Technical Report No. 225

004860-3-T

SIMULTANEOUS DETECTION AND ESTIMATION: THE USE OF SUFFICIENT STATISTICS AND REPRODUCING PROBABILITY DENSITIES

by

Jurgen O. Gobien

Approved by: Theodore G. Birdsall

COOLEY ELECTRONICS LABORATORY

Department of Electrical and Computer Engineering
The University of Michigan
Ann Arbor, Michigan

for

Contract No. N0014-67-A-0181-0035
Office of Naval Research
Department of the Navy
Arlington, Va. 22217

November 1973

ABSTRACT

The problem considered postulates two mutually exclusive and exhaustive statistical hypotheses, under each of which the probability distribution on the observation space is known except for (conditional to) a finite-dimensional parameter. The parameters may or may not have common components, and are considered as random variables and assigned a priori probability density functions (p.d.f.'s). It includes the standard signal-detection problem where either the signal, or the noise, or both, may contain uncertain parameters.

Estimation is defined as knowledge of the a posteriori parameter p.d.f. given by Bayes' rule. The likelihood ratio of marginal observation distributions is considered an optimal detection statistic. It is shown that this statistic can be found by using the two separate estimation results to modify a related simple-hypothesis detection statistic. Thus, estimation and detection occur simultaneously in a very natural fashion.

The concept and existence of necessary and sufficient statistics are investigated. If the conditional observation distribution under either hypothesis admits a sufficient statistic of fixed dimension, then a natural conjugate family of parameter densities exists

and is indexed by a "conjugate parameter" of the same dimension. Explicit relations can be found to "update" the conjugate parameter based on the sufficient statistic; usually the procedure is recursive. Explicit use of Bayes' rule becomes unnecessary and the estimation problem is reduced to a tractable, fixed-dimensional procedure. The detection problem is similarly simplified.

It is shown that any p.d.f. nonzero on the same parameter space as the natural conjugate density also reproduces. Much of the signal processing is shown to be independent of the a priori parameter densities.

All results are rigorously extended to include observations which are continuous-parameter random processes. To illustrate the theory, the problem of detecting a known signal in and simultaneously estimating the parameters of M^{th}-order stationary autoregressive Gaussian (Gauss-Markov) noise is addressed. Solutions are found for both the discrete (sampled) and the continuous case. The estimation solution is tractable; the detection statistic is complicated. It is written in closed form for $M = 1$. For arbitrary (known) values of M, it contains integrals which are quite difficult and are left unevaluated.

FOREWORD

This doctoral thesis represents a major theoretical breakthrough in the fields of detection theory and estimation. Previous research has been based on the use of the Shannon sampling theorem or the Karhanen-Loeve theorem; both require knowledge of the noise autocorrelation function. In practice, this knowledge is often unavailable. In situations where this lack of knowledge may critically influence the design and performance of equipment, it is necessary to "tell the mathematics" about this uncertainty of the noise characteristics.

This doctoral thesis develops both the foundation and the techniques for working with uncertainty about the noise process. It is hoped that its distribution to research workers in detection and estimation theory will spur renewed interest and progress in extending theories toward handling more realistic situations and thereby be of more direct help to the practical equipment designer.

TABLE OF CONTENTS

TABLE OF CONTENTS (Cont.)

TABLE OF CONTENTS (Cont.)

LIST OF ILLUSTRATIONS

LIST OF APPENDICES

LIST OF SYMBOLS AND ABBREVIATIONS[1]

a.e., a.s.	almost everywhere, almost surely; true except on a set of measure zero [#B.1].
$\mathscr{A}$	the σ-algebra induced on the observation space $\mathscr{Y}$ by the observed random process [#B.1, #2.2].
$\mathscr{A}_0$, $\mathscr{A}_*$	sub σ-algebras of $\mathscr{A}$ [#B.5].
c.d.f.	cumulative distribution function [#B.2].
D_0, D_1	the decision that H_0 [resp. H_1] is true.
$\{e_k\}$	a white sequence of unit Gaussian random variables.
$E(\cdot)$	mathematical expectation.
$E^{\mathscr{A}_0}(\cdot)$	conditional expectation w.r.t. the sub-σ-algebra $\mathscr{A}_0$ [#B.5, (B.13)].
f	frequency.
$f(\cdot)$	a probability density function identified by its arguments [#1.1].
$g[t(y); \theta]$	the factor of the conditional observation p.d.f. which depends on the parameter [(2.10)].

[1] Where necessary, references to the text are made in square brackets. Section numbers are preceded by # , and equation numbers placed in parentheses.

$G(y)$	the factor of the conditional observation p.d.f. which does not depend on the parameter [(2.10)] .
$G(\cdot\,,\,\cdot)$	a detection-theory goal functional satisfying certain regularity properties [#1.3.1, #1.3.3] .
$h(\cdot\,;\theta)$	an alternate way of writing the transition density of a Markov process [(3.35)] .
H_0, H_1	mutually exclusive and exhaustive statistical hypotheses, usually representing "noise only" and "signal plus noise" resp. in communications problems.
$H(j2\pi f)$	a Fourier transfer function.
j	the square root of -1 .
$J(\cdot)$	a cost functional for Bayesian estimation.
$L(\theta\,;\underset{\sim}{y})$	the natural log of the transition likelihood ratio function for a Markov process [(2.28), (2.29)] .
$\mathscr{L}$	Lebesgue measure.
$\ell(\cdot)$	the likelihood ratio [(1.14)] .
MAP	maximum a posteriori, an estimate based on the mode of the a posteriori p.d.f.
M-SAG	M^{th}-order stationary autoregressive Gaussian.
$\mathscr{M}$	the set of measures $\mathscr{P}_\theta$ on the observations, indexed by $\theta \in \Theta^\theta$ [#B.4] .
n_t	a continuous-parameter noise process.
$\{n_i\}$	a discrete noise sequence.

$N(\cdot, \cdot)$	the Gaussian distribution with mean given by the first argument and variance by the second.
$\mathcal{O}(\cdot)$	of the order of, in the sense that $\delta \mathcal{O}(\delta^{-1}) \to M < \infty$ as $\delta \to 0$.
p_i	the upper-half-plane poles of a rational spectral density [(5.2), (5.3)].
p.d.f.	probability density function.
$p(\theta ; \gamma)$	the natural conjugate density on the parameters of H_1 [(3.8)].
$P(\cdot)$	a probability
$P_{k, \theta}$	the Borel probability measure given by the p.d.f. $f(\underline{Y}_k \mid \theta)$ [#2.2, (2.19)].
$\mathscr{P}, \mathscr{P}_\theta$	probability measures on the observation space [#B.1].
q_i	coefficients of the polynomial $Q(\cdot)$.
$q(\eta ; \psi)$	the natural conjugate density on the parameters of H_0.
$Q(\cdot)$	a polynomial with roots whose real parts are negative [(5.4), (5.5)].
r_i	the cross-correlation of sequential samples separated by i instants [(4.10)].
$\underline{r}$	the M-vector of cross-correlations of a stationary M-th order Markov sequence.
$r_\theta(\cdot), r_\eta(\cdot)$	the R-N derivatives relating the actual and natural conjugate a priori densities under H_1 and H_0 resp. [(3.19)].

R-N	Radon-Nikodym.
R	a covariance matrix [(4.9), (4.11)].
R^n	n-dimensional Euclidean space.
R^+	the complement of the negative half-line.
$R_y(\tau)$	The autocorrelation function $E\, y_t y_{t+\tau}$.
$\mathscr{R}$	Bayes' risk, the expectation of $J(\cdot)$.
$\mathscr{R}^n$	the Borel sets on R^n
s(t)	an exactly-known deterministic signal.
$S_y(f^2)$	the spectral density function of the process y_t .
$\mathscr{S}$	the admissible set of values for the auto-regressive parameters $\underline{\beta}$ [#4.1.1, (4.3)].
t	time.
$t(\cdot)$	a sufficient statistic.
T, T_i	fixed instants denoting the beginning or end of an observation interval.
$\mathscr{T}$	the space in which the sufficient statistic $t(\cdot)$ takes values.
w.r.t.	with respect to.
y	a scalar observation.
y_t	a continuous-time observation.

LIST OF SYMBOLS AND ABBREVIATIONS (Cont.)

$\underline{Y}_k$	a k-vector of sequential scalar observations [(1.1)] .
$\mathcal{Y}, \mathcal{Y}^k$	the observation space and its k-fold Cartesian product.
$z(\cdot)$	the natural log of the likelihood ratio.
α	the intensity parameter of the discrete autoregression [(4.2), (4.6)] ; the threshold for the likelihood ratio [(1.15)] .
β_i , $\underline{\beta}$	parameters of the discrete autoregression [(4.2), (4.6)] .
γ, Γ	the conjugate parameter under H_1 and the space in which it assumes values [#3.1, (3.8)] .
δ	the sampling interval.
ϵ	belongs to, is a member of.
η, $\mathcal{N}$	the uncertain parameter of the observation under H_0 , and the associated parameter space [#1.2.3] .
θ , Θ	The uncertain parameter of the observation under H_1 , and the associated parameter space [#1.2.3] .
$\hat{\theta}$, θ^*, θ_0	an estimate, the true value, and a fixed value of θ .
$\lambda(y_t, \theta)$	The R-N derivative given by the limit of the likelihood ratio function [(2.21)] .
$\Lambda_k(\underline{Y}_k, \theta)$	the likelihood ratio function of k sequential samples [(1.30)] .

$\Lambda_0(\underset{\sim}{\tilde{y}}, \theta)$	the transition likelihood ratio function of a Markov process [(2.27)] .
$\Omega(\cdot)$	the logarithmic derivative of the unit Gaussian c.d.f. [#5.4.1] .
$\phi(\cdot)$	the unit Gaussian p.d.f.
$\Phi(\cdot)$	the unit Gaussian c.d.f.
ρ	an alternate intensity parameter for the discrete autoregression [(4.22)] .
[ ]	modulo, with respect to; also, a reference to the bibliography or a closed interval on the real line.
$\forall$	for any, for all.
|	given, conditional to.
*	transpose, complex conjugate
$\rightarrow$	tends to, has the limit.
$\subset$	is a subset of .
$\underline{\underline{d}}$	is defined as.
$\equiv$	denotes equivalence of measures [#B.3] .
$\perp$	denotes singularity of measures [#B.3] .
$<<$	denotes absolute continuity of measures [#B.3] .
■	denotes the end of theorems, proofs, and samples.

CHAPTER I

INTRODUCTION

The theory of signal detection and estimation is the study of methods for determining the presence of, and extracting useful information from, communication signals corrupted by random interference. As such, it is cast within the framework of the statistical theories of hypothesis testing and point estimation and shares a great deal of mathematical ground with other disciplines cast in that framework; for example, pattern recognition and statistical decision theory. Both of the latter disciplines explicitly apply and make use of some statistical results concerning sufficient statistics and reproducing probability densities. It is the aim of this dissertation to apply those same concepts to the theory of signal detection and estimation. To this end, it is necessary to rederive the results in terms of the language and notation of detection theory. Then, the available results are considerably extended by applying them to the infinite-dimensional function spaces generated by random processes. Finally, they are applied to the detection and estimation problem and some significant new results are obtained.

This work concerns itself with problems in which the signal, the corrupting noise, or both contain a finite number of parameters

which are not known; it is especially significant that the methods developed are easily applied to unknown noise parameters. As will be shown, the results allow application of the classical detection theory results without the assumption that a complete statistical description of the noise is available. The viewpoint taken is essentially Bayesian; that is, unknown parameters are considered as random variables with probability distributions which reflect a composite of the observer's subjective opinion and of past observation of their state.[1] As further observations are made, these distributions are modified according to Bayes' Rule. This viewpoint is a matter of some controversy in the statistical literature (see, e.g., Savage [53]) and certainly there are cases where it is unjustified. For many problems of interest in detection theory, the Bayesian approach is quite valuable. If nothing else, the domain of an unknown parameter can often be bounded by practical considerations and ignorance beyond this point expressed as a uniform a priori distribution.[2] Further, it is shown in Appendix C that Bayesian techniques are just as

---

[1] This is the only intended implication of the word "Bayesian"; in particular, the results are not restricted to the linear or quadratic cost functionals which "Bayesian" is sometimes taken to imply.

[2] See, for example, Kashyap [31].

applicable if the unknown parameters are not random but if one admits a utility function which represents a specification of the estimator and detector performance as a function of the "true" parameter values.

Since Bayesian methods in general utilize and manipulate probability distributions rather than just numbers, the data storage and calculation necessary to their use often seem prohibitive. One purpose of this dissertation is to demonstrate how, for a large class of problems, these difficulties can be surmounted.

A word is in order concerning the mathematical level of this work. It is intended that a first-year graduate education in communications engineering which includes one or two courses in probability and statistics be a sufficient background to read and apply the results. Thus, the use of analysis or measure theory is specifically avoided whenever possible. Occasionally (e.g., in sections of Chapters II, III, and V) it is not; the offending sections are marked with an asterisk. To read them, one needs the background of Appendix B, as well as some knowledge of random process theory. This dichotomy in mathematical levels has the acknowledged effect of making the material somewhat longer and occasionally more tedious than necessary, but is considered worthwhile.

The notation is standard. Footnotes are referenced in raised parentheses$^{(\ )}$, references to the bibliography are made in square brackets [ ], equation numbers are placed on line in parentheses ( ), and the end of theorems, proofs, examples, etc., is indicated by a square block. ■ An effort has been made to use conventional engineering symbols, abbreviations, and notation whenever possible. For convenience, a list of symbols is in the preliminaries.

The material is organized as follows. Chapter I first states and then presents traditional solutions to three problems: simple-hypothesis signal detection, Bayesian point estimation, and compound-hypothesis detection. An interesting new model, useful for the solution of certain simple-hypothesis detection problems, is given in Appendix A and is used to obtain solutions needed in later chapters. In addition to stating the problems in some detail, Chapter I establishes the notation to be used and concludes with a brief historical review of the subject.

Chapter II addresses the concept, properties, and existence of necessary and sufficient statistics for a family of probability distributions. After a presentation of the usual statistical concepts, some modern measure-theoretic results concerning sufficient statistics are applied to the function spaces generated by a random process; this allows application of these concepts to observations

made continuously in time. The chapter concludes by reconsidering finite-dimensional observations; many of the classical results are rederived for the case where discrete samples possess an M^{th}-order Markov dependence, as is often true in communications and control problems.

