ISIT'98 Plenary Lecture Report:

From Matched Filters to Martingales

Thomas Kailath

1. Introduction

The hopes of the organizers for this special session were that it would cover statistical detection and estimation theory, topics that were major areas of investigation in the first three decades of Information Theory. In recent years, most of the activity in these areas has been reported elsewhere.

The symposium talk covered a number of topics (some old, some new, some borrowed, none blue), going beyond the advertised title. Here we briefly present a few of them. The first is the rapidly growing area of techniques for blind channel equalization using second-order statistics, commonly thought only to apply to minimum phase channels. For potentially non-minimum-phase channels, the only option seemed to be to use higher-order statistics, but these need more data to estimate well and more complicated algorithms, both of which are unreasonable in rapidly changing environments, as encountered, for example, in mobile wireless systems. In Sec. 3, we note the origin of the matched filter and the early (1947) work of Kotel'nikov on optimal signal detection in additive white Gaussian noise. We remark how close, and yet how far, Kotel'nikov was to Shannon's channel capacity concept, even for this special, but important, channel. The final topic is an even briefer review of the search for insight into the structure of likelihood ratios for signal detection. A key concept in uncovering such structure is a generalization, using the modern (post 1967) theory of martingale processes, of the concept of innovations introduced by Bode and Shannon (1950) to provide a more insightful derivation of Wiener's celebrated results on the prediction and filtering of stationary stochastic processes.

2. Blind Channel Equalization

The basic equalization problem is indicated in Fig. 1.

If the channel $H(z)$ is known, or can be identified, we can choose the equalizer (in the absence of noise) as $G(z) = H^{-1}(z)$. Of course $G^{-1}(z)$ will be an IIR (infinite impulse response) filter, even when the channel is

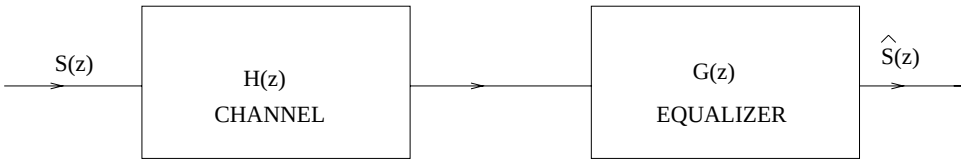

Figure 1:

(modeled as) an FIR filter. Moreover when $G(z)$ is not minimum-phase, $H(z)$ will be noncausal, but this can be accommodated by introducing a sufficient delay in the equalizer. The problem is to identify $H(z)$. If we make the (common) assumption that the input sequence is an uncorrelated unit variance random process, then the power spectral density of the output of the channel will be $H(z)H^*(z^{-1}) + \sigma^2 I$, if we also have additive white noise of intensity σ^2. While σ^2 can be determined fairly easily, the problem is that we cannot recover $H(z)$ from the product $H(z)H^*(z^{-1})$, unless $H(z)$ is minimum-phase. However, on further reflection, all we have shown is that phase information cannot be recovered from *stationary* second-order statistics, such as the power spectral density function. As mentioned in the introduction, it turns out that phase information can (often) be recovered from *nonstationary* second-order statistics. And in particular from cyclostationary (or periodically correlated) second-order processes.

A New Solution : Use oversampling (when excess BW is available), as is done already for other reasons, e.g., 'clock recovery,' leading to what are called fractionally spaced equalizers. We demonstrate now a deeper reason for using such equalizers.

To present the main ideas in the simplest context, consider oversampling the received signal by a factor of two. The transmitted signal is kept at the original rate and to accommodate the oversampling we can repeat the information, so that (see Fig. 2) the oversampled signal can be written as

$$s(z) = (1+z)S(z^2) = (1+z)(S_0 + S_2 z^2 + \ldots)$$

We now separate out the even and odd samples of the received signal,

$$\begin{aligned} y(z) &= H(z)(1+z)S(z^2) \stackrel{\triangle}{=} h(z)S(z^2) \\ &\stackrel{\triangle}{=} (h_e(z^2) + z h_o(z^2))S(z^2) \end{aligned}$$

