(m)^{T} \right\rangle = 0, \quad m = 0,1 \cdots n$$

We see from (81) that the Kalman filtering process represents the current estimate of **x** as a predicted contribution, formed from the previous estimate, and a correction term incorporating the difference between the current measurement and the value predicted from the previous estimate of **x**. This latter quantity has been termed the innovation, as it incorporates the 'new' information in the current measurement, beyond that which might be deduced prior to the receipt of the current measurement. Furthermore, if we were to construct our least squares estimate in terms of some orthogonal basis, the innovation derived from that estimate would be orthogonal to the basis functions used in its construction. This implies that a suitable basis set might be constructed from the successive innovations. These two observations focus our attention on the innovations sequence; to exploit this most effectively we now consider one-step prediction processing, rather than filtering. The innovations sequence arises most naturally in this context; the one-step prediction algorithm can then be modified quite straightforwardly to yield a Kalman filtering process.

We now consider a non-stationary process, characterized by a state vector **x**, which evolves over discrete time intervals as

$$\mathbf{x}(n+1) = \mathbf{\Phi}(n+1, n)\mathbf{x}(n) + \mathbf{v}_1(n) \tag{87}$$

Measurements are derived from the process **x** through

$$\mathbf{y}(n) = \mathbf{C}(n)\mathbf{x}(n) + \mathbf{v}_2(n), \tag{88}$$

C need not be a square matrix: the number of components in the state vector need not be equal to the number of components of the measurement vector. The evolution matrix **Φ** has some simple properties that will be exploited shortly; we note that

$$\begin{aligned} &\mathbf{\Phi}(n, n) = \mathbf{1} \\ &\mathbf{\Phi}(m, n)^{-1} = \mathbf{\Phi}(n, m) \end{aligned} \tag{89}$$

We assume that we know **Φ, C** and the correlation properties of the process and measurement noises. Thus we have

$$\begin{aligned}
&\left\langle \mathbf{v}_1(n)\mathbf{v}_1(m)^H \right\rangle = \delta_{nm}\mathbf{Q}_1(n) \\
&\left\langle \mathbf{v}_2(n)\mathbf{v}_2(m)^H \right\rangle = \delta_{nm}\mathbf{Q}_2(n) \\
&\left\langle \mathbf{v}_1(n)\mathbf{v}_2(m)^H \right\rangle = \mathbf{0}
\end{aligned} \tag{90}$$

Now let us make a least squares estimate of the *n*th measurement, on the basis of the previous *n*-1 measurements $\mathbf{Y}(n-1)$. This estimate is denoted by $\hat{\mathbf{y}}(n \mid \mathbf{Y}(n-1))$. The difference between this estimate and the actual value is referred to as the innovation, i.e. it gives us a measure of the new information in this measurement, beyond that provided by the previous measurements.

$$\mathbf{a}(n) = \mathbf{y}(n) - \hat{\mathbf{y}}(n \mid \mathbf{Y}(n-1)) \tag{91}$$

This innovation is orthogonal to the previous measurements, and to the previous innovations

$$\begin{aligned}
&\left\langle \mathbf{a}(n)\mathbf{y}(m)^T \right\rangle = \mathbf{0}, m = 1 \cdots n-1 \\
&\left\langle \mathbf{a}(n)\mathbf{a}(m)^T \right\rangle = \mathbf{0}, m = 1 \cdots n-1
\end{aligned} \tag{92}$$

There is also a one to one invertible correspondence between the measurements and the innovations; the orthogonal innovations sequence can be constructed from the measurement sequence by a Gram-Schmidt procedure.

We now wish to construct the covariance or the correlation matrix of the innovations process. If we have an estimate $\bar{\mathbf{x}}(n \mid \mathbf{Y}(n-1))$ of the current state vector based on the previous measurements then we can construct the innovation and evaluate its correlation matrix

$$\begin{aligned}
&\mathbf{a}(n) = \mathbf{y}(n) - \mathbf{C}(n)\bar{\mathbf{x}}(n \mid \mathbf{Y}(n-1)) = \mathbf{C}(n)\boldsymbol{\varepsilon}(n, n-1) + \mathbf{v}_2(n) \\
&\boldsymbol{\varepsilon}(n, n-1) = \mathbf{x}(n) - \bar{\mathbf{x}}(n \mid \mathbf{Y}(n-1)) \\
&\boldsymbol{\Sigma}(n) = \left\langle \mathbf{a}(n)\mathbf{a}(n)^T \right\rangle = \mathbf{C}(n)\mathbf{K}(n \mid n-1)\mathbf{C}(n)^T + \mathbf{Q}_2(n) \\
&\mathbf{K}(n \mid n-1) = \left\langle \boldsymbol{\varepsilon}(n, n-1)\boldsymbol{\varepsilon}(n, n-1)^T \right\rangle
\end{aligned} \tag{93}$$

Here we have adopted a notation that anticipates the identification of the results of the innovations analysis with those based on likelihood maximization.