Chapter III defines and demonstrates the existence of classes of reproducing probability densities, and then applies the properties of such classes to the Bayesian sequential and "one-shot" detection and estimation problems; these are shown to be simplified considerably. The detection problem is shown to partition in such a way that a great deal of the signal processing becomes independent of the a priori distributions. Finally, the results are extended to observations made continuously in time, and the sequential treatment of such observations is addressed.

Chapter IV applies the theory to the problem of detecting a known sure signal in, and simultaneously estimating the spectral parameters of, discrete M^{th}-order stationary autoregressive Gaussian (M-SAG) noise. General solutions are derived, and the case $M = 1$ is done in detail. The results are quite complicated; no attempt is made to simplify or approximate the discrete solution. Instead, it is left to stand in contrast to the material of Chapter V, which solves the same problem for signal and noise

observed continuously in time. Although the theory is much harder, the results are considerably simpler than the corresponding "sampled" results of Chapter IV.

The word "solutions" above should be interpreted loosely. The purpose of Chapters IV and V is to illustrate the application of and to further clarify the theory; no claim as to the practicality of the results is made. Estimation, for instance, will consist of finding the parameters which specify the a posteriori p.d.f. as a member of some known family. This is fine in theory, but actually making a sensible estimate based on that density (i.e., on its parameters) is a completely new problem, and is not one which this work intends to address. The same holds for detection; usually the detection statistic, though optimal, is so complicated as to be impractical. Again, no attempts will be made to approximate or simplify the results.

One final apology is necessary at the outset. A key feature of the theory of Chapters II and III is that it is in no way restricted to problems which are linear, Gaussian, involve quadratic costs, or any of the other usual constraints; yet, the examples of Chapters IV and V are both linear and Gaussian. There are two reasons for this: First is the usual one that Gaussian noise is indeed a very realistic model, especially in communications problems. Second, it was desirable that one unified example serve for both

the discrete and the continuous case. In the latter, the singularity of (or the existence of densities for) abstract measures become important topics for which very few practical results exist once one leaves the realm of Gaussian measures.

1.1 General Notation and Assumptions

Throughout this work, observations will be denoted as y and will lie in a set $\mathcal{Y}$ which represents the totality of all possible observations. They are generated by some sort of random mechanism (process, experiment), and one wishes to use them to make inferences (of a nature to be made clear in Section 1.2) about that generating mechanism. There are two broad categories of observations which will be considered:

The first is that y is a sample function from a real-valued random process $\{y_t, t\in[0, T]\}$, where t takes on continuous values in the fixed, finite interval $[0, T]$. The sample space is a measure space $(\mathcal{Y}, \mathcal{A}, \mathcal{P}_\theta)$ of real-valued functions on $[0, T]$; $\mathcal{P}_\theta$ is a probability measure belonging to a family $\mathcal{M} = \{\mathcal{P}_\theta; \theta\in\Theta\}$ indexed by a finite-dimensional parameter. Solution of this case is quite difficult and will usually be approached by taking a suitable limit of solutions obtained for the second category of observations, namely:

The observation is a real number belonging to some domain $\mathcal{Y} \subset R^1$. All results could just as well have been derived for

$y \in R^n$, but at considerable notational expense. If a single observation is to be processed, the Borel probability measure on $\mathcal{Y}$ will be considered as given by a member of a family $\{f(y|\theta);\theta \in \Theta\}$ of probability density functions (p.d.f.'s) indexed by a finite-dimensional parameter. Throughout this dissertation, it will be assumed that all Borel probability measures are absolutely continuous with respect to Lebesgue measure and hence given by p.d.f.'s. A totally rigorous and unambiguous notation would require that these p.d.f.'s be identified by a subscript, and that different symbols be used to distinguish random variables, dummy variables, and constants in their arguments. Such a notation makes equations rather cumbersome, especially where conditional p.d.f.'s are employed; further, it can be difficult to read without practice. Hence, this dissertation will use the shorter but admittedly ambiguous notation of letting $f(\cdot)$ denote a p.d.f. which is identified by its arguments, e.g.,

$$f(\theta) \neq f(\eta)|_{\eta=\theta}$$

Further, the notation will not explicitly show in what sense those arguments are to be considered; ambiguities will be resolved by context or by an explicit comment.

Often, one is interested not in a single observation y but

in a finite-length sequence of observations which the process has generated. Denote the first k observations of interest as

$$\underline{Y}_k = (y_1, y_2, \cdots, y_k) \tag{1.1}$$

which is, if appropriate, considered a column vector. Clearly, $\underline{Y}_k \in \mathcal{Y}^k \subset R^k$. If no assumptions are made regarding their statistical dependence, its elements have a joint p.d.f. which belongs to $\{f(\underline{Y}_k|\theta); \theta \in \Theta\}$; these families may be different for different values of k, and hence require an arbitrarily large "hard memory" to store.[1] The densities can be written sequentially using the relation

$$f(\underline{Y}_k|\theta) = \prod_{i=1}^{k} f(y_i|\underline{Y}_{i-1}, \theta) \tag{1.2}$$

but this does not lessen the storage requirement.

The situation improves if more structure is placed on the problem. Consider, for instance, that the samples possess an M^{th}-order Markov dependence with stationary transitions.[2] For reasons apparent later, it is then convenient to include the

---

[1] "Hard" memory contains information inherent in the problem statement and stays fixed as θ is learned.

[2] Each observation is statistically dependent upon only the preceding M.

samples $(y_{-M+1}, \cdots, y_0)$ and write

$$f(y_{-M+1}\cdots y_0\cdots y_k|\theta) = f(y_{-M+1}\cdots y_0|\theta) \prod_{i=1}^{k} f(y_i|y_{i-1}\cdots y_{i-M},\theta) \tag{1.3}$$

This is simplified by defining the state vector

$$\underline{y}_i = (y_i\cdots y_{i-M+1}) \in \mathcal{Y}^M \tag{1.4}$$

Conditioned upon $\underline{y}_0$, (1.3) becomes

$$f(\underline{Y}_k|\underline{y}_0,\theta) = \prod_{i=1}^{k} f(y_i|\underline{y}_{i-1},\theta) \tag{1.5}$$

The structure of this equation will be important in the sequel, and it will often be convenient to condition all results upon the "initial observations" $\underline{y}_0 = (y_{-M+1}, \cdots, y_0)$ in this fashion. It is important to note that the results will not in general be the same as if unconditioned, and the conditioning must finally be undone or justified in some way. The case of statistically independent observations is given by $M = 0$, i.e.,

$$f(\underline{Y}_k|\theta) = \prod_{i=1}^{k} f(y_i|\theta) \tag{1.6}$$

Note that (1. 5) and (1. 6) present two simplifications. First, they are inherently sequential so that the joint density of k observations can be found using only the p. d. f. of $(k - 1)$ observations and the preceding M observations themselves. Second, a saving of "hard memory" is apparent since knowledge of the family of transition densities $\{f(y|\underset{\sim}{y}, \theta); \theta \in \Theta\}$ yields the (conditioned) family of joint p. d. f. 's for any k .

Many results for the continuous process $\{y_t; t \in [0, T]\}$ will be obtained by considering that k samples are drawn from the interval. Let T_k denote the set of k sampling instants

$$T_k = \{t_i = \frac{i}{k} T \; ; \; i = 1, 2, \ldots, k\} \tag{1.7}$$

equally spaced with sampling interval $\delta_k = T/k$. One obtains a discrete process $\{y_t, \; t \in T_k\}$; it will be assumed that the distribution of $\underset{\sim}{Y}_k$, the vector of (1. 1) with elements $y_i = y_{t_i}$, satisfies all requirements posed above for the finite-dimensional case. As $k \to \infty$ (i. e., $\delta_k \to 0$) , the samples grow dense in $[0, T]$ and, under suitable restrictions, inferences can be made about the continuous process.[1]

---

[1] Inferences made by this scheme are actually based on observing the half-open interval $(0, T]$. If the sample functions are a. s. continuous, this is clearly equivalent to observing $[0, T]$. If not, it may be desirable to include the point $t = 0$ in the finite sample sets.

If the continuous process is M^{th}-order Markov[1], then it is again convenient to condition all expressions upon y_0 ; these will not be identical to the expressions given above until the conditioning is accounted for. There is an important distinction between inferences which include initial observations and those which are conditioned on initial observations.

1.2 Statement of the Problem

This section describes the problem to be solved by first stating two sub-problems: the simple-hypothesis detection problem , and the Bayesian estimation problem. Finally a combination of the two, the doubly compound hypothesis detection problem will be posed. It is this last problem of simultaneous detection and estimation which is of concern; the sub-problems are separated as above for historical reasons and because the solution will separate similarly. Section 1.3 will be organized similarly and will present general solutions to the problems; Chapter III will later apply the concepts of sufficient statistics which are introduced in Chapter II.

---

[1] This means that any collection of sequential samples is M^{th}-order Markov as previously defined.

1.2.1 The Simple-Hypothesis Detection Problem. The basic signal detection problem postulates two mutually exclusive and exhaustive hypotheses, denoted as H_0 ("noise alone is present") and H_1 ("signal and noise are present"). The "true" hypothesis activates some sort of probabilistic mechanism (communications channel); the result is mathematically described as placing a corresponding probability measure on the space of observables $\mathcal{Y}$; i.e., one of two p.d.f.'s is active on $\mathcal{Y}$: $f(y|H_0)$ or $f(y|H_1)$.[1] By observing $y \in \mathcal{Y}$, the "detector" must decide which hypothesis is true; the two possible decisions are denoted D_0 and D_1 .

The detector will be designed to maximize a "goal functional" $G[P(D_1|H_1), P(D_1|H_0)]$ which is a real function of the "detection probability" $P(D_1|H_1)$ and the "false-alarm probability" $P(D_1|H_0)$. The only restriction on $G(\cdot, \cdot)$ is that it be monotone nondecreasing in its first argument and monotone nonincreasing in the second: It does not penalize correct decisions or reward false ones.

---

[1] These statements should be interpreted as referring to measures instead of p.d.f.'s if the observations are continuous time functions. This liberty with notation will be taken throughout.

The following examples of simple-hypothesis detection problems establish notation to be used in later chapters.

Example 1. 1(a)

The observation is a continuous random process, given under each hypothesis by:

$$H_0 : y_t = n_t$$

$$H_1 : y_t = s(t) + n_t \quad , \quad t\epsilon[0, T]$$

where $s(t)$ is an exactly-known deterministic signal, and n_t is first-order stationary autoregressive Gaussian (1-SAG) noise of zero mean with known spectral density

$$S_n(\omega^2) = \frac{a^2}{\omega^2 + p_1^2} \; ; p_1 > 0 \qquad (1.8)$$

and autocorrelation function

$$R_n(\tau) = \frac{a^2}{2p_1} e^{-p_1|\tau|} \qquad (1.9)$$

This noise process is often called the Ornstein-Uhlenbeck process.

The observation y_t , $t\epsilon[0, T]$, is to be processed to determine whether or not the signal is present; "optimality" is determined in accordance with a goal functional G which satisfies the assumptions made above. ∎

Example 1.1(b)

Sample the observation of the preceding example as described in (1.7). Then

$$H_0 : \underline{Y}_k = \underline{N}_k$$

$$H_1 : \underline{Y}_k = \underline{S}_k + \underline{N}_k$$

and the observation is a vector-valued random variable which is easily treated by classical methods. ∎

1.2.2 The Bayesian Estimation Problem. In the preceding section, the family of distributions on $\mathcal{Y}$ had two members. Now, assume that this family is, as stated in Section 1.1, given by the p.d.f.'s $\{f(y|\theta); \theta \in \Theta\}$ where the parameter set Θ is a domain in R^m. Let $f(y|\theta^*)$ be the member of the family which is active on $\mathcal{Y}$.

Before making an observation, the subjective and objective knowledge about θ^* is summarized as an a priori p.d.f. $f_0(\theta)$, $\theta \in \Theta$. This may be an actual explicitly known probability density, or it may be a density function chosen from some class for its

ability to model the observer's a priori knowledge.[1][2]

Observations are used to "update" the probability distribution on Θ, resulting in a new a posteriori p.d.f. after each observation is made. This p.d.f. may then be used to make an optimal estimate $\hat{\theta}(y)$ of θ^* with respect to any desired criterion as shown in Section 1.3.2; the results of this dissertation will not be restricted to specific criteria. Instead, "estimation" will always imply explicit knowledge of the a posteriori p.d.f.

Example 1.2(a)

Put

$$\theta = \begin{bmatrix} a^2 \\ p_1 \end{bmatrix}, \quad p_1 > 0$$

and let the p.d.f. $f_0(\theta)$ summarize all previous knowledge about θ^*. The observation is continuous,

---

[1] Such modeling techniques constitute an active field of study; see, e.g., Kashyap [31] or De Groot [10], Ch. 6.

[2] Appendix C shows that Bayesian techniques are as applicable (though the philosophy is different) if one considers θ a fixed but unknown constant and uses an integrable utility or performance function in place of the a priori p.d.f., or even if one has a bona-fide a priori p.d.f. and wishes to combine a utility specification with it.

$$y_t = n_t \quad , \quad t\epsilon[0, T]$$

where n_t is 1-SAG noise as described in Example 1.1(a); "a^2" and "p_1" are the parameters of the noise spectral density, and they are to be estimated to minimize the cost functional

$$J = E_{\mathcal{Y}_{x\Theta}} \| \hat{\theta}(y) - \theta \|^2 \quad \blacksquare$$

Example 1.2(b)

Sample the observation of Example 1.2(a), so that the parameters to be estimated are unchanged but $\mathcal{Y} = R^1$ and the observation is $\underline{Y}_k = \underline{N}_k \epsilon \mathcal{Y}^k$. The cost is

$$J = E_{\mathcal{Y}^k x\Theta} \| \hat{\theta}_k(\underline{Y}_k) - \theta \|^2 \quad \blacksquare$$

Example 1.2(c)

Chapter V will show that the preceding example is equivalent to:

Let $\{y_i, i = 1, 2, \ldots\}$ be the first-order autoregressive process which satisfies

$$y_i + \beta_1 y_{i-1} = \alpha e_i \quad ; \quad -1 < \beta_1 < 0$$

with initial condition selected at random from a $N(0, \frac{\alpha^2}{1-\beta_1^2})$ distribution. $\{e_k\}$ is a sequence of independent $N(0, 1)$ random variables.