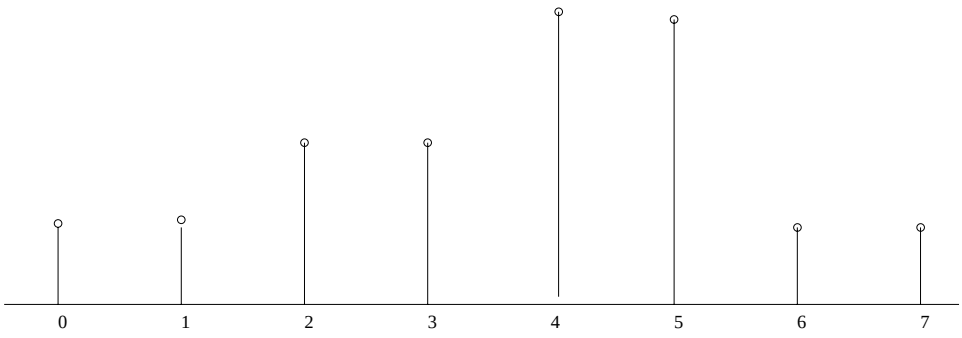

Figure 2: Oversampled transmitted sequence

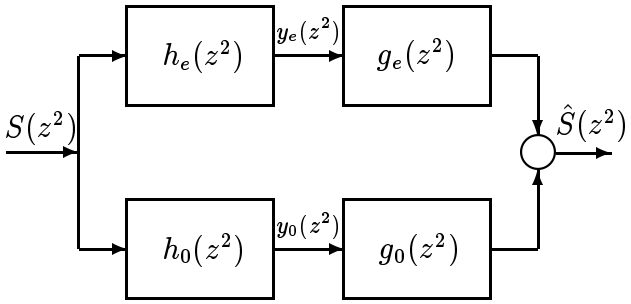

Figure 3: An equivalent single-input, two-output channel model

and process them individually before reconstructing the results (see Fig. 3) to obtain

$$\hat{S}(z^2) = \left[g_e(z^2)h_e(z^2) + g_o(z^2)h_o(z^2)\right] S(z^2)\,.$$

But now if $h_e(z^2)$ and $h_o(z^2)$ are coprime polynomials, then we can choose polynomials $\{g_e(z^2),\, g_o(z^2)\}$ such that (the Bezout identity)

$$g_e(z^2)h_e(z^2) + g_o(z^2)h_o(z^2) = 1(\text{ or } z^d)$$

holds, which means that the signal can be recovered perfectly (or with a delay d). So, in the absence of noise, we shall have perfect equalization, and moreover using FIR filters, provided of course that we can identify $h_e(z^2)$ and $h_o(z^2)$. For this we first form the covariance matrix of the received pair of sequences $\{y_e(k), y_o(k)\}$, and take its z-transform to obtain the power

spectral density matrix. Assuming that the input signal sequence can be modeled as uncorrelated equal variance (unity for convenience), this matrix will be as shown below:

$$\mathcal{Z}\left\{\begin{bmatrix} Ey_e(k)y_e^*(k-i) & Ey_e(k)y_o^*(k-i) \\ Ey_o(k)y_e^*(k-i) & Ey_o(k)y_o^*(k-i) \end{bmatrix}\right\} = \begin{bmatrix} h_e(z^2)h_e^*(z^{-2}) & h_e(z^2)h_o^*(z^{-2}) \\ h_o(z^2)h_e^*(z^{-2}) & h_o(z^2)h_o^*(z^{-2}) \end{bmatrix}$$

Now when $h_e(z^2)$ and $h_o(z^2)$ are coprime, we can find $h_e(z^2)$ as the common factor of the $(1,1)$ and $(1,2)$ entries. Similarly for $h_o(z^2)$! So, in the ideal case, we have shown that we can equalize a nonminimum-phase channel using oversampling FIR filters and second-order statistics.

The basic ideas behind this surprising result were first presented in the paper of Tong, Xu, Kailath (Asilomar Conf. Proceedings, 1991; IT Trans. 94); the presentation given above also uses ideas from further joint work with Hassibi (IT, Jan 95; Asilomar, 93). We should mention that the multiple channels that oversampling allows us to define are directly available when antenna arrays are used; this is explained in the Asilomar 93 paper.

Of course the above procedure is sensitive to the effects of noise and of error in estimating the covariance functions using a finite amount of data. The noise can be accounted for by using least-squares or the more recent H_∞ (minimax) filtering criteria. The development of effective algorithms in the finite data case is currently an area of active research. There are three classes of techniques for approaching this problem: Sylvester Matrix Techniques, Subspace Techniques, and Linear Prediction and Smoothing Techniques.