We now set about constructing an estimate of a state vector, based on the set of *n* innovations; this is equivalent, through the invertibility of the Gram-Schmidt process, of making an estimate in terms of the *n* measurements that we have. Thus we write

$$\hat{\mathbf{x}}(i \mid \mathbf{Y}(n)) = \sum_{k=1}^{n} \mathbf{B}_i(k)\mathbf{a}(k) \tag{94}$$

The orthogonality principle tells us that

$$\left\langle \left(\mathbf{x}(i) - \hat{\mathbf{x}}(i \mid \mathbf{Y}(n))\right)\mathbf{a}(m) \right\rangle = \mathbf{0}, m = 1 \cdots n \tag{95}$$

so that, on substituting our expansion of the estimate in terms of the innovation sequence, we find that

$$\begin{aligned}\left\langle \mathbf{x}(i)\mathbf{a}(m)^T \right\rangle &= \mathbf{B}_i(m)\left\langle \mathbf{a}(m)\mathbf{a}(m)^T \right\rangle = \mathbf{B}_i(m)\boldsymbol{\Sigma}(m) \\ \mathbf{B}_i(m) &= \left\langle \mathbf{x}(i)\mathbf{a}(m)^T \right\rangle \boldsymbol{\Sigma}(m)^{-1}\end{aligned} \tag{96}$$

$\boldsymbol{\Sigma}(m)$ is the covariance matrix of the mth innovation. Thus we can write our estimate as

$$\hat{\mathbf{x}}(i \mid \mathbf{Y}(n)) = \sum_{m=1}^{n} \left\langle \mathbf{x}(i)\mathbf{a}(m)^T \right\rangle \boldsymbol{\Sigma}(m)^{-1}\mathbf{a}(m) \tag{97}$$

In particular we have

$$\hat{\mathbf{x}}(n+1 \mid \mathbf{Y}(n)) = \sum_{m=1}^{n} \left\langle \mathbf{x}(n+1)\mathbf{a}(m)^T \right\rangle \boldsymbol{\Sigma}(m)^{-1}\mathbf{a}(m) \tag{98}$$

Now we can express this estimate recursively in terms of its predecessor, and a correction:

$$\begin{aligned}\hat{\mathbf{x}}(n+1 \mid \mathbf{Y}(n)) &= \left\langle \mathbf{x}(n+1)\mathbf{a}(n)^T \right\rangle \boldsymbol{\Sigma}(n)^{-1}\mathbf{a}(n) + \boldsymbol{\Phi}(n+1,n)\sum_{m=1}^{n-1} \left\langle \mathbf{x}(n)\mathbf{a}(m)^T \right\rangle \boldsymbol{\Sigma}(m)^{-1}\mathbf{a}(m) \\ &= \boldsymbol{\Phi}(n+1,n)\hat{\mathbf{x}}(n \mid \mathbf{Y}(n-1)) + \mathbf{G}(n)\mathbf{a}(n) \\ \mathbf{G}(n) &= \left\langle \mathbf{x}(n+1)\mathbf{a}(n)^T \right\rangle \boldsymbol{\Sigma}(n)^{-1}\end{aligned} \tag{99}$$

This is the fundamental Kalman recursion relation, where we predict the next estimate in terms of the previous one, then correct it in terms of the innovation in the new measurement.