The parameters of this problem are related to those of Example 1.2(b) by

$$\beta_1 = -e^{-p_1\delta}$$

$$\alpha^2 = \frac{a^2}{2p_1}(1 - e^{-2p_1\delta})$$

where δ is the sampling interval. ∎

1.2.3 The Compound-Hypothesis Detection Problem. The situation is similar to the simple-hypothesis problem, except that under either or both hypotheses the active probability measure on $\mathcal{Y}$ is one of a family indexed by an unknown parameter. Under H_1, this parameter is denoted $\theta \in \Theta$ and under H_0 it is $\eta \in \mathcal{N}$; these may or may not have common components. The corresponding families of densities $\{f(y|\theta, H_1); \theta \in \Theta\}$ and $\{f(y|\eta, H_0); \eta \in \mathcal{N}\}$ are known. In essence, this problem is a combination of the two already defined.

The reward of a Bayesian approach to this problem is two-fold: First, it allows detection and estimation to occur simultaneously in a natural way, especially for sequential observations.

Second, it permits sidestepping a difficult problem of statistical decision theory; namely, it avoids the issue of finding a "uniformly most powerful" decision statistic (one which is optimal for any admissible θ *) . This is accomplished by optimizing the detector with respect to a goal functional exactly as was done for the simple-hypothesis problem of Section 1. 2. 1, except that detection and false-alarm probabilities are now defined to be the marginal probabilities

$$P(D_1|H_1) = \int_{\Theta} P(D_1|H_1, \theta)\, f_0(\theta)\, d\theta \tag{1.10}$$

$$P(D_1|H_0) = \int_{\mathcal{N}} P(D_1|H_0, \eta)\, f_0(\eta)\, d\eta \tag{1.11}$$

where $f_0(\theta)$, $f_0(\eta)$ are the a priori p. d. f. 's.

Once the detector makes a decision, that hypothesis is assumed true and the corresponding parameter is to be estimated precisely as in Section 1. 1. 2.

Example 1.3

In the previous notation

$$H_0 : y_t = n_t$$

$$H_1 : y_t = \alpha s(t) + n_t \quad , \quad t \in [0, T]$$

where $s(t)$ is a signal known exactly and n_t is 1-SAG noise (Example 1.1). The values of α, a^2, and p_1 are not known, so

$$\theta = \begin{bmatrix} \alpha \\ a^2 \\ p_1 \end{bmatrix} \quad , \quad \eta = \begin{bmatrix} a^2 \\ \\ p_1 \end{bmatrix}$$

Note that some parameters are common to both hypotheses. ∎

1.3 Theoretical Foundations: the Traditional Solution

1.3.1 Simple Hypothesis Detection Theory. The problem of Section 1.2.1 has four possible decision/hypothesis combinations. Since a decision is forced, the probabilities of two of these (usually the "detection probability" $P(D_1|H_1)$ and the "false-alarm probability" $P(D_1|H_0)$) are sufficient to characterize the performance of any detection scheme.

It is a basic fact of classical detection theory (see [18], [43], [39]) that optimal processing of the observation consists

of computing a one-dimensional statistic (the likelihood ratio). This is trivial to demonstrate for a simple goal functional; suppose one desires to maximize

$$G = P(D_1|H_1) - \alpha\, P(D_1|H_0) \quad , \quad \alpha \geq 0 \tag{1.12}$$

Many linear goal functionals including those commonly referred to as "Bayes' criteria," can be written thus. Clearly,

$$\begin{aligned} G &= \int_{\mathcal{Y}} P(D_1|y)\, d\, \mathcal{P}_1(y) - \alpha \int_{\mathcal{Y}} P(D_1|y)\, d\, \mathcal{P}_0(y) \\ &= \int_{\mathcal{Y}} P(D_1|y) \left[\frac{d\, \mathcal{P}_1(y)}{d\, \mathcal{P}_0(y)} - \alpha \right] d\, \mathcal{P}_0(y) \end{aligned} \tag{1.13}$$

where $\mathcal{P}_1(y)$ and $\mathcal{P}_0(y)$ are probability measures on $\mathcal{Y}$ and $P(D_1|y)$ is a randomized decision rule. If the Radon-Nikodym derivative (likelihood ratio)

$$\begin{aligned} \ell(y) &\stackrel{d}{=} \frac{d\, \mathcal{P}_1(y)}{d\, \mathcal{P}_0(y)} \\ &= \frac{f(y|H_1)}{f(y|H_0)} \qquad \text{if densities exist} \end{aligned} \tag{1.14}$$

makes sense, then (1. 1. 3) is maximized by choosing the decision rule as follows:

$$P(D_1|y) = \begin{cases} 1 & \ell(y) > \alpha \\ r \in [0, 1] & \ell(y) = \alpha \\ 0 & \ell(y) < \alpha \end{cases} \tag{1.15}$$

If y is finite-dimensional and densities exist as in (1. 14), then application of these concepts is straightforward; more general situations, however, can become quite complex.

Birdsall [5] has demonstrated that the likelihood ratio (1. 14) is an optimal decision statistic, and that the decision rule has the form of (1. 15), for any goal function possessing the properties set forth in Section 1. 2. 1; only the threshold α depends upon the actual criterion. This is a powerful result since it permits design of a detector which will be optimal for a very wide class of criteria.

In infinite-dimensional spaces, the situation is much more difficult because p. d. f. 's fail to exist. Most of the available theory pertains to Gaussian processes, in which case the mean and covariance functions constitute a complete statistical description. One basic result concerns the detection of a signal known exactly

in known Gaussian noise; it is obtained as follows.[1] Suppose

$$H_0 : y_t = n_t$$
$$H_1 : y_t = s(t) + n_t \quad , \quad t \in [0, T] \tag{1.16}$$

where $s(t)$ is a sure signal and n_t is zero-mean stationary Gaussian noise with autocorrelation function $R_n(\tau)$. Let $\{\lambda_k; k = 1, 2, \ldots\}$ be the eigenvalues of the kernel $R_n(t - s)$ and expand y_t, $s(t)$, and n_t in terms of the corresponding normalized eigenfunctions $\{\varphi_k(t)\}$; if n_t has rational spectral density, this is a complete orthonormal (c. o. n.) set of functions in $L_2[0, T]$, the space of all square-integrable functions.[2]

Let $\{y_i\}$, $\{s_i\}$, and $\{n_i\}$ be the corresponding sets of (possibly random) Fourier coefficients, e. g.,

$$n_i = \int_0^T n_t \varphi_i(t)\, dt \tag{1.17}$$

---

[1] See Grenander [18] or Davenport and Root [9], Art. 14. 5.

[2] Davenport and Root [9], Appendix 2.

By the Karhunen-Loeve theorem, $\{n_i/\sqrt{\lambda_i}\}$ is a c. o. n. set of random variables; since Gaussian, they are also statistically independent. Let $\underline{Y}_k$ denote the finite collection $(y_1, \ldots, y_k)$. It is trivial to verify that $z(\underline{Y}_k)$, the natural logarithm of $\ell(\underline{Y}_k)$, is given as

$$z(\underline{Y}_k) = \sum_{i=1}^{k} \frac{y_i s_i}{\lambda_i} - \tfrac{1}{2} \sum_{i=1}^{k} \frac{s_i^2}{\lambda_i} \qquad (1.18)$$

Here the first term represents the necessary signal processing and the second is a measure of the performance of the detector:

$$\sum_{i=1}^{k} \frac{s_i^2}{\lambda_i} = d \qquad (1.19)$$

where d is the "detectability index".[1] For the general Gaussian

---

[1] Van Trees [61], p. 100 or Peterson, Birdsall, and Fox [43].

problem (i.e., z(y) is a Gaussian random variable under either hypothesis), it is always true that

$$d = E[\, z(y) | H_1 \,] - E[\, z(y) | H_0 \,] \tag{1.20}$$

where $E(\cdot)$ denotes mathematical expectation. Under suitable regularity conditions,[1] the sums in (1.18) converge in L_2 and

$$z(y) = \int_0^T y_t \, s_2(t) \, dt - \tfrac{1}{2} \int_0^T s_1^2(t) \, dt \tag{1.21}$$

where

$$s_2(t) = \sum_{i=1}^{\infty} \frac{s_i}{\lambda_i} \varphi_i(t) \tag{1.22}$$

$$s_1(t) = \sum_{i=1}^{\infty} \frac{s_i}{\sqrt{\lambda_i}} \varphi_i(t) \tag{1.23}$$

---

[1] It is necessary that $\sum \frac{s_i^2}{\lambda_i^2}$ converge; see e.g., Kelly, Reed, and Root [32].

It is easily verified using (1.22) that the function $s_2(t)$ is the solution to a Fredholm integral equation of the first kind,

$$s(t) = \int_0^T s_2(\lambda)\, R_n(t - \lambda)\, d\lambda \quad , \quad 0 \leq t \leq T \qquad (1.24)$$

and in fact, most classical solutions solve (1.24) using an associated differential equation with suitable boundary conditions;[1] the necessary signal processing is then given by the first part of (1.21).

A model first suggested by K. Metzger [38] and extensively refined by T. Birdsall allows finding $s_2(t)$ and the quadratic content of $s_1(t)$ through methods more familiar to the engineer. It applies to cases where n_t has a rational spectral density: in engineering terms, n_t can be represented as "white Gaussian noise" filtered by a causal, linear, time-invariant system with rational transfer function $H(s)$. Details of the model are derived in Appendix A. As an example of its application, Appendix A

---

[1] For example, Van Trees [61], Art. 4.3.6 or Helstrom [26], Art. IV.5.

solves the problem of detecting an exactly-known sure signal in additive M-SAG noise; for $M = 1$, one obtains the solution to Example 1.1:

$$z(y) = a^{-2}\left\{ \int_0^T [p_1^2\, s(t) - s''(t)]\, y_t\, dt + [p_1 s(0) - s'(0)]\, y_0 + [p_1 s(T) + s'(T)\, y_T] \right\} - \frac{d}{2} \tag{1.25}$$

where d is the detectability index

$$d = a^{-2}\left\{ p_1^2 \int_0^T s^2(t)\, dt + p_1[s^2(T) + s^2(0)] + \int_0^T [s'(t)]^2\, dt \right\} \tag{1.26}$$

Note that the last term of (1.25) does not involve the observation y_t; in most classical solutions this term is either omitted and considered as part of the likelihood-ratio threshold or is kept separate as a measure of performance.[1] [2] For reasons which

---

[1] Van Trees [61], p. 318.

[2] Helstrom [26], p. 136.

will be apparent in Chapter V, it is here retained as an explicit term in the likelihood ratio.

It is well known that sample functions of the process n_t, and hence y_t, are continuous with probability 1. Thus, the likelihood ratio (1.25) is the same whether one considers $\{y_t, t\epsilon[0,T]\}$ as done here, or $\{y_t, t\epsilon(0,T]\}$.

1.3.2 Bayesian Estimation Theory.

a. Bayes' Rule. Consider the problem of Section 1.2.2 for sequential discrete observations. Suppose $\underline{Y}_k$ has been observed. Given the a priori p.d.f. $f_0(\theta)$, one can use the information in the observation to form an a posteriori p.d.f. $f(\theta|\underline{Y}_k)$ by using Bayes' rule, which is merely a restatement of the definition of conditional p.d.f.'s:

$$f(\theta|\underline{Y}_k) = \frac{f(\underline{Y}_k|\theta)f_0(\theta)}{f(\underline{Y}_k)} \tag{1.27}$$

Once an observation is made, the denominator is a number which serves to normalize the a posteriori p.d.f.,

$$K(\underline{Y}_k) = f^{-1}(\underline{Y}_k) = \left[\int_{\Theta} f(\underline{Y}_k | \theta) f_0(\theta) \, d\theta \right]^{-1} \tag{1.28}$$

Prior to any observations, $\underline{Y}_k$ is a random variable and its marginal p.d.f. may be written

$$f(\underline{Y}_k) = \frac{f_0(\theta)}{f(\theta | \underline{Y}_k)} \; f(\underline{Y}_k | \theta) \tag{1.29}$$

As an incidental observation, suppose one defines

$$\Lambda_k(\underline{Y}_k;\theta) \stackrel{d}{=} \frac{f(\underline{Y}_k | \theta)}{f(\underline{Y}_k | \theta_0)} \tag{1.30}$$

where θ_0 is a fixed value in Θ such that the expression exists. Observe that Bayes' rule (1.27) holds if $\Lambda_k(\underline{Y}_k;\theta)$ replaces $f(\underline{Y}_k | \theta)$, since the denominator of (1.30) cancels out of (1.27) and (1.28). The function $\Lambda_k(\cdot)$ is called the <u>likelihood ratio function</u> and will play a significant role in later chapters.

Equations (1.27) and (1.29) can be put into sequential form; using (1.2) in (1.27) gives

$$f(\theta \mid \underline{Y}_k) = \frac{f_0(\theta) \prod_{i=1}^{k} f(y_i \mid \underline{Y}_{i-1}, \theta)}{\int_{\Theta} (\text{Numerator})\, d\theta} \tag{1.31}$$

Separating the first factor, one finds

$$f(\theta \mid \underline{Y}_k) = \frac{f(y_k \mid \underline{Y}_{k-1}, \theta)\, f(\theta \mid \underline{Y}_{k-1})}{\int_{\Theta} (\text{Numerator})\, d\theta} \tag{1.32}$$

Comparison with (1.27) shows that, at each stage, the previous a posteriori p.d.f. should be used as a prior for the next increment. A sequential analog to (1.29) follows similarly:

$$f(y_k \mid \underline{Y}_{k-1}) = \frac{f(\theta \mid \underline{Y}_{k-1})}{f(\theta \mid \underline{Y}_k)}\, f(y_k \mid \underline{Y}_{k-1}, \theta) \tag{1.33}$$

For the reasons given in Section 1.1, all this is in general

intractable: The hard memory (that which retains information which stays fixed throughout the problem) must store arbitrarily many different conditional p.d.f.'s $f(y_k | \underline{Y}_{k-1}, \theta)$, and the soft memory (that which changes as θ is learned) must update and retain the a posteriori p.d.f. as a function. Further, all previous observations must be retained in order to select the proper conditional p.d.f.