We refer for details on these results to the now-extensive literature, which appears largely in signal processing journals. Recent survey articles include Liu et al., Signal Processing, 1996, and a follow-up survey by Tong and Perreau, appearing in a special Oct.98 issue of the IEEE Proceedings on Blind System Identification and Estimation.

3. **Matched Filters; North, Kotel'nikov and Shannon**

The origins of signal detection theory go back to World War II when researchers began to explore the possibilities of replacing human decision makers peering at a radar screen with an automated decision making device. One of the most famous early results here is the matched filter, introduced by the physicist D.O. North in a 1943 RCA Princeton lab report (reprinted in the Proc. IEEE Jul. 1963). The matched filter maximizes the output SNR for

a known signal corrupted by additive noise. North makes a remarkably advanced analysis of radar problems, showing great facility with statistical calculations (the Rice distribution appears here) and physical approximations. Unfortunately for us, "at the end of the war, solid-state physics beckoned, and [North] turned to it." [North passed away, in his nineties, a few weeks before the symposium; however, Vince Poor and Sergio Verdú did manage to talk to him by phone a few weeks earlier.] While North computed the probabilities of detection and of false alarm, he was not aware of the Neyman Pearson lemma showing that calculating the likelihood ratio enabled an optimal tradeoff between these two probabilities. The famous mathematician, M. Kac, used to joke that his main contribution to the war effort was providing a reference to the Neyman-Pearson theory to A.J.F. Siegert. In digital communications problems (e.g., FSK, PSK) the criterion is minimizing the overall error probability, and here again the likelihood ratio is the key statistic to compute. This was perhaps first recognized by V.A. Kotel'nikov (in the USSR) in a remarkable doctoral dissertation submitted in 1947, The Theory of Optimum Noise Immunity (translated into English and published by McGraw-Hill in 1960). [Kotel'nikov's 90th birthday was celebrated in Moscow on Sept. 6 and also acknowledged at the IT Symposium.] The thesis treats binary and multiple signal detection in additive white Gaussian noise, and also parameter estimation problems in digital and analog communication systems; many of the results in it were only rediscovered several years later. In particular, Kotel'nikov used geometrical arguments and interpretations very effectively. Among these one finds the nice geometric interpretation of the threshold effect in bandwidth-expanding modulation schemes such as FM, which was made famous in Shannon's 1949 paper and further elaborated in the classic textbook of Wozencraft and Jacobs. In Kotel'nikov's words: "However, when the length of the curve is increased, the distance between separate "twists" or sections of the curve must decrease, which perforce increases the probability of anomalous errors."

Though Kotel'nikov was very close to the notion of channel capacity for the wideband AWGN channel, he missed it, because (as he mentioned in a conversation on the occasion of the First (and only) Joint IEEE-USSR Academy of Sciences Symposium on Information Theory in Moscow, Dec. 1975), he never even dreamt of the possibility that one could have communication at a nonzero rate with arbitrarily small probability of error. The point is that Kotel'nikov often used the simple "union bound" on the error probability for M equal energy orthogonal signals in white Gaussian noise

(in a standard notation, $P_e \leq M \exp -(P_{av}/2N_0)T$) and used it to study the advantages of multiple versus binary signaling. It was Shannon's great insight that by allowing M to increase with time, and by using a logarithmic measure of signaling rate,

$$M = e^{RT}, R = (\ln M)/T$$

one could make P_e go to zero as $T \to \infty$, provided that R was not too high:

$$\begin{aligned} P_e \quad &\leq M \exp -(P_{av}/2N_0)T \\ &= \exp -[R - (P_{av}/2N_0)]T \\ &\to 0 \text{ for all} R < P_{av}/2N_0 \end{aligned}$$

In fact, of course, the actual capacity of the wideband $AWGN$ channel is higher, $C = P_{av}/N_0$, as was simply demonstrated in the lecture using a result on the asymptotic estimate of the maximum of M i.i.d. normal random variables as $M \longrightarrow \infty$. Briefly, the point is that the matched filter output corresponding to the actual transmitted signal is a $N(P_{av}, \sqrt{P_{av}N_o/2T})$ random variable, while the outputs of the $M-1$ other matched filters are $N(0, \sqrt{P_{av}N_o/2T})$. It might seem that for large T we will rarely make a mistake, no matter how large M is. However while the "incorrect" matched filter outputs are all very close to zero for large T, their *maximum* value tends, not to zero, but to $\sqrt{P_{av}N_o/2T} \cdot \sqrt{2 \ln M} = \sqrt{P_{av}N_o R}$. Hence there will certainly be an error unless $\sqrt{P_{av}N_o R} < P_{av}$, i.e., unless $R < P_{av}/N_o = \lim_{W \longrightarrow \infty} W \log (1 + \frac{P_{av}}{N_o W})$, the capacity of the wideband channel! This pretty argument was shown to me by Jack Stiffler at JPL in 1962. It provides a fine illustration of Shannon's fundamental observation that "Delay has the (additional) function of allowing a large sample of noise to affect the signal before any judgment is made at the receiving point as to the original message. Increasing the sample size always sharpens the possible statistical assertions." (Shannon, 1948, Sec. 19).