To make further progress we now evaluate the Kalman gain matrix **G**. To this end we note that

$$\begin{aligned}&\left\langle \mathbf{x}(n+1)\mathbf{a}(m)^T \right\rangle = \boldsymbol{\Phi}(n+1,n)\left\langle \mathbf{x}(n)\mathbf{a}(n)^T \right\rangle + \left\langle \mathbf{v}_1(n)\mathbf{a}(n)^T \right\rangle; \quad \left\langle \mathbf{v}_1(n)\mathbf{a}(n)^T \right\rangle = \mathbf{0} \\ &\left\langle \mathbf{x}(n)\mathbf{a}(n)^T \right\rangle = \left\langle \mathbf{x}(n)(\mathbf{C}(n)\boldsymbol{\varepsilon}(n,n-1) + \mathbf{v}_2(n))^T \right\rangle = \left\langle \boldsymbol{\varepsilon}(n,n-1)\boldsymbol{\varepsilon}(n,n-1)^T \right\rangle \mathbf{C}(n)^T \\ &= \mathbf{K}(n \mid n-1)\mathbf{C}(n)^T; \mathbf{K}(n \mid n-1) = \left\langle \boldsymbol{\varepsilon}(n,n-1)\boldsymbol{\varepsilon}(n,n-1)^T \right\rangle\end{aligned} \tag{100}$$

Now we make the identification

$$\mathbf{G}(n) = \boldsymbol{\Phi}(n+1,n)\mathbf{K}(n \mid n-1)\mathbf{C}^T(n)\boldsymbol{\Sigma}(n)^{-1} \tag{101}$$

recalling that this is the gain for the one-step prediction process.

Now we wish to establish a recurrence relation satisfied by the $\mathbf{K}(n \mid n-1)$; the earlier analysis of the stationary Gaussian model provides us with a few clues to what we should do. Thus we form the error in the one-step prediction and establish a recurrence relation:

$$\begin{aligned}\boldsymbol{\varepsilon}(n+1,n) &= \mathbf{x}(n+1) - \hat{\mathbf{x}}(n+1 \mid \mathbf{Y}(n)) \\ &= \boldsymbol{\Phi}(n+1,n)\left[\mathbf{x}(n) - \hat{\mathbf{x}}(n \mid \mathbf{Y}(n-1))\right] - \mathbf{G}(n)\left[\mathbf{y}(n) - \mathbf{C}(n)\hat{\mathbf{x}}(n \mid \mathbf{Y}(n-1))\right] + \mathbf{v}_1(n) \\ &= \left[\boldsymbol{\Phi}(n+1,n) - \mathbf{G}(n)\mathbf{C}(n)\right]\boldsymbol{\varepsilon}(n,n-1) + \mathbf{v}_1(n) - \mathbf{G}(n)\mathbf{v}_2(n)\end{aligned} \tag{102}$$

Squaring and averaging we find that

$$\mathbf{K}(n+1 \mid n) = [\mathbf{\Phi}(n+1,n) - \mathbf{G}(n)\mathbf{C}(n)]\mathbf{K}(n \mid n-1)[\mathbf{\Phi}(n+1,n) - \mathbf{G}(n)\mathbf{C}(n)]^T + \mathbf{Q}_1(n) + \mathbf{G}(n)\mathbf{Q}_2(n)\mathbf{G}(n)^T \tag{103}$$

while expanding out and tidying things up gives

$$\mathbf{K}(n+1 \mid n) = \mathbf{\Phi}(n+1,n)\mathbf{K}(n)\mathbf{\Phi}(n+1,n)^T + \mathbf{Q}_1(n) \tag{104}$$

where

$$\mathbf{K}(n) = (\mathbf{1} - \mathbf{\Phi}(n,n+1)\mathbf{G}(n)\mathbf{C}(n))\mathbf{K}(n \mid n-1) \tag{105}$$

The result (104) is often referred to as the Riccatti difference equation; should everything settle down to an equilibrium then the predicted state error correlation matrix will tend to a constant value **K** satisfying the identity

$$\mathbf{K}\mathbf{C}^T\left(\mathbf{C}\mathbf{K}\mathbf{C}^T + \mathbf{Q}_2\right)^{-1}\mathbf{C}\mathbf{K} - \mathbf{Q}_1 = \mathbf{0} \tag{106}$$