If the observations are M^{th}-order Markov with stationary transitions, (see Section 1.1, Eq. (1.3) ff.), things improve markedly. Equation (1.31) becomes

$$f(\theta | \underline{Y}_k, \underline{y}_0) = \frac{f_0(\theta) \prod_{i=1}^{k} f(y_i | \underline{y}_{i-1}, \theta)}{\int_{\Theta} (\text{Numerator})\, d\theta} \qquad (1.34)$$

and most of the hard-memory problems have been alleviated.

Under quite general conditions, the a posteriori p.d.f. defined by (1.27) or its equivalents approaches a delta-function about the "true" value θ^* as $k \to \infty$; this has been studied, for example, by Liporace [37] and Le Cam [34]. Note that the conditional p.d.f. $f(\underline{Y}_k | \theta)$ is precisely the classical likelihood function altered through multiplication by the a priori p.d.f. If the

a priori p.d.f. is nonzero in a neighborhood of θ^*, one expects that the maximum likelihood and Bayes estimators will have similar asymptotic properties; this has been verified by Levin and Shinakov [36] for a large class of problems.

b. Bayesian Estimation. Suppose that the a posteriori p.d.f. has been formed and that a cost functional $J[\hat{\theta}(y), \theta^*]$ is given ($\hat{\theta}$ is an estimate of the true value θ^*); the object is to find a $\hat{\theta}(y)$ which minimizes the risk $\mathscr{R}$, defined as the mathematical expectation of the cost. Conditional to θ^*, this expectation is

$$E_Y \left[J(\hat{\theta}, \theta^*) | \theta^* \right] = \int J[\hat{\theta}(y), \theta^*] \, f(y|\theta^*) \, dy$$

where $\hat{\theta}(y)$ is the estimator being used. Since θ^* is not known, the risk is in fact given by averaging this over all possible θ^* using the known (a priori) distribution on θ :

$$\mathscr{R} = \int_\Theta f_0(\theta) \int J[\hat{\theta}(y), \theta] \, f(y|\theta) \, dy \, d\theta$$

Note from (1.27) that, upon interchanging the order of integration, this can be written

$$\mathscr{R} = \int_y f(y) \int_\Theta J[\hat{\theta}(y), \theta] \, f(\theta|y) \, d\theta \, dy \qquad (1.35)$$

Since $f(y) \geq 0$, it is sufficient to choose $\hat{\theta}(y)$ in a way which minimizes the a posteriori expected cost; for a large class of cost functionals this estimate is the mean of the a posteriori density $f(\theta | y)$.[1] If a cost functional is not given, the estimate is often chosen to be the mode of $f(\theta | y)$; this is called maximum a posteriori (MAP) estimation.

For purposes of this dissertation, the Bayesian estimation problem is considered solved when the a posteriori density is found, and no specific cost functionals will be considered.

A final concern is to determine if and when one can sample a continuous observation as discussed in Section 1. 1 and define an a posteriori p. d. f. (and hence a Bayes' estimate) by taking a limit as the samples grow dense in the observation interval. This question will be posponed until Chapter II (Section 2. 2).

1. 3. 3 Compound-Hypothesis Detection Theory. The problem was stated in Section 1. 2. 3; one desires a decision to maximize $G[P(D_1|H_1), P(D_1|H_0)]$ where $G(\cdot, \cdot)$ satisfies the consistency conditions stated in Section 1. 2. 1 and where the detection and false-alarm probabilities are marginal probabilities as defined

---

[1] Van Trees [61], pp. 60-63.

in (1.10) and (1.11).[1] For discrete observations it is easily shown that the optimal decision is based on the likelihood ratio of marginal p.d.f.'s

$$\ell(\underline{Y}_k) = \frac{f(\underline{Y}_k|H_1)}{f(\underline{Y}_k|H_0)} = \frac{\int_{\Theta} f(\underline{Y}_k|H_1,\theta)\, f_0(\theta)\, d\theta}{\int_{\mathcal{N}} f(\underline{Y}_k|H_0,\eta)\, f_0(\eta)\, d\eta} \tag{1.36}$$

If hypothesis H_0 is simple this reduces to the well-known result

$$\ell(\underline{Y}_k) = \int_{\Theta} \ell(\underline{Y}_k|\theta)\, f_0(\theta)\, d\theta \tag{1.37}$$

---

[1] It should be noted that defining the goal functional in this way sidesteps one of the primary concerns of statistical decision theory, namely the existence of a uniformly most powerful test; see, e.g., Van Trees [61], p. 85 or Lehmann [35], Ch. 3. This is possible only because θ is considered a random variable. (See also Appendix C).

Section 1.2.3 claimed that using the Bayesian approach results in a natural relation between detection and estimation which allows the two to be done concurrently. This is dramatically illustrated by using (1.29) to write the marginal densities in the likelihood ratio (1.36):

$$\ell(\underline{Y}_k) = \frac{f_0(\theta)}{f_0(\eta)} \frac{f(\eta|\underline{Y}_k, H_0)}{f(\theta|\underline{Y}_k, H_1)} \ell(\underline{Y}_k|\theta, \eta) \qquad (1.38)$$

where

$$\ell(\underline{Y}_k|\theta, \eta) \stackrel{d}{=} \frac{f(\underline{Y}_k|\theta, H_1)}{f(\underline{Y}_k|\eta, H_0)} \qquad (1.39)$$

Note that θ and η must "cancel out" of the right side of (1.38), since the left side does not depend on the parameters. This implies that <u>any fixed, convenient, admissible value of the parameters may be used to evaluate the expression</u>, and $\ell(\underline{Y}_k|\theta, \eta)$ is merely the simple-hypothesis likelihood ratio for those fixed values. Thus, the compound-hypothesis detection problem can be solved using the

simple-hypothesis result, modified by the a posteriori p.d.f.'s which are found in the Bayesian estimation problem. Equation (1.38) will be a basic relation in the work which follows.

A sequential analog to (1.38) is again easily found; it is given by

$$\ell(\underline{Y}_k) = \prod_{i=1}^{k} \ell(y_i | \underline{Y}_{i-1}) \tag{1.40}$$

where

$$\ell(y_i | \underline{Y}_{i-1}) = \frac{f(\theta | \underline{Y}_{i-1}, H_1)}{f(\theta | \underline{Y}_i, H_1)} \frac{f(\eta | \underline{Y}_i, H_0)}{f(\eta | \underline{Y}_{i-1}, H_0)} \ell(y_i | \underline{Y}_{i-1}, \theta, \eta) \tag{1.41}$$

1.4 Historical Background

The "classical" theory of signal detection, based primarily upon the statistical theory of hypothesis testing, was developed in the early 1950's; basic papers in the field are those of Grenander [18], Peterson, Birdsall, and Fox [43], and Middleton and Van Meter [39]. This early work was primarily concerned with detection; although the presence of unknown parameters in the signal

was addressed, no attempt was made to simultaneously estimate these.

The earliest treatments of simultaneous estimation and detection usually proposed suboptimal solutions based upon using a maximum likelihood estimate of the parameters in a detection scheme; see, e.g., Kelly, Reed, and Root [32] or Price [45]. Only quite recently has the simultaneous optimality of detection and estimation schemes been addressed; even then, this has usually been done in a limited context which tends to obscure the overall simplicity of the problem.

For instance, Middleton and Esposito [40] developed a Bayesian solution for signal parameters, but from a nonrecursive viewpoint and through use of a cost structure which explicitly considers coupling of the detection and estimation costs. Scharf and Lytle [54] address Gaussian noise of unknown level, thus including noise parameters in the problem. Their solution is nonrecursive, and investigates the existence of uniformly most powerful (UMP) or UMP-invariant tests; this question can be avoided by adoption of a Bayesian approach. Spooner [57] [58] considered unknown noise parameters in detail; his work falls closest to that presented here, being a specific application of the principles involved. Jaffer and Gupta [29] [30] consider the recursive Bayesian problem

using a quadratic cost, Gaussian Markov processes, and estimating only signal parameters. Roberts [47] [48] considered signals of unknown phase or amplitude. Though much more specialized, his work also is quite close to the results given here; he appears to have been the first to employ reproducing probability densities (a phenomenon he called "closure of the distribution") for the uncertain parameters. Spooner [58] used the same technique to estimate the spectral density of sampled white noise; the results were subsequently generalized by Birdsall [6], whose work provided much of the motivation for Chapter III of this dissertation.

Many other pertinent papers exist in the recent literature; among them, the multiple-hypothesis work of Fredriksen, Middleton, and Vandelinde [17] and the paper by Nahi [41]. A good discussion of earlier papers on the problem is contained in [40].

CHAPTER II

NECESSARY AND SUFFICIENT STATISTICS

Most results in the sequel depend on the existence of sufficient statistics for the parameter of a family of probability distributions; this chapter is dedicated to the presentation of results concerning such statistics and the families of distributions which admit them. The emphasis is different from that usually found in treatments of the subject; for a more conventional approach, the reader is referred to Ferguson [16] Ch. 3, or Hogg and Craig [28] Ch. 8. A rigorous, measure-theoretic treatment based on the work of Halmos and Savage [24] and Bahadur [3] can be found in Appendix B and is necessary to a full understanding of Section 2. 2. 2 and portions of Chapters III and V. The mathematics involved in that treatment are sufficiently difficult to be incompatible with the rest of this work; hence, the results were relegated to an appendix. All results in Sections 2. 1 and 2. 3 can be derived from the work in Appendix B; in keeping with the spirit of this dissertation, they are here developed independently at a lower level of rigor.

Section 2. 1 will introduce the concept of necessary and sufficient statistics by considering them in the context of finite parameter and observation spaces (which means that all probabilities are discrete). In addition to motivating the various

definitions, this discussion forms a good heuristic basis for the measure-theoretic work in Appendix B. The section then proceeds to formally define necessary and sufficient statistics (using the factorization criterion, since all spaces are assumed finite-dimensional and all probability measures absolutely continuous), and concludes by demonstrating a function which, under suitable restrictions, always satisfies the definitions.

Section 2.2 extends the results to continuous-parameter random processes. This is done by assuming such a process to be observed on a fixed, finite interval and sampled at evenly-spaced points on that interval. The samples are made to grow dense, and limits for the sufficient statistics and the a posteriori p.d.f. generated by Bayes' rule are investigated.

Finally, Section 2.3 assumes that observations are generated sequentially by some process which causes them to possess an Mth-order Markov dependence. This generalizes, and includes as a special case, the situation where samples are independent. Under certain restrictions, only exponential families of distributions will be seen to admit sufficient statistics of fixed, finite dimension.(1)

---

(1)This was first proved for the case of independent samples by Koopman [33].

The general theory developed in Chapter III will depend primarily upon the contents of Sections 2.1 and 2.2. Section 2.3 treats a special class of processes and is presented here for reference in Chapter IV where the Mth-order Gaussian Markov process will be considered.

2.1 Definition and General Results for Finite-Dimensional Observations.

Recall that $y \in \mathcal{Y}$ is observed; for notational simplicity, $\mathcal{Y}$ is considered a domain in R^1. If k sequential observations are made, their ordered column vector is denoted $\underline{Y}_k \in \mathcal{Y}^k \subset R^k$. Suppose that $P_{k,\theta} \sim \{f(\underline{Y}_k|\theta); \theta \in \Theta \subset R^m\}$ is a known family of Borel probability measures given by the indicated piecewise-smooth p.d.f.'s on $\mathcal{Y}^k$. (These results are generalized in Section 2.2 and Appendix B.) Occasionally in the sequel, it will be assumed that each $f(\underline{Y}_k|\theta)$ is strictly positive on $\mathcal{Y}^k$; this will be referred to as the regular case. The assumption, when made, simplifies results significantly but rules out some (though not many) distributions of practical interest; e.g.,

$$f(\underline{Y}_k|\theta) = \begin{cases} \theta^{-k}; \ 0 \leq y_i \leq \theta \quad , \quad i = 1 \ldots k \\ 0 \qquad \text{elsewhere} \end{cases}$$

2.1.1 Basic Concepts. The concept of a necessary and sufficient statistic is easy when stated with enough generality: Suppose the observations $\underline{Y}_k$ have been made. A mapping $t(\underline{Y}_k)$ is a sufficient statistic if its value contains, in a sense to be defined, as much information about the true value of θ as did $\underline{Y}_k$ itself; a mapping $t(\underline{Y}_k)$ is a necessary statistic for θ if none of the information it contains about θ is redundant.[1] The definitions in existence all seek to express these concepts in various forms and degrees of generality.

Usually the task is complicated by the necessity of admitting the non-Bayesian, so that θ may not be considered a random variable. That is not a problem here; accordingly, $t(\underline{Y}_k)$ will be considered a sufficient statistic (of a sample of size k, for the parameter θ) if

$$f(\theta \mid \underline{Y}_k) = f[\theta \mid t(\underline{Y}_k)] \qquad \forall \underline{Y}_k \in \mathcal{Y}^k \tag{2.1}$$

where the a posteriori p.d.f.'s are given by Bayes' rule (1.27).

The following discussion, due to Dynkin [13], makes the

---

[1] The sense in which this may be interpreted is a subject of continuing debate by mathematical statisticians. For a good discussion, see the conclusion of the paper by Halmos and Savage [24].

results in the sequel heuristically clear. Suppose that all probability distributions are discrete and finite, i.e., one has a finite set of random outcomes

$$\mathcal{Y} = \{y_1, y_2, \cdots y_N\}$$

On these observations define a finite family of probability distributions, each member of which is

$$\{P(y_1|\theta), \ldots, P(y_N|\theta)\}$$

where $\theta \in \Theta = \{\theta_1, \ldots \theta_s\}$, a finite parameter set. Given an a priori distribution $\{P(\theta = \theta_i) = p_i\}_{i=1 \ldots s}$ on Θ, and assuming an outcome y was observed, one uses Bayes' rule

$$P(\theta_i|y) = \frac{p_i P(y|\theta_i)}{\sum_{k=1}^{s} p_k P(y|\theta_k)}, \quad i = 1 \ldots s \qquad (2.2)$$

to construct the a posteriori distribution. Now suppose that for some $y', y'' \in \mathcal{Y}$,

$$P(y'|\theta_i) = P(y''|\theta_i), \quad i = 1 \ldots s \qquad (2.3)$$

Then from (2.2),

$$P(\theta_i|y') = P(\theta_i|y'') \quad i = 1 \ldots s \tag{2.4}$$

and the outcomes $\{y', y''\}$ are equivalent for constructing the a posteriori distribution on Θ. Use (2.3) to define a class of sets on $\mathcal{Y}$:

$$B_k = \{y : P(y|\theta_i) = \gamma_k, \ i = 1 \ldots s\}$$

The class $\{B_k\}$ is a sufficient class of sets for θ, since one need only know the set into which a particular observation fell in order to construct the a posteriori distribution on Θ. A mapping $t : \mathcal{Y} \to \mathcal{T}$ which is constant on the sets B_k is a sufficient statistic for θ ;[1] clearly, $P(\theta_i|y) = P(\theta_i|t(y))$, $i = 1 \ldots s$.