4. **The Structure of Likelihood Ratios**

Interest in likelihood ratios is again increasing in the information theory community, in part because of the importance of soft decoding in the new

turbocodes. A long survey paper in the special Oct. 98 issue of the IT Transactions (with Vince Poor) gives a detailed account of this topic. So, as in the talk, here we even more briefly outline the main message.

Given complete statistical information and adequate computational resources one can always evaluate the L.R. as (in standard notation)

$$L.R. = \lim_{n\to\infty} \frac{w_1(y(t_0^{(n)}), ..., y(t_n^{(n)}))}{w_0(y(t_0^{(n)}), ..., y(t_n^{(n)}))}$$

However, as with all applications of mathematics to engineering problems, we need to understand enough of the structure of mathematical solution that we can make intelligent approximations when the solution is too complicated to actually evaluate or realistically implement, especially when we only have inaccurate or incomplete knowledge of the parameters in our model. The only recourse we have is to look for structure and insight in the exact (analytic) solutions to as many special cases as possible.

Chief among these is the L.R., first given by Kotel'nikov, for the problem of choosing between the hypotheses

$$H_1 : y(t) = m(t) + v(t) \text{ and } H_0 : y(t) = v(t), \qquad 0 \le t \le T$$

where $m(\cdot)$ is a completely known signal of energy E and $v(\cdot)$ is unit intensity white Gaussian noise (WGN):

$$L(T) = \exp\left[\int_0^T m(t)y(t)dt - \frac{1}{2}\int_0^T m^2(t)dt\right] \tag{1}$$

The basic operation on the data is correlating the possible transmitted signal waveform $m(\cdot)$ against the received waveform $y(\cdot)$. As is well known, this correlation integral can also be computed as the output at time T of a filter *matched* to $m(\cdot)$ and (i.e., with impulse response $m(T..)$) driven by $y(\cdot)$. In other words, North's matched filter derived under a SNR criterion is in fact optimal in the stronger sense (of minimum probability of error).

Exact but more complicated L.R. formulas can be found when the signal is known except say for phase or for phase and amplitude. A very widely studied case is that of Gaussian signals. Here the usual hypotheses are:

$$H_1 : y(t) = z(t) + v(t), H_0 : y(t) = v(t)$$

where $v(\cdot)$ is again zero-mean unit-intensity WGN and $z(\cdot)$ is a zero-mean Gaussian random process independent of $v(\cdot)$ and having a continuous covariance function, $K(t,s)$. Then Price (IT, 1956) showed that the L.R. could be calculated as

$$L(T) = (\text{F. Det.}) \cdot \exp \int_0^T \int_0^T y(t)H(t,s)y(s)dt\,ds \tag{2}$$

where $H(.,.)$ is the solution to the integral equation

$$H(t,s) + \int_0^T H(t,\tau)K(\tau,s)dt = K(t,s),\ 0 \le t,s \le T \le \infty,$$

and F. Det. is the so-called Fredholm determinant, $\Pi_1^\infty(1+\lambda_i)$, where $\{\lambda_i\}$ are the eigenvalues of $K(.,.)$ over [0,T] x [0, T].

This is quite an explicit formula, but it illustrates some of the issues we mentioned earlier. First of all, explicit solutions of the integral equation are only available for a very few known functions $K(.,.)$. Are there good approximate solutions for other $K(.,.)$? What can we do if we only have a rough idea of what $K(.,.)$ is? Or when $z(\cdot)$ is a stationary process and we only have a general idea of what its power spectral density function is? How do we compute the Fredholm determinant? And so on.