The results derived thus far allow us to set up a recursive one-step prediction algorithm. The output of this is readily modified to give us a filtering algorithm, which can be compared directly with that derived from the Gaussian likelihoods model. Given an estimate of $\mathbf{x}(n)$ given $\mathbf{Y}(n)$ we form the corresponding estimate of $\mathbf{x}(n+1)$ from the evolution equation as follows

$$\hat{\mathbf{x}}(n+1 \mid \mathbf{Y}(n)) = \mathbf{\Phi}(n+1,n)\cdot\hat{\mathbf{x}}(n \mid \mathbf{Y}(n)) + \hat{\mathbf{v}}(n) = \mathbf{\Phi}(n+1,n)\cdot\hat{\mathbf{x}}(n \mid \mathbf{Y}(n)) \tag{107}$$

as the estimate $\hat{\mathbf{v}}(n)$ of a zero mean Gaussian process uncorrelated with the process **x** is zero. Thus we can write

$$\begin{aligned}\hat{\mathbf{x}}(n \mid \mathbf{Y}(n)) &= \mathbf{\Phi}(n,n+1)\cdot\hat{\mathbf{x}}(n+1 \mid \mathbf{Y}(n)) \\ &= \hat{\mathbf{x}}(n \mid \mathbf{Y}(n-1)) + \mathbf{\Phi}(n,n+1)\cdot\mathbf{G}(n)\cdot\mathbf{a}(n) \\ &= \mathbf{\Phi}(n,n-1)\cdot\hat{\mathbf{x}}(n-1 \mid \mathbf{Y}(n-1)) + \mathbf{K}(n \mid n-1)\cdot\mathbf{C}(n)^T\cdot\mathbf{R}(n)^{-1}\cdot\mathbf{a}(n)\end{aligned} \tag{108}$$

This can be seen to map directly onto the results (93) obtained earlier, and to generalise them to the non-stationary case.

Retracing the steps that led from (75) to (79), but in the reverse order, we see that

$$\mathbf{K}(n \mid n-1)\cdot\mathbf{C}(n)^T\cdot\mathbf{R}(n)^{-1} = \mathbf{K}(n)\cdot\mathbf{C}(n)^T\cdot\mathbf{Q}_2(n)^{-1} \tag{109}$$

Finally we note that (105) expresses $\mathbf{K}(n)$, which is necessarily positive, as the difference of two matrices; numerical errors can result in $\mathbf{K}(n)$ becoming negative. This instability can be avoided by expressing $\mathbf{K}(n)$ in the so-called Joseph stabilised form, which displays it as a sum of two necessarily positive forms:

$$\begin{aligned}\mathbf{K}(n) &= (\mathbf{1} - \mathbf{G}'(n)\cdot\mathbf{C}(n))\cdot\mathbf{K}\cdot(\mathbf{1} - \mathbf{G}'(n)\cdot\mathbf{C}(n))^T + \mathbf{G}'(n)\cdot\mathbf{Q}_2(n)\cdot\mathbf{G}'(n)^T \\ \mathbf{G}'(n) &= \mathbf{\Phi}(n,n+1)\cdot\mathbf{G}(n)\end{aligned} \tag{110}$$

It is hoped that this preliminary account of the Kalman filter has demonstrated how its predictor-corrector form combines prior knowledge (from the underlying model of the process and earlier measurements) with the new information (derived from the most recent measurement) to give a constantly up-dated estimate of the state of the underlying process. Some formal manipulations of covariance matrices, that are frequently omitted, albeit for rather different reasons, in both elementary and advanced texts, can nonetheless be rather unfamiliar and have been discussed here in sufficient detail to make them more readily accessible. Even so, it should be stressed that this brief account has done no more than scratch the surface of the vast body of work that has grown over the past forty or so years. Obviously, any choice of review material is a matter of personal taste: van Trees, '*Detection, Estimation and Modulation Theory, Part 1*', John Wiley, New York, 1968, as always, provides a thorough grounding in the relevant estimation theory and discusses the early work on the subject. Haykin's book '*Adaptive Filter Theory*', Prentice Hall, New Jersey, 1996, provides a great deal of useful background material on the Kalman filter and its generalisations, showing how the principles of its derivation and implementation underpin much current work on adaptive filters. Finally a closer examination of the mathematical foundations of the subject can be found in Kailath's '*Lectures on Wiener and Kalman filtering'* Springer, New York, 1981. It is interesting to note how like a lot of economics modeling the Kalman approach is; an exhaustive account is given in 'Dynamic Econometrics', D. F. Hendry, Oxford University Press, 1997 which nonetheless contrives to omit the words Kalman and filter from all of its 900 or so pages.