It is not clear that (2.3) generates the coarsest class which is sufficient to construct $\{P(\theta_i|y); i = 1 \ldots s\}$. In fact, it does not in general. Suppose $\theta_0 \in \Theta$ is a value of the parameter for which $P(y_i|\theta_0) > 0$, $i = 1 \ldots N$; define the probability ratios

---

[1] In Appendix B, $\{B_\gamma\} = \mathcal{A}_0$ will be called a "sufficient sub-σ-algebra" for $\{\mathcal{P}_\theta\}$, and $t(\cdot)$ is a statistic which generates $\mathcal{A}_0$.

$$\Lambda(y_j|\theta_i) = \frac{P(y_j|\theta_i)}{P(y_j|\theta_0)} \quad , \quad j = 1 \ldots N, \; i = 1 \ldots s \qquad (2.5)$$

Bayes' rule (2.2) is unchanged if all $P(y|\theta_j)$ are replaced by $\Lambda(y|\theta_j)$. Hence, the class of sets $\{A_k\}$,

$$A_k = \{ y : \Lambda(y|\theta_i) = \lambda_k \, , \; i = 1 \ldots s \} \qquad (2.6)$$

is also sufficient for θ.[1] It can be shown that $\{A_k\}$ is the coarsest (i.e., the minimal) class which is sufficient; accordingly, it is called the <u>necessary and sufficient class of sets</u> for θ, and a mapping T which is constant on $\{A_k\}$ is a <u>necessary and sufficient statistic.</u>

Now let $\{A_k\}$ be an <u>arbitrary</u> sufficient class for θ and let t(y) be a sufficient statistic which is constant on members of this class, say $t(y) = t_k$ for $y \in A_k$. By assumption, then,

---

[1] It is clear that a sufficient class of sets also results if the denominator of the ratios (2.5) is any arbitrary probability distribution on $\mathcal{Y}$ which does not depend on θ; this class will not, however, necessarily be minimal.

$$P(\theta_i | y) = P(\theta_i | t(y)) \tag{2.7}$$

for $i = 1 \ldots s$. Let $\{p_i, i = 1 \ldots s\}$ be a known assignment of a priori probabilities. It follows directly from the definition of conditional probabilities that

$$p_i P(y | \theta_i) = P(\theta_i | y) P(y)$$

Hence,

$$P(y | \theta_i) = g[t(y), \theta_i] G(y) \tag{2.8}$$

where, from (2.7),

$$g[t(y), \theta_i] = \frac{P(\theta_i | t(y))}{p_i}$$

depends on y only through $t(\cdot)$, and where

$$G(y) = \sum_{r=1}^{s} p_r P(y | \theta_r)$$

does not depend on θ_i. The converse is trivial to verify: Using Bayes' rule, any family satisfying (2.8) satisfies (2.7). Thus, the factorization (2.8) provides an alternate characterization of a

sufficient statistic.

Finally, compute the conditional probabilities:

$$P(y|t(y) = t_k, \theta_i) = \begin{cases} \dfrac{P(y|\theta_i)}{\sum_{y_r \in A_k} P(y_r|\theta_i)} & \text{if} \quad y \in A_k \\ 0 & \text{otherwise} \end{cases}$$

Using (2.8) and the fact that $t(y) = t_k$ $\forall y \in A_k$,

$$P(y|t_k, \theta_i) = \begin{cases} 1/n_k & \text{if} \quad y \in A_k \\ 0 & \text{otherwise} \end{cases} \tag{2.9}$$

where n_k is the number of observations in A_k . This motivates yet a third characterization of sufficient statistics; namely, $t(y)$ is a sufficient statistic for θ if the conditional distribution $P(y|t, \theta)$ does not depend on θ .

This completes the elementary discussion using finite spaces.

2.1.2 Some General Results. For this and the next section,

the factorization criterion will serve as a useful characterization for sufficient statistics:

DEFINITION 2.1

A mapping $t_k : \mathcal{Y}^k \to \mathcal{T}_k$, $t_k = t_k(\underline{Y}_k)$, is said to be a sufficient statistic of a sample of size k, for the family $\{f(\underline{Y}_k|\theta) ; \theta \in \Theta\}$ if there exist:

-- A nonnegative function $g[t_k(\underline{Y}_k);\theta]$ which depends on $\underline{Y}_k$ only through $t_k(\cdot)$, and

-- A function $G(\underline{Y}_k)$ which does not depend on θ, such that

$$f(\underline{Y}_k|\theta) = g[t_k(\underline{Y}_k);\theta]G(\underline{Y}_k) \quad \blacksquare \tag{2.10}$$

The value of a statistic which satisfies (2.10) is clearly sufficient to evaluate the a posteriori p.d.f. via Bayes' rule; using (2.10) in (1.27) and (1.28),

$$f(\theta|\underline{Y}_k) = \frac{g[t_k(\underline{Y}_k);\theta]f_0(\theta)}{\int_\Theta (\text{Numerator})\, d\theta} \tag{2.11}$$

$$= f(\theta|t_k(\underline{Y}_k))$$

The set $\mathcal{T}_k$ is purposely unspecified as yet; if it has enough structure and dimension less than k, then a clear saving in "soft memory" requirements results from using $t_k(\underline{Y}_k)$ rather than $\underline{Y}_k$ to estimate θ.

DEFINITION 2.2

A system of sufficient statistics $\{ t_k : \mathcal{Y}^k \to \mathcal{T}_k \}_{k=1,2,\ldots}$ is said to be of fixed dimension r if, for any k, elements of the set $\mathcal{T}_k$ can be indexed by a parameter of dimension r. ∎

All sufficient statistics of practical interest will be of fixed dimension; further, it will be possible to update them sequentially: $t_{k+1}(\underline{Y}_{k+1})$ can be evaluated using only $t_k(\underline{Y}_k)$ and y_{k+1}. These two properties, and the fact that sufficient statistics exist in many problems of practical interest, lend significance to their study.

So far, nothing has been said about necessary statistics. To discuss the topic without use of measure theory, the following concept is needed:[1]

DEFINITION 2.3

Let $s_1(\cdot)$, $s_2(\cdot)$ be any mappings defined on a set $\mathcal{X}$.

---

[1] The following definitions should be compared with Definition B.2 of Appendix B.

Then one says

s_1 is dependent on s_2 if $s_2(x') = s_2(x'')$ implies $s_1(x') = s_1(x'')$.

s_1 is equivalent to s_2 if each is dependent on the other.

s_1 is trivial if it is equivalent to $I(x)$, the identity function on $\mathscr{X}$. ■

DEFINITION 2.4

$t_k : \mathscr{Y}^k \to \mathscr{T}_k$ is a necessary statistic of a sample of size k for the family $\{f(\underline{Y}_k | \theta ; \theta \in \Theta\}$ if it is dependent on every sufficient statistic. ■

The following facts follow trivially from the definitions:

a. The elementary necessary statistic is a constant on $\mathscr{Y}^k$.

b. The elementary sufficient statistic is $\underline{Y}_k$.

c. All necessary and sufficient statistics are equivalent.

d. T_k is a sufficient statistic for θ , T_k is dependent on $S_k \Rightarrow S_k$ is a sufficient statistic for θ .

e. T_k is a necessary statistic for θ , S_k is dependent on $T_k \Rightarrow S_k$ is a necessary statistic for θ .

The definitions agree with the concepts present in Section 2.1.1. An elementary discussion using classes of sets is not possible here because $\mathscr{Y}^k$ is not finite, or even countable. Thus, it is necessary to resort to the properties of functions to describe the desired concept.

DEFINITION 2.5

Let $\theta_0 \in \Theta$ be a fixed value of the parameter for the regular family of p.d.f.'s $\{f(\underline{Y}_k|\theta); \theta \in \Theta\}$.(1) The likelihood ratio function of a sample of size k, $\Lambda_k : \mathcal{Y}^k \times \Theta \to R^+$, is defined as(2)

$$\Lambda_k[\underline{Y}_k, \theta] \stackrel{d}{=} \frac{f(\underline{Y}_k|\theta)}{f(\underline{Y}_k|\theta_0)} \quad \blacksquare \tag{2.12}$$

Recall from the comment made following (1.30) that Bayes' rule can be written

$$f(\theta|\underline{Y}_k) = \frac{\Lambda_k(\underline{Y}_k, \theta) f_0(\theta)}{\int_\Theta (\text{Numerator})\, d\theta} \tag{2.13}$$

Thus, if $t_k(\cdot)$ is a sufficient statistic for θ, it follows (and is an immediate consequence of the factorization criterion (2.10)) that

---

(1) "Regular" means that each member of the family is strictly positive on $\mathcal{Y}^k$; see Section 2.1.

(2) Compare this with (2.5).

$$\Lambda_k[\underline{Y}_k, \theta] = \frac{g[t(\underline{Y}_k);\theta]}{g[t(\underline{Y}_k);\theta_0]} \qquad (2.14)$$

The significance of the likelihood ratio function lies in the following result:

THEOREM 2.1[1]

The likelihood ratio function, considered as a mapping from $\mathcal{Y}^k$ into piecewise smooth, positive functions on Θ, is a necessary and sufficient statistic of a sample of size k for θ.

Proof:

Sufficiency follows by rearranging (2.12); necessity from (2.14), whence Λ_k is dependent on any sufficient statistic $t_k(\cdot)$.[2] ∎

Clearly, the same properties hold for the function

---

(1) Dynkin [13] p. 23.

(2) The remark made following (2.5) holds here also. If the denominator of the likelihood ratio function is an arbitrary p.d.f. of $\underline{Y}_k$ which does not depend on θ, then the function yields a sufficient statistic; choosing the denominator from the family $\{f(\underline{Y}_k|\theta)\}$ yields a necessary and sufficient statistic. In fact, it is easy to see that if the denominator is a linear combination of members of the family, then one still obtains a necessary and sufficient statistic.

$$L_k(\underline{Y}_k, \theta) \stackrel{d}{=} \ell n\, \Lambda_k(\underline{Y}_k, \theta) \qquad (2.15)$$

In either case, the range $\mathcal{T}_k$ of the statistic so defined is a function space; suppose it has finite dimension r. Then the likelihood ratio function can be expressed in terms of a given basis in the space; the coefficients of such an expansion will constitute a mapping $\mathcal{Y}^k \rightarrow R^r$, and are a set of necessary and sufficient statistics in the usual sense. This will be made more precise in a later section; to do so, it is beneficial to restrict attention to a smaller class of observed processes. First, the case where $\mathcal{Y}$ is a function space will be investigated.

2.2 Continuous Parameter Processes

Suppose the observed process is $\{y_t,\ t \in [0, T]\}$, a continuous-parameter random process on a fixed finite interval whose sample functions, elements of a measure space $(\mathcal{Y}, \mathcal{A}, \mathcal{P}_\theta)$, are real-valued on $[0, T] \subset R^1$. $\mathcal{M} = \{\mathcal{P}_\theta,\ \theta \in \Theta\}$, is a family of probability measures on $(\mathcal{Y}, \mathcal{A})$. The definition of this measure space is beyond the scope of this dissertation;[1] elements of $\mathcal{Y}$ are functions, which is inconsistent with previous

---

[1] See, e.g., Doob [11] pp. 47-50, of Wong [62] pp. 37-41.

use of the symbol but should cause no confusion.

Section 2.2.1 considers that a finite number of samples are drawn from the observation, and previously established results are applied to the vector of those samples. The limiting procedure as sampling grows dense in $[0, T]$ will be discussed. Section 2.2.2 actually develops the continuous result; a working knowledge of the contents of Appendix B is presupposed. Finally, Section 2.2.3 presents a brief summary of the results, stated in terms which do not require a knowledge of measure theory.

2.2.1 Sampling the Observation. Suppose that discrete observations are generated by drawing k evenly spaced samples from a realization y_t as discussed in Section 1.1 (see Eq. (1.7) ff.).

As before, the column vector of samples is denoted

$$\underset{\sim}{Y}_k = (y_1, y_2, \cdots y_k)^*$$

In some cases (e.g., Markov processes) it may again be convenient to omit a number of "initial" samples from consideration and proceed with all expressions conditioned on these samples.[1] This

---

[1] Recall the discussion of Section 1.1.

will be done in applications which follow in later chapters, but will not be explicitly indicated here. The joint p.d.f. of k samples, $f(\underline{Y}_k|\theta)$, is usually easy to find.[1]

One wishes to investigate limiting behavior as the samples grow dense in $[0, T]$ or $(0, T]$; i.e., as $\delta_k \to 0$ or $k \to \infty$. Specifically, it is desired to investigate the limits, should they exist, of the sufficient statistics and the a posteriori p.d.f. This requires some care, since subtle mathematical singularities can occur in such a procedure;[2] mathematical rigor is indispensable, and liberal use of the measure-theoretic results of Appendix B will thus be made.

A direct attempt at finding a limiting a posteriori p.d.f. by use of Bayes' rule, say

$$f(\theta|y_t) = \lim_{k\to\infty} \frac{f(\underline{Y}_k|\theta)\, f_0(\theta)}{\int_\Theta (\text{Numerator})\, d\theta} \tag{2.16}$$

---

[1] It is, in fact, the theoretic basis for being able to define the measures $\{ \mathscr{P}_\theta \}$; see Wong [62] p. 40, or Doob [11] pp. 47-48.

[2] See, for example, Slepian [56].

is usually doomed to failure. The conditional p.d.f. $f(\underline{Y}_k|\theta)$ is defined on a space of dimension k, but p.d.f.'s in the usual sense (i.e., with respect to Lebesgue measure) only exist on finite-dimensional spaces; the problem arises because of the required normalization. Thus, the numerator of the right-hand side of (2.16) does not in general have a limit; the limit of the entire term, including denominator, may be extremely difficult to evaluate.

Recall from (2.13) that if the family of distributions is regular, Bayes' rule can be written using the likelihood ratio function $\Lambda_k(\underline{Y}_k;\theta)$ as defined in (2.12). Moreover, this function will be shown to have a limit under quite general conditions; if it does, one can define an a posteriori p.d.f. on Θ for continuous observations by

$$f(\theta|y_t) = \lim_{k\to\infty} \frac{\Lambda_k(\underline{Y}_k;\theta) f_0(\theta)}{\int_{\Theta} (\text{Numerator})\, d\theta} \qquad (2.17)$$

Suppose that the families of density functions $\{f(\underline{Y}_k|\theta),\ \theta\in\Theta\}_{k=1,2,\ldots}$ have fixed rank $r < \infty$; i.e., they admit a family of sufficient statistics $\{t_k(\cdot)\}$ of dimension r. Under suitable conditions (e.g., Section 2.3; in particular, Eq. (2.34)) the likelihood ratio function is

$$\Lambda_k(\underline{Y}_k;\theta) = \Lambda(t_k(\underline{Y}_k);\theta) \tag{2.18}$$

and it appears the sufficient statistics may themselves possess a limit. This will also be investigated.

2.2.2 Sufficient Statistics for Continuous Processes. The convergence of the likelihood-ratio function can be established using a basic result for random processes. Assume that $\mathscr{P}_{\theta_0}$ is a probability measure which dominates $\mathscr{M}$, $\mathscr{P}_\theta << \mathscr{P}_{\theta_0}$ for all $\theta \in \Theta$.(1) Suppose $(\mathscr{Y}, \mathscr{A}, \mathscr{P}_{\theta_0})$ is complete.

Let T_k denote the set of k sampling instants, see (1.7), and $P_{k,\theta}$ the Borel probability measure generated by $\{y_t, t \in T_k\}$ under $\mathscr{P}_\theta$. By assumption, $P_{k,\theta}$ is given by the p.d.f. $f(\underline{Y}_k|\theta)$ and so

$$\Lambda_k(\underline{Y}_k;\theta) = \frac{dP_{k,\theta}}{dP_{k,\theta_0}}(\underline{Y}_k) \quad \text{a.s.}[\mathscr{P}_{\theta_0}] \tag{2.19}$$

---

(1) Theorem 2.2 will hold whether or not $\mathscr{P}_{\theta_0} \in \mathscr{M}$. In practice, things are greatly simplified if $\mathscr{M}$ is an equivalent family and $\theta_0 \in \Theta$.

THEOREM 2.2[1]

Let

(i) $(\mathcal{Y}, \mathcal{A}, \mathcal{P}_\theta)$, $\mathcal{P}_\theta \in \mathcal{M}$ be as above.

(ii) $(\mathcal{Y}, \mathcal{A}, \mathcal{P}_{\theta_0})$ be complete, and $\mathcal{P}_\theta << \mathcal{P}_{\theta_0}$ $\forall \theta \in \Theta$.

(iii) The process $\{y_t, t \in [0, T]\}$ be separable and continuous in probability $[\mathcal{P}_\theta]$ for all θ .

Then

for each $\theta \in \Theta$,

$$\Lambda_k(\underline{Y}_k;\theta) = \frac{dP_{k,\theta}}{dP_{k,\theta_0}} \xrightarrow[k\to\infty]{} \frac{d\mathcal{P}_\theta}{d\mathcal{P}_{\theta_0}}(y_t) \quad \text{a.s. } [\mathcal{P}_{\theta_0}] \quad \blacksquare \tag{2.20}$$

Assumption (iii) is quite weak and, as a precept, its satisfaction is a primary property of probabilistic processes possessing prominent practicality. The requirement which is stronger and

---

(1) This result was rigorously established by Striebel [60], who obtained it as a direct consequence of Doob's convergence theorem for martingales ([12], Ch. 7, Thm. 4.1). In her work Striebel also observed that sufficient statistics can often be found by inspection of the resulting density.

may be more difficult to establish is that $\mathscr{P}_\theta << \mathscr{P}_{\theta_0}$ for all $\theta \in \Theta$. Fortunately, a large literature exists on the singularity of Gaussian measures, and Gaussian processes are of primary interest in many applications.

The convergence of (2.20) does not imply that the limit (Radon-Nikodym derivative) is easy to evaluate or even exists in closed form. The existence of sufficient statistics can simplify the matter considerably; the following theorem summarizes the results of Appendix B for the case where $\mathscr{P}_{\theta_0} \in \mathscr{M}$ dominates $\mathscr{M}$.[1]

THEOREM 2.3

Put

$$\frac{d\mathscr{P}_\theta}{d\mathscr{P}_{\theta_0}}(y_t) \triangleq \lambda(y_t ; \theta) \tag{2.21}$$

Let $A_\theta(r)$ be the inverse image of the Borel set $\lambda(\cdot\ ; \theta) < r$ in R^1 , and $\mathscr{A}^*$ the minimal σ-algebra which contains all $A_\theta(r)$, $\theta \in \Theta$, $0 < r < \infty$. The sub-algebra $\mathscr{A}^*$ is necessary and sufficient

---

[1] Recall footnote (1) following Theorem 2.1, which applies here as well.

for $\mathcal{M}$.

In particular, $t : (\mathcal{Y}, \mathcal{A}) \to (\mathcal{T}, \mathcal{B})$ is a necessary and sufficient statistic for $\mathcal{M}$ (i.e., for θ) iff there exists a family $g(\cdot\,;\theta) : (\mathcal{T}, \mathcal{B}) \to (R^+, \mathcal{R})$ such that

$$\lambda(y_t;\theta) = g[\,t(y_t);\theta\,]\; G(y_t) \quad \blacksquare \tag{2.22}$$

Here, "y_t" denotes a sample function and "$t(\cdot)$" the sufficient statistic. Since $\lambda(y_t;\theta)$ is the probability density of $\mathcal{P}_\theta$ with respect to $\mathcal{P}_{\theta_0}$, Theorem 2.3 is a generalization of the classical factorization criterion; often, the sufficient statistics can be found by inspection. The definitions of necessary and sufficient σ-algebras which were used to obtain the theorem are direct extensions of the set-theoretic notions discussed in Section 2.1.1.

2.2.3 Summary. The procedure for the continuous case is as follows. Once it has been verified that $\{y_t,\ t\in[0,\ T]\}$ is separable and continuous in probability,[1] then a suitable $\mathcal{P}_{\theta_0}$

---

[1] A separable version of any process exists; Wong [62] p. 42.

with respect to which each member of $\mathcal{M}$ is absolutely continuous must be found. If $\mathcal{P}_{\theta_0} \epsilon \mathcal{M}$ or is a linear combination of members of $\mathcal{M}$, the procedure is guaranteed to yield a necessary and sufficient statistic. If not, only sufficiency can be insured. Commonly, $\mathcal{M}$ will be an equivalent set of measures and $\theta_0 \epsilon \Theta$ chosen arbitrarily; as demonstrated in Chapter V, this has an additional advantage for certain Gaussian processes.[1]

Next, one writes the joint p.d.f. of $k < \infty$ evenly spaced samples of y_t, and forms $\Lambda_k(\underline{Y}_k;\theta)$ as in (2.12). At this point, it is helpful to recognize the sufficient statistics $t_k(\underline{Y}_k)$ if they exist, and to arrange the expression for Λ_k (usually through judicious multiplication and division by the sampling interval) so that the statistics will converge to a well-defined function of y_t.

The limit of $\Lambda_k(\underline{Y}_k;\theta)$ is evaluated,[2] and the sufficient statistics for continuous observations are recognized directly (as above) or through factoring as in Theorem 2.3. The a posteriori

---

[1] The arbitrariness of $\mathcal{P}_{\theta_0}$ may seem bothersome since it will usually introduce extraneous parameters into the resulting density. Theoretically, this is no problem since those parameters are constants and do not affect the sufficient statistics. Practically, it is also no problem since they will generally cancel in Bayes' rule (2.23).

[2] As will be seen in Chapter V, this also is not a trivial task.

p.d.f. of θ is given by (2.17), which may clearly be written

$$f(\theta \mid y_t) = \frac{\lambda(y_t;\theta) f_0(\theta)}{\int_{\Theta} (\text{Numerator})\, d\theta} \tag{2.23}$$

provided that the integral exists.

The procedure will be illustrated in Chapter V. As a final comment, it is pointed out that other procedures than the limiting described above exist for evaluating Radon-Nikodym derivatives. Any such procedure may, of course, be employed in lieu of the above.

2.3 Sequential Samples from an M^{th}-Order Markov Process

To conclude this chapter, attention is focused on a special class of discrete (i.e., $\mathcal{Y} \subset R^1$) processes. One wants to investigate the conditions under which such processes admit necessary and sufficient statistics, and what form those statistics may take.

Most sequential results available in the statistical literature address sufficient statistics in the context of independent sampling, i.e., successive observations are assumed statistically independent. Many, especially those concerning the form of distributions which admit sufficient statistics, are easily extended

to the case where the samples possess an M^{th}-Order Markov dependence as discussed in Section 1.1 (Equations (1.3) through (1.5)). Note that (1.5) is functionally similar to (1.6), the expression for the joint p.d.f. of k independent samples, except that each factor of (1.5) is a function of $M+1$ variables $\tilde{y}_i = (y_i \cdots y_{i-M})$. Many results concerning necessary and sufficient statistics are derived from the functional (as opposed to probabilistic) properties of these factors; thus, it is just as easy to use vector notation and treat the M^{th}-Order Markovian case; the "independent" result may be recovered by putting $M = O$ if necessary.

The work presented parallels Dynkin [13]; if proofs are sufficiently similar to be obvious, they will only be sketched here. Define state vectors of sequential observations as in (1.4) and also define an $(M+1)$-vector

$$\tilde{y}_i = (y_i, y_{i-1}, \cdots, y_{i-M})^* \in \mathcal{Y}^{M+1} \qquad (2.24)$$

where (*) denotes transpose. Now assume that

a. The conditional Borel probability measure of the i^{th} observation is given by one of a family of p.d.f.'s

$$\{f(y_i | y_{i-1} \cdots y_{i-M}, \theta) ; \theta \in \Theta\}$$
$$= \{f(y_i | \underline{y}_{i-1}, \theta) ; \theta \in \Theta\}$$

where θ is an m-dimensional parameter.

b. $f(y_i|\underline{y}_{i-1}, \theta)$ does not depend on i ; that is, the observations have stationary transitions.

c. $f(y|\underline{y}, \theta)$ is piecewise smooth and nonzero in, and integrable on, $\mathcal{Y} \times \mathcal{Y}^M \times \Theta = \mathcal{Y}^{M+1} \times \Theta$.

d. The value of θ stays fixed as sequential observations are made.

Some remarks are necessary: Whenever the vector $\tilde{\underline{y}}$ is used, its first component is understood to be the "present" observation y , and its other components are the previous M observations $\underline{y}$ which are needed in the transition density. To start the process, one needs $\underline{y}_0$; these will be assumed given and all expressions which follow should be interpreted as conditioned upon the "initial observation" $\underline{y}_0$ (see Section 1.1).

Henceforth, the transition density will be denoted as

$$h(\tilde{\underline{y}};\theta) \stackrel{d}{=} f(y|\underline{y}, \theta) \tag{2.25}$$

This is convenient because, as discussed, most results depend on the functional properties of $h(\tilde{\underline{y}};\theta)$ rather than its probabilistic interpretation.

Note that assumption c. implies the "regular case" : the domain in which the transition density is nonzero does not depend

on the value of θ .

If the transition density can be factored in a manner analogous to (2. 10),

$$h(\tilde{\underline{y}};\theta) = g[\, t(\tilde{\underline{y}}),\, \theta\,]\; G(\tilde{\underline{y}}) \tag{2.26}$$

then the joint density of k observations (1. 5) factors similarly. Thus, the results which follow will be based on the properties of the transition density; $t(\tilde{\underline{y}})$ above will be called a sufficient statistic of y for θ ; a necessary statistic is similarly defined (Definition 2. 4).

Consider the transition likelihood ratio function

$$\Lambda_0(\tilde{\underline{y}};\theta) = \frac{f(y|\underline{y},\theta)}{f(y|\underline{y},\theta_0)} = \frac{h(\tilde{\underline{y}};\theta)}{h(\underline{y};\theta_0)} \tag{2.27}$$

where $\theta_0 \epsilon \Theta$ is an arbitrary fixed element. Λ_0 is well-defined (see assumption c.); its k-fold product is, from (1. 5) , seen to be the likelihood ratio function $\Lambda_k(\underline{Y}_k;\theta)$. Also,

$$L(\theta;\tilde{y}) \triangleq \ell n\, \Lambda_0(\tilde{y};\theta) \qquad (2.28)$$

is well defined; its use will turn out to be convenient in what follows.

THEOREM 2.4

The mapping, $L : \mathcal{Y}^{M+1} \to \mathcal{S}$

$$L(\theta;\tilde{y}) = \ell n\, h(\tilde{y};\theta) - \ell n\, h(\tilde{y};\theta_0) \qquad (2.29)$$

is a necessary and sufficient statistic of y for θ.

Proof:

From (2.29)

$$h(\tilde{y};\theta) = h(\tilde{y};\theta_0) \exp\{L(\theta;\tilde{y})\}$$

so L is a sufficient statistic.

Let $t(\tilde{y})$ be any sufficient statistic; then (2.26) is satisfied for some functions $g(\cdot)$ and $G(\cdot)$. Using (2.29),

$$L(\theta;\tilde{y}) = \ell n\, g[t(\tilde{y});\theta] - \ell n\, g[t(\tilde{y});\theta_0]$$

which is dependent on $t(\tilde{y})$. Thus L is a necessary statistic. ∎

Note again that L maps $\mathcal{Y}^{M+1}$ into piecewise smooth functions on Θ .