Being well aware of such issues, Price was very happy to find that under his assumptions, he could show the following: Denote the inner integral in (2) as

$$\int_0^T H(t,s)y(s)ds \stackrel{\triangle}{=} z_e(t)\,.$$

Then for each t, it turns out that $z_e(t)$ is the least-squares estimate of $z(t)$ given *all* (past and future) the observations $y(t) = z(t)+v(t)$, $0 \le t \le T$. The double integral in (2) then becomes $\int y(t)z_e(t)dt$, which can be implemented by a filter matched to $z_e(\cdot)$. This is a nice tie-in to the known signal case – when $z(\cdot)$ is random and therefore unknown, we form the mean-square estimate $z_e(\cdot)$ and then proceed (almost) as in the known signal case. An immediate bonus of this interpretation is that it provides a reasonable answer to the previous questions – in the face of limited knowledge, we put in the best estimator we can produce. If for example all we know is that the power spectrum has roughly a certain shape and a certain bandwidth, a first cut at an estimating filter is one with a transfer function roughly the same shape and bandwidth. This may or may not sound reasonable to all readers, but suffice it to say that it was precisely intelligent approximations of this kind that

were used in the now-famous RAKE anti-multipath receiver (Price and Green (Proc. IEEE,1958)), which is now again gaining attention in the wireless field.

A question is whether such interpretations are available for non-Gaussian signals. In fact they are, and actually in a form much closer to the original L.R. formula (1) for the known signal case. Before presenting this result, however, a further note on the much studied Gaussian case will be useful.

First of all it is important to allow for correlation between the signal and noise processes. For example, in feedback systems the present signal is a function of past observations. It turns out that Price's interpretation breaks down in this case – $z_e(\cdot)$ is no longer an estimate. Secondly, for greatest generality one should allow square-integrable covariance functions. In this case the Fredholm determinant may not exist and the formula (2) breaks down. The appropriate generalization was found by Shepp (Ann. Stat., 1966)

$$L(T) = (\text{F.C. Det.}) \exp \int_0^T c \int_0^T y(t) H(t,s) y(s) dt\, ds \tag{3}$$

Here, F.C. Det. $= \Pi_1^\infty (1+\lambda_i) e^{-\lambda_i}$, a so-called Fredholm-Carleman determinant [The point is that in the general case, $\sum |\lambda_i|$ may diverge, making the usual Fredholm determinant infinite. However the Fredholm-Carleman determinant will exist whenever the L.R. is well defined (see Shepp (1966) or Kailath (IT, May 1970)).] As for the other term, the "c" between the integrals is used to indicate something that most engineers have never had to face before – the fact that a "new" kind of integral has to be used, in this case a so-called multiple Wiener stochastic integral. There is really no problem with this – either in theory (we just have to introduce the appropriate definitions) or in practice (the new integral can be (approximately) calculated using available hardware); the same comments apply to the so-called Ito stochastic integral mentioned below. For more on these issues at a tutorial level, see Kailath (IT, 1969, May 1970a).

Here we go on to the promised result for (Gaussian and) non-Gaussian $z(\cdot)$. Following some fundamental work by F.C. Schweppe (IT, 1965), by R.L. Stratonovich and his colleagues (e.g., Stratonovich and Sosulin (1964), and in the Stanford dissertation of T.E. Duncan (1967), the following general result was presented in Kailath (IT, 1969), with a rigorous proof using modern martingale theory in Kailath (IT, July 1970b).

Assume that the signal $z(\cdot)$ has finite energy, but is not necessarily Gaussian, while the noise $v(\cdot)$ is unit intensity white and Gaussian. Also that $z(\cdot)$ and $v(\cdot)$ may be dependent, as long as future $v(\cdot)$ are independent of

past signal $z(\cdot)$ (as in feedback communications). Let $\hat{z}_1(t)$ = the causal least-squares estimate of $z(t)$ given past $y(\cdot)$, and assuming H_1 holds (i.e., $y(t) = z(t) + v(t)$). Then the L.R. has exactly the same form as for known signals in WGN:

$$L(T) = \exp \int\!\!\!\!\!\!- \int_0^T \hat{z}(t)y(t)dt - \frac{1}{2}\int_0^T \hat{z}^2(t)dt \tag{4}$$

where $\int\!\!\!\!\!\!-(\ )$ denotes a so-called Ito stochastic integral. With this definition, (3) can be shown to be equivalent to all earlier explicit L.R. formulas, including (2) and (3). However the real point is that (3) gives a universal structure into which we can insert our best available causal signal estimate to obtain a reasonable approximation to the L.R. Such structural information is the most valuable information mathematical results can give about real world problems! A further indication that this is a basic result is that a similar estimator-correlator structure also holds for the apparently very different non-Gaussian detection problems using jump-process observations. e.g., choosing between Poisson processes with different random rates.