Exercises

The material covered today makes much closer contact with 'practical' radar problems than has been achieved in earlier sessions. This new realism is reflected in the exercises, which encourage you to fill in the details we have glossed over in the notes. If you are pressed for time you might identify the application closest to your own interests and concentrate on that. A more leisurely run through of the full set of exercises should give you a working knowledge of detection theory, as applied to a variety of radar systems, and put you in pretty good shape for attacking the research literature.

1. Draw out some ROC curves based on the result (21) for a fixed value of $\langle I_a \rangle$ and several values of $\langle I_t \rangle$. Check the derivation of (25) and use this to evaluate P_D and P_F, based on the same $\langle I_a \rangle, \langle I_t \rangle$ and different *N*. Present your results as ROC curves and comment on them. Fill in the details of the derivations of (26), (30) and (31). Plot out the pdfs for various $A^2/\langle I \rangle$ and *N*. Compare the means and normalised variances of the clutter and target plus clutter pdfs for a fixed $A^2/\langle I \rangle$ and varying *N*. Comment on your results. (Mathematica is very helpful when it comes to evaluating and plotting out the modified Bessel functions.)

2. In (30) we derived the form of the Rice distribution; try out the following alternative derivation. If $\mathbf{E} = \mathbf{n} + \mathbf{A}$ and $\mathbf{n}$ has the pdf

$$P(n_I, n_Q) = \frac{1}{\pi \langle I \rangle} \exp\left(-\left(n_Q^2 + n_I^2\right)/\langle I \rangle\right)$$

what is the pdf $P(E_I, E_Q)$? Express this in terms of the amplitude and phase of $\mathbf{E}$ to give

$$P(E, \theta) = \frac{E}{\pi \langle I \rangle} \exp\left(-\left(E^2 + A^2\right)/\langle I \rangle\right) \exp\left(2EA\cos\theta/\langle I \rangle\right)$$

the phase θ being measured relative to that of **A.** Evaluate the marginal distributions of *E* and θ. Relate the former to (30) through $I = E^2$. Here we have regarded E_I, E_Q as components of a vector in a two dimensional space. By considering a vector in a 2*N* dimensional space extend the above argument to derive (31). Spherical polar co-ordinates in an arbitrary number of dimensions are required for this; one possible reference is Sommerfeld: Partial differential equations in physics, Academic Press, 1949 Chapter 5, Appendix 4 (p227).

3 Starting with the joint distributions (54) of *t* and *s*, evaluate the marginal distribution of *t* when *m*, the mean of the underlying Gaussian distribution, is zero. You should get

$$P(t) = \frac{1}{\mathrm{B}((n-1)/2, 1/2)\left(1 + t^2\right)^{n/2}}$$

Here we have introduced the beta function

$$\mathrm{B}(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)}$$

which lends its name to this distribution. You might recognise this as the Student *t* distribution. What form does the pdf of *t* take when *m* is not zero? You might like to avail yourself of Mathematica as a short-cut through the fancy sums. Derive the following approximation, valid when m^2/σ^2 is large

$$P(t)=\frac{2}{\Gamma((n-1)/2)\Gamma(1/2)}\frac{\exp\left(-nm^2/\left(2\sigma^2\left(1+t^2\right)\right)\right)}{\left(1+t^2\right)^{\frac{n}{2}}}\sum_{q=0}^{\left[\frac{n-1}{2}\right]}\binom{n-1}{2q}\Gamma(q+1/2)\left(mt\sqrt{\frac{n}{2\sigma^2\left(1+t^2\right)}}\right)^{(n-1-2q)} \tag{45}$$

([*z*] represents the integer part of *z*).. Check the derivation of (55) and bash out a few ROC curves of your own.

4 Working your way through the details of the Kalman filter derivations is a worthwhile exercise in itself. To provide some motivation, check that, if the predicted state error correlation matrix 'settles down', it satisfies the equilibrium Ricatti difference equation (106). In addition, show that (105) is equivalent to the Joseph stabilised form (110).

5 The following has nothing to do with the material in the session, but is quite neat. Consider the mean absolute difference between two identical, independent random variables, which can be written as

$$\Delta_1=\int_{-\infty}^{\infty}dx\int_{-\infty}^{\infty}dy|x-y|p(x)p(y).$$

where *p* is the pdf of an individual random variable. Show that

$$\Delta_1=2\int_{-\infty}^{\infty}F(y)(1-F(y))dy$$

where *F* is the cumulative probability

$$F(x)=\int_{-\infty}^{x}p(y)dy\,.$$