Corollary

A necessary and sufficient statistic for a sample of size k , $\underline{Y}_k = \{y_1, \cdots, y_{k-1}\}$, is

$$L_k(\theta ; \underline{Y}_k) = \sum_{i=1}^{k} L(\theta ; \tilde{y}_i) \quad \blacksquare \tag{2.30}$$

This follows from Theorem 2.2 and (1.5).

THEOREM 2.5

Let

V_L denote the minimal linear space of functions, defined on $\mathcal{Y}^{M+1}$, spanned by constants and the functions $\{L(\theta ; \tilde{y}) ; \theta \in \Theta\}$.

Suppose dim $V_L = r + 1$ (possibly $r = \infty$) .

Then

a. For every finite $k \leq r$, any sufficient statistic for a sample of size k is trivial.[1]

---

[1] See Definition 2.3.

b. If the functions $\{1, \varphi_1(\tilde{y}), \ldots, \varphi_r(\tilde{y})\}$ are a basis for V_L, then $\forall k \geq r$ the r-vector of functions $\underline{t}_k(\underline{Y}_k) = [t_{ki}(\underline{Y}_k) ; i = 1 \ldots r]$, where

$$t_{ki}(\underline{Y}_k) = \varphi_i(\tilde{y}_1) + \varphi_i(\tilde{y}_2) + \ldots + \varphi_i(\tilde{y}_k) \qquad (2.31)$$

is functionally independent and forms a necessary and sufficient statistic of $\underline{Y}_k$ for θ. ∎

The function $\underline{t}_k(\underline{Y}_k)$ is a mapping from $\mathcal{Y}^k$ to R^r; it is the quantity usually written directly as a sufficient statistic. The approach here is more general and yields the concept of necessity as a side benefit.

The proof is a generalization of Dynkin's proof for independent observations ([13] p. 24) and is sketched in Appendix D. In essence, it is only necessary to replace Dynkin's observation x with the vector $\tilde{y}$ and to redefine the spaces accordingly. Care must be taken, however, to recall from (2.24) how the vectors $\tilde{y}$ were defined; for example, the pair of vectors $\{\tilde{y}_i, \tilde{y}_{i+1}\}$ has M common components and thus $\{\tilde{y}_i, \tilde{y}_{i+1}\} \in \mathcal{Y}^{M+2}$; similarly, $\{\tilde{y}_i \ldots \tilde{y}_{i+k}\} \in \mathcal{Y}^{M+1+k}$.

Theorem 2.5 inspires the following definition:

DEFINITION 2.6

The rank of the family $\{f(y_i|y_{i-1}, \theta)\ \theta\in\Theta\}$ in the domain $\mathcal{Y}$ is the greatest number r such that, for any finite $k \leq r$, there is no nontrivial sufficient statistic of a sample of size k for θ. ∎

THEOREM 2.6(a)

Suppose $\{f(y_i|y_{i-1}, \theta)\ ;\ \theta\in\Theta\}$ has finite rank r for $y\in\mathcal{Y}$. Then the transition density $f(y|\underset{\sim}{y}, \theta)$ has the form

$$h(\tilde{y};\theta) = \exp\left\{ \sum_{i=1}^{r} \varphi_i(\tilde{y})\, c_i(\theta) + c_0(\theta) + \varphi_0(\tilde{y}) \right\} \quad (2.32)$$

where the functions $\{\varphi_1(\tilde{y}) \ldots \varphi_r(\tilde{y})\}$ are piecewise smooth in $\mathcal{Y}^{M+1}$, and where the systems of functions $\{1, \varphi_1 \ldots \varphi_r\}$ and $\{1, c_1, \ldots c_r\}$ are linearly independent. ∎

THEOREM 2.6(b)

Suppose

$$f(y_i | \underline{y}_i, \theta) = \exp \left\{ \sum_{j=1}^{r} \varphi_j(\tilde{\underline{y}}_i)\, c_j(\theta) + \varphi_0(\tilde{\underline{y}}_i) + c_0(\theta) \right\}$$

$$\underline{\text{for}} \quad (y_i, \underline{y}_{i-1}) \in \mathcal{Y}^{M+1}$$

$$\theta \in \Theta \qquad (2.33)$$

Then the rank of the family of densities does not exceed r. If the systems of functions $\{1, \varphi_1 \ldots \varphi_r\}$ and $\{1, c_1, \ldots c_r\}$ are linearly independent, then the rank equals r and, for $k \geq r$,

$$\underline{t}_k(\underline{Y}_k) = \left[\sum_{j=1}^{k} \varphi_i(\tilde{\underline{y}}_j) \right]_{i=1 \ldots r}$$

is a necessary and sufficient statistic for θ. ∎

The proofs will be omitted; they are analogous to Dynkin [13] pp. 26-27, and are easy consequences of Theorem 2.5 and Equation (2.29). Only the statements concerning linear independence require some effort.

Before illustrating all this with an example, two comments

should be made:

a. Recall from Theorem 2.5 that V_L included the constant functions. In general, the constant in the likelihood ratio function for a sample of size k will explicitly depend on k which must therefore be known to have totally sufficient statistic. Although it is pedagogically questionable, k will henceforth be included as a component of $\underline{t}_k$ (even though it does not depend upon the observation) whenever the discrete case is being investigated. This clearly makes no sense if one intends only to pass to the limit $k \to \infty$.

b. If (2.32) and (2.25) are used in (2.24), it is clear that the likelihood ratio function for a sample of size k can be written

$$\Lambda_k(\underline{Y}_k, \theta) = \Lambda[\underline{t}_k(\underline{Y}_k), \theta] \tag{2.34}$$

where the functional form of Λ depends only on the transition density, and not on k.

Example 2.1

Suppose that $\{y_i\}$ is a stationary, discrete, M-th order autoregressive Gaussian process as defined in Section 4.1; the

parameter θ has components denoted α and $\underline{\beta} = (\beta_1 \cdots \beta_M)^*$, and the transition density is (see (4.16))

$$f(y_i | \underline{y}_{i-1}, \underline{y}_0, \underline{\beta}, \alpha)$$

$$= \exp \left\{ -\tfrac{1}{2} [\ell n(2\pi\alpha^2) + \frac{1}{\alpha^2} (y_i + \underline{\beta}^* \underline{y}_{i-1})^2] \right\}$$

$$= \exp \left\{ -\tfrac{1}{2} [\ell n\, 2\pi + 2\ell n\, \alpha + \frac{1}{\alpha^2} y_i^2 + \frac{1}{\alpha^2} \underline{\beta}^* y_i \underline{y}_{i-1} + \frac{1}{\alpha^2} \underline{\beta}^* \underline{y}_{i-1} \underline{y}_{i-1}^* \underline{\beta}] \right\} \tag{2.35}$$

It is convenient to retain vector-matrix notation. Define

$$\varphi_0(\tilde{\underline{y}}) = y_i^2 \tag{2.36}$$

$$\varphi_M(\tilde{\underline{y}}) = y_i \underline{y}_{i-1}$$

$$= (y_i y_{i-1}, \cdots, y_i y_{i-M})^* \tag{2.37}$$

$$\Phi_M(\tilde{\underline{y}}) = \underline{y}_{i-1} \underline{y}_{i-1}^*$$

$$= \begin{bmatrix} y_{i-1}^2 & y_{i-1} y_{i-2} \cdots y_{i-1} y_{i-M} \\ y_{i-2} y_{i-1} & y_{i-2}^2 \quad \cdots y_{i-2} y_{i-M} \\ \vdots & \vdots \quad \vdots \\ y_{i-M} y_{i-1} & y_{i-M} y_{i-2} \cdots y_{i-M}^2 \end{bmatrix} \tag{2.38}$$

Φ_M is symmetric and has $\frac{1}{2}M(M+1)$ linearly independent components; thus, there is a total of $\frac{1}{2}(M+1)(M+2)$ linearly independent statistics defined above, and a set of necessary and sufficient statistics for a sample of size k is

$$t_0(\underline{Y}_k) = \sum_{i=1}^{k} \varphi_0(\tilde{\underline{y}}_i) = \sum_{i=1}^{k} y_i^2 \tag{2.39}$$

$$\underline{t}_M(\underline{Y}_k) = \sum_{i=1}^{k} \underline{\varphi}_M(\tilde{\underline{y}}_i) = \sum_{i=1}^{k} y_i \underline{y}_{i-1} \tag{2.40}$$

$$T_M(\underline{Y}_k) = \sum_{i=1}^{k} \Phi_M(\tilde{\underline{y}}_i) = \sum_{i=1}^{k} \underline{y}_{i-1} \underline{y}_{i-1}^{*} \tag{2.41}$$

In Chapter IV, these will collectively be referred to as $t(\underline{Y}_k)$. ■

CHAPTER III

SUFFICIENT STATISTICS AND REPRODUCING DENSITIES IN SIMULTANEOUS DETECTION AND ESTIMATION

Chapter I, specifically Sections 1.3.2 and 1.3.3, presented general Bayesian solutions to the estimation and compound-hypothesis detection problems; these were given in both one-shot and sequential formulations. Section 2.1 then introduced the concept of sufficient statistics; it was shown (Eq. (2.11)) how the existence of sufficient statistics of fixed dimension could significantly reduce the amount of "soft memory" necessary to store the observation, since one need only save the (r-dimensional) updated value of $t_k(\underline{Y}_k)$, rather than saving the observations $\underline{Y}_k$ themselves, in order to estimate θ .

The definitions of Chapter II always admit the observation itself as an "elementary" sufficient statistic. To save verbiage, it will henceforth be convenient to reserve the term "sufficient statistic" for those statistics which are of fixed minimal dimension r , i.e., which are necessary and sufficient and of fixed dimension. If no such statistic exists, then the family of distributions will be said to not admit a sufficient statistic. As a further notational convenience, one subscript will henceforth be eliminated and $t_k(\underline{Y}_k)$ written as $t(\underline{Y}_k)$.

The remainder of this dissertation will deal only with classes of probability distributions which admit sufficient statistics in the sense just discussed; this chapter presents the general theory, and Chapters IV and V a specific application. Section 3.1 treats only the estimation problem, since its solution is basic to solving the detection problem (see Section 1.3.3); it will be shown that not only does existence of a sufficient statistic eliminate the need to retain all past observations, but it also obviates the need to store the a posteriori p.d.f. on Θ as a function. Instead, this p.d.f. is itself indexed by a parameter of fixed dimension, regardless of the number of observations upon which it is based.

Section 3.2 applies the results to the detection problem; finally, Section 3.3 will treat the case of continuous observation.

3.1 Bayesian Estimation

Suppose that $t(\cdot)$: $\mathcal{Y}^k \rightarrow R^r$, $t(\underline{Y}_k) = t_k$, is a sufficient statistic of fixed dimension; then by Definition 2.1,

$$f(\underline{Y}_k | \theta) = g[\, t(\underline{Y}_k), \theta \,]\; G(\underline{Y}_k) \tag{3.1}$$

for all k, and Bayes' rule can be written as in (2.11). The following example illustrates the concept and will be re-examined later in this section:

Example 3.1

Let the observations y_i be generated by

$$y_i = \theta + n_i \quad , \quad i = 1, 2, 3 \ldots$$

where θ is an unknown scalar and the n_i are independent "noise" samples from a $N(0, \sigma^2)$ distribution. Conditioned on θ, the samples are independent $N(\theta, \sigma^2)$ and their joint conditional p.d.f. can be written as follows:

$$f(\underline{Y}_k|\theta) = (2\pi\sigma^2)^{-\frac{k}{2}} \exp\left[-\frac{1}{2\sigma^2}\left(k\theta^2 - 2\theta \sum_{i=1}^{k} y_i\right)\right]$$

$$\cdot \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^{k} y_i^2\right]$$

$$= g[t(\underline{Y}_k);\theta]\, G(\underline{Y}_k) \qquad (3.2)$$

Thus, the sufficient statistic $t(\underline{Y}_k)$ consists of the sum of the observations and the total number of observations,

$$t(\underline{Y}_k) = \left[k, \sum_{i=1}^{k} y_i\right]^* \qquad (3.3)$$

where * denotes transpose. Obviously, this can be "updated" sequentially as the sequence $\{y_i\}$ is received. Including k as a "statistic" is, as discussed in the paragraph preceding (2.34), pedagogically questionable but will turn out to be a convenience. ∎

The function $g[t(\underline{Y}_k;\theta)]$ will always be assumed maximally factored (i.e., it contains no factors which do not depend on θ).

3.1.1 Natural Conjugate Densities. The sufficient statistic can serve to also characterize the a posteriori p.d.f.'s on Θ, eliminating the need to store them as functions in soft memory.

Example 3.2

Consider the situation of Example 3.1:

Suppose the a priori p.d.f. on Θ has the form

$$f_0(\theta) = K_0 \exp\left\{ -\frac{\gamma_{01}\theta^2 - 2\gamma_{02}\theta}{2\sigma^2} \right\}, \quad -\infty < \theta < \infty \tag{3.4}$$

where $\gamma_{01} > 0$. This is the class of Gaussian densities on $\Theta = R^1$; its members are indexed by the parameter $\gamma_0 = (\gamma_{01}, \gamma_{02})$. Note the similarity between (3.4) and the first term of (3.2). Now suppose $\underline{Y}_k$ is observed; applying Bayes' rule (1.27) in one-shot form yields

$$f(\theta \mid \underline{Y}_k) = K_k \exp \left\{ - \frac{(\gamma_{01}+k)\theta^2 - 2(\gamma_{02}+\sum_{i}^{k} y_i)\theta}{2\sigma^2} \right\} \qquad (3.5)$$

where all terms not involving θ are included in the normalizing constant K_k. This is the member of the class of (3.4) indexed by

$$\gamma_k = \gamma_0 + t(\underline{Y}_k) \qquad (3.6)$$

Choosing $f_0(\theta)$ in the class of (3.4) causes the a posteriori p.d.f. to remain in that class; the indexing parameter γ is updated through (3.6). Note that the updating can be done recursively. ■

The concept of a reproducing density function, illustrated above, will be defined shortly. One might wonder whether the parametrization established in (3.4) is necessary to the phenomenon; the answer is no. For example, had the a priori p.d.f. been parametrized by its mean and variance, $f_0(\theta) \sim N(m_0, d_0^2)$, then the a posteriori density would be Gaussian with

$$\begin{bmatrix} m_k \\ d_k^2 \end{bmatrix} = \frac{1}{\sigma^2 + k d_0^2} \begin{bmatrix} m_0\sigma^2 + d_0^2 \sum_{i=1}^{k} y_i \\ \sigma^2 d_0^2 \end{bmatrix} \qquad (3.7)$$

These parameters are updated using the same set of statistics (3.3), and hence can also be formed sequentially; only the "updating equations" (compare (3.6) and (3.7)) change.

DEFINITION 3.1

Let $\mathcal{H}_\Gamma(\Theta) = \{h(\theta;\gamma); \gamma \in \Gamma \subset R^m, \theta \in \Theta\}$ be a family of p.d.f.'s on Θ which is indexed by the m-dimensional parameter γ. $\mathcal{H}_\Gamma(\Theta)$ is said to be a reproducing class of probability densities under $\{f(\underline{Y}_k|\theta)\}$ if, for any k, whenever the a priori p.d.f. on Θ is

$$f_0(\theta) = h(\theta;\gamma_0) \quad , \quad \gamma_0 \in \Gamma$$

there exists a $\gamma_k = \gamma_k(\gamma_0, \underline{Y}_k) \in \Gamma$ such that the a posteriori p.d.f. is

$$f(\theta|\underline{Y}_k) = h(\theta;\gamma_k) \quad , \quad \gamma_k \in \Gamma \ \blacksquare$$

If such a class exists and is used, $f(\theta \mid \underline{Y}_k)$ need not be stored as a function; γ_k completely characterizes the a posteriori p. d. f. , and its dimension is fixed. The following theorem generalizes the result of Example 3. 2.

THEOREM 3. 1

Suppose $f(\underline{Y}_k \mid \theta)$ admits a nontrivial sufficient statistic of fixed dimension, $t(\underline{Y}_k)$, for θ and hence can be factored as in (3. 1); let the function $g(\cdot, \cdot)$ be as defined there, and, provided the integral exists, put

$$p(\theta ; \gamma) = \frac{g[\gamma, \theta]}{\int_{\Theta} g[\gamma, \theta'] \, d\theta'} , \quad \gamma \in \Gamma \tag{3.8}$$

where Γ is the image of the space of observations under $t(\cdot)$.[1]

Then $\{p(\theta ; \gamma), \gamma \in \Gamma\}$ is a reproducing class of densities under $f(\underline{Y}_k \mid \theta)$. ∎

---

[1] Γ can actually be taken to include all values of γ for which $p(\theta ; \gamma)$ retains the mathematical properties of a p. d. f. ; see Raiffa and Schlaiffer [46], p. 50.

The proof is given in Appendix D. The class thus defined is called the natural conjugate class of p.d.f.'s under $f(\underline{Y}_k|\theta)$; existence of a sufficient statistic implies existence of such a class. The parameter γ which indexes its members will be called the conjugate parameter to distinguish it from the parameter θ. The class may be quite rich,[1] or it may be restrictive and not contain a satisfactory model for the a priori p.d.f. The next section will show that this class is only a small subset of all densities which reproduce; first, some relations which apply if the a priori p.d.f. is in the natural conjugate class.

Consider one-shot processing and that $f_0(\theta) = p(\theta;\gamma_0)$, $\gamma_0 \in \Gamma$. Bayes' rule (1.27) becomes

$$p(\theta;\gamma_k) = \frac{p(\theta;\gamma_0)\, f(\underline{Y}_k|\theta)}{f(\underline{Y}_k)} \tag{3.9}$$

where

---

[1] See, for example, Howard [27].

$$f(\underline{Y}_k) = \int_{\Theta} p(\theta ;\gamma_0) f(\underline{Y}_k | \theta) d\theta \qquad (3.10)$$

Alternately, (3.9) may be rearranged to yield the analog to (1.29),

$$f(\underline{Y}_k) = \frac{p(\theta ;\gamma_0)}{p(\theta ;\gamma_k)} f(\underline{Y}_k | \theta) \qquad (3.11)$$

Since the updated conjugate parameter γ_k is found from γ_0 using the sufficient statistics (see e.g., (3.6)), this result can be valuable in forming the marginal p.d.f. of $\underline{Y}_k$.

Analogous expressions can be written for <u>sequential processing</u>. From (1.32), Bayes' rule is

$$p(\theta ;\gamma_{k+1}) = \frac{p(\theta ;\gamma_k) f(y_{k+1} | \underline{Y}_k, \theta)}{f(y_{k+1} | \underline{Y}_k)} \qquad (3.12)$$

and the marginal observation p.d.f. is

$$f(y_{k+1} | \underline{Y}_k) = \frac{p(\theta ; \gamma_k)}{p(\theta ; \gamma_{k+1})} \; f(y_{k+1} | \underline{Y}_k, \theta) \tag{3.13}$$

The comment made following (1.39), which applies to (3.11) and (3.13) as well, indicates that these equations are not in their simplest form (although they will be useful as written); indeed, one may use the factorization criterion (3.1) and the definition of $p(\theta ; \gamma)$ (3.8) to rewrite them. First, using (3.8), define the "conjugate normalizing constant" as

$$K(\gamma) = \left[\int_{\Theta} g[\gamma, \theta] \, d\theta \right]^{-1} \tag{3.14}$$

Use of the symbol $K(\cdot)$ will henceforth be consistent with this definition. Now use (3.1), (3.8), and (3.14) to rewrite Bayes' rule (3.9):

$$p(\theta ; \gamma_k) = \frac{g[t(\underline{Y}_k), \theta] \, g[\gamma_0, \theta]}{\int_{\Theta} (\text{Numerator}) \, d\theta} \tag{3.15}$$

$$= K(\gamma_k) \, g[\gamma_k, \theta] \tag{3.16}$$

The second equality follows because $g(\cdot,\cdot)$ was assumed maximally factored, so that

$$g[\gamma_k, \theta] = g[t(\underline{Y}_k), \theta] \, g[\gamma_0, \theta] \tag{3.17}$$

Finally, use (3.14) - (3.17) to rewrite (3.11)

$$f(\underline{Y}_k) = \frac{K(\gamma_0)}{K(\gamma_k)} \; G(\underline{Y}_k) \tag{3.18}$$

This, as claimed, does not depend on θ .

Equation (3.18) implies that, though $t(\underline{Y}_k)$ is sufficient for θ , it is <u>not</u> sufficient for the marginal density of the observation (and hence, in a later section, for the marginal likelihood ratio); this in turn means that (3.15) and (3.17) can be made sequential, but this need not be so for (3.18).

<u>3.1.2 Other Reproducing Densities</u>. Suppose now that the actual <u>a priori</u> p.d.f. on Θ is not a natural conjugate prior but

can, for some value $\gamma_0 \in \Gamma$ of the conjugate parameter, be written[1]

$$f_0(\theta) = r(\theta)\, p(\theta ; \gamma_0) \tag{3.19}$$

where $r(\theta)$ is a positive function defined on Γ. The following result will show that $f_0(\theta)$ also reproduces with the parameter γ.[2]

THEOREM 3.2

Let $f_1(\theta)$, $f_2(\theta)$ be two p.d.f.'s on Θ which can be written

$$f_2(\theta) = r(\theta)\, f_1(\theta) \tag{3.20}$$

---

[1] Note that $r(\theta)$ is a Radon-Nikodym derivative (see Appendix B). If $\mu_p(\theta)$ is the Borel measure on Θ represented by the p.d.f. $p(\theta ; \gamma_0)$ and $\mu_f(\theta)$ the measure represented by $f_0(\theta)$, then $r(\theta)$ is the Radon-Nikodym derivative of μ_f with respect to μ_p. The positivity of $r(\theta)$ can be relaxed to requiring it nonnegative but nonzero in some neighborhood of the "true" value θ^*; this is necessary so that the actual a priori p.d.f. does not exclude the true value.

[2] This can be shown more directly (see, e.g., Spragins [59]) by applying Bayes' rule to (3.19). Clearly, the reproducing class is the natural conjugate class with each member multiplied by $r(\theta)$. The approach here was chosen because the equations developed in the theorem are needed for later work.

Suppose that using $f_1(\theta)$ as an a priori p.d.f. results in an a posteriori p.d.f. $f_1(\theta \mid \underline{Y}_k)$ and a marginal p.d.f. $f_1(\underline{Y}_k)$. Similarly, using $f_2(\theta)$ results in $f_2(\theta \mid \underline{Y}_k)$ and $f_2(\underline{Y}_k)$. Then

$$f_2(\theta \mid \underline{Y}_k) = \frac{r(\theta) f_1(\theta \mid \underline{Y}_k)}{\int_{\Theta} r(\theta') f_1(\theta' \mid \underline{Y}_k) d\theta'} \tag{3.21}$$

$$f_2(\underline{Y}_k) = f_1(\underline{Y}_k) \int_{\Theta} r(\theta') f_1(\theta' \mid \underline{Y}_k) d\theta' \quad \blacksquare \tag{3.22}$$

The proof is given in Appendix D.

Corollary

An a priori p.d.f. which can be written as in (3.19) reproduces with parameter γ. $\blacksquare$

For, in an obvious notation, (3.21) becomes

$$f(\theta \mid \underline{Y}_k, f_0) = \frac{r(\theta) p(\theta; \gamma_k)}{\int_{\Theta} (\text{Numerator})\, d\theta} \tag{3.23}$$

Since $r(\theta)$ is known, the product function $\{ r(\theta)\, p(\theta;\gamma) ; \gamma \in \Gamma \}$ is a class of (unnormalized) p.d.f.'s which reproduces under $f(\underline{Y}_k|\theta)$ with the conjugate parameter γ. Applying (3.22) yields

$$f(\underline{Y}_k|f_0) = f(\underline{Y}_k|p_0) \int_{\Theta} r(\theta)\, p(\theta;\gamma_k)\, d\theta \qquad (3.24)$$

where p_0 means "using $p(\theta;\gamma_0)$ as an a priori p.d.f.". Now suppose $f_0(\theta)$ is given: One can choose a convenient γ_0, define $r(\theta)$ by (3.19), and proceed (using the theory of the preceding section) on the assumption that a priori p.d.f. is $p(\theta, \gamma_0)$. To account for the actual a priori p.d.f. $f_0(\theta)$, the results must then be modified using (3.23) and (3.24).

If it is possible to define reproducing classes for which the integral in the last two equations exists in closed form, the use of this technique can be as tractable as the use of natural conjugate priors. If the integral must be evaluated numerically, the resultant processing is probably too complex to be worthwhile.

3.2 Compound-Hypothesis Detection

The problem was stated in Section 1.2.3 and its solution begun in 1.3.3. The notation is more complicated than that of the preceding section since there are essentially two separate estimation problems to be solved. Under hypothesis H_1, the

symbols introduced for the estimation problem will be used; under H_0, an analogous set of symbols is introduced. This is summarized in Table 3.1.

Table 3.1 Detection Problem Notation

	Hypothesis H_1	Hypothesis H_0
Uncertain Parameter	θ	η
Conditional p.d.f. on the Observations	$f(\underline{Y}_k \mid \theta, H_1)$	$f(\underline{Y}_k \mid \eta, H_0)$
Sufficient Statistic	$t_1(\underline{Y}_k)$	$t_0(\underline{Y}_k)$
Natural Conjugate p.d.f.'s	$p(\theta;\gamma)$, $\gamma \in \Gamma$	$q(\eta;\psi)$, $\psi \in \Psi$
Other Reproducing p.d.f.'s	$f_0(\theta) = r_\theta(\theta)p(\theta;\gamma_0)$	$f_0(\eta) = r_\eta(\eta)q(\eta;\psi_0)$

Recall that θ and η may represent completely different parameters or may have components which represent the same parameters (as would, for instance, be the case in the classical "signal or signal-plus-noise" problem with unknown noise parameters). Even if they represent the same parameters, the natural conjugate classes may be different (and hence, since they are

computed under different statistical hypotheses, their a posteriori p.d.f.'s will evolve differently).

The optimal detection statistic is the likelihood ratio of marginal observation p.d.f.'s (1.36). This was rewritten in (1.38) and formulated sequentially in (1.40). Recall from these equations and the comment following (3.18) that if a "parameters fixed" (simple hypotheses) solution is known, then the a posteriori p.d.f.'s (i.e., the updated parameters of the natural conjugate p.d.f.'s) are sufficient to solve the compound-hypotheses problem. If not, then this may not be the case.

3.2.1 Natural Conjugate Densities. Suppose that both hypotheses admit sufficient statistics and that the a priori p.d.f.'s are members of the corresponding natural conjugate classes. The detector can be written in two equivalent forms: The first is useful if a solution to the corresponding simple-hypotheses problem already exists (especially if the results are to be extended to the case of continuous observation); the second eliminates redundancies in the first and is especially useful if the sequential (or the one-shot discrete) solution is desired for its own sake.

a. Form #1 of the Detector: Apply (3.13) and its analog for hypothesis H_0 to (1.38):

$$\ell(y_i \mid \underline{Y}_{i-1}) = \frac{p(\theta ;\gamma_{i-1})}{p(\theta ;\gamma_i)} \frac{q(\eta;\psi_i)}{q(\eta;\psi_{i-1})} \ell(y_i \mid \underline{Y}_{i-1}, \theta , \eta) \tag{3.25}$$

Recall again the comment following (1.39); the computation required to implement (3.25) can often be simplified by proper choice of the fixed values of the parameters θ and η. Updating of the conjugate parameters γ and ψ takes place through the sufficient statistics, precisely as described for the estimation problem. If the logarithm of the likelihood ratio is denoted $z(\cdot)$, then Fig. 3.1 illustrates the processing necessary to find $z(\underline{Y}_k)$ by use of (3.25); the tilde on θ and η indicates that fixed values of those parameters are used. The updated values of γ and ψ completely characterize the a posteriori p.d.f.'s, so the conjugate parameters are sufficient outputs to an external estimator (recall the comment at the end of Section 1.3.2(b)).

Finally, note that the "parameters given" likelihood ratio is explicitly required; this, again, is because γ and ψ are not sufficient for the marginal densities.

b. Form #2 of the Detector. If the simple hypothesis solution is not available there is no advantage to explicitly retaining