The basic idea underlying the proof is the following result:

Given a process $y(t) = z(t) + v(t)$, introduce the innovations process,

$$\begin{aligned} i(t) &= y(t) - \hat{y}(t|t-) = y(t) - \hat{z}(t) \\ &= \text{the new information in } y(\cdot) \text{ at time t} \end{aligned}$$

Perhaps not surprisingly, this process is white (i.e., its values at different times are *uncorrelated* with each other) but in fact it is also Gaussian, so that they are *independent* of each other. Moreover $i(\cdot)$ has the same intensity as $v(\cdot)$. Therefore the original hypothesis

$$H_1 : y(t) = z(t) + v(t), \qquad H_0 : y(t) = v(t),$$

can be replaced by

$$H_1 : y(t) = \hat{z}(t) + i(t), \qquad H_0 : y(t) = i(t).$$

But $i(\cdot)$ is WGN, and $\hat{z}_1(\cdot)$ is *conditionally known, given past* $y(\cdot)$. Hence it is reasonable that $L(T)$ is the same as in the known signal case, thus leading up to the formula (4).

Of course this heuristic (but rigorizable) argument raises several questions, even before the issue of making it precise. For example (just to begin):

Why *least-squares* signal estimates rather than any others? Why is $i(\cdot)$ Gaussian, though neither $y(\cdot)$ nor $\hat{z}(\cdot)$ need be? Moreover, according to the traditional definition, in which a Gaussian process is completely defined by its mean and covariance function, the WGN processes $v(\cdot)$ and $i(\cdot)$ should be indistinguishable. But they are clearly not the same! In fact, $i(t) = y(t) - \hat{z}(t) = (z(t) - \hat{z}(t)) + v(t) = \tilde{z}(t) + v(t) \neq v(t)$. So how can we distinguish them?

To answer these, and several further related questions, we have to bring in some concepts not usually covered in first courses on random processes; in particular, the concepts of sigma fields of events, and of martingales with respect to increasing families of sigma fields. Here we shall assume knowledge of them in order to outline an answer to our last question: how to distinguish the Gaussian processes $v(\cdot)$ and $\hat{z}(\cdot)$?

Given the processes $\{z(\cdot), v(\cdot)\}$, we introduce the increasing family of sigma fields generated by $y(\cdot) = z(\cdot) + v(\cdot)$, $\mathcal{F}_t = \sigma\{y(\tau), \tau \leq t\}$, $0 \leq t \leq T$, and also the larger family $\mathcal{B}_t = \sigma\{z(\tau), v(\tau), z \leq t\}$. [The larger family corresponds to the state of knowledge of an omniscient observer, involved in setting up the original model!] Then we may note that, for $s < t$,

$$E[v(t)|\mathcal{B}_s] = 0 \text{ but } E[v(t)|\mathcal{F}_s] \neq 0$$

while

$$E[i(t)|\mathcal{B}_s] \neq 0, \text{ but } E[i(t)|\mathcal{F}_s] = 0\,.$$

So this may be one way in which the processes may be distinguished.

In more traditional language, one would introduce the integrated processes

$$V(t) = \int_0^t v(\tau)d\tau \text{ and } I(t) = \int_0^t i(\tau)d\tau$$

in which case the above statements are equivalent to

$$E[V(t)|\mathcal{B}_s] = V(s)\,,\; E[V(t)|\mathcal{F}_s] \neq V(s)$$

while

$$E[I(t)|\mathcal{F}_s] = I(s)\,, E[I(t)|\mathcal{B}_s] \neq V(t)\,.$$

In other words, even though $V(\cdot)$ and $I(\cdot)$ are both Gaussian with the same mean and covariance,they are different because $I(\cdot)$ is a *martingale* with respect to the family of sigma fields $\{\mathcal{F}_t\}$ generated by the observations, but is not a martingale with respect to the fields $\{\mathcal{B}_t\}$; the opposite is true

for the process $V(\cdot)$. Many other beautiful results arise from martingale theory (as largely developed by French and Japanese probabilists beginning in the late 1960s) in establishing (4) and its generalizations (e.g., Kailath and Duttweiler, IT Nov. 73; Segall and Kailath, Ann. Prob., 1975), discussion of which we must forego here.

However, a final thought, especially appropriate as we celebrate the Golden Jubilee of Information **Theory**, is to recall the words of Ludwig Boltzmann: There is nothing so practical as a good theory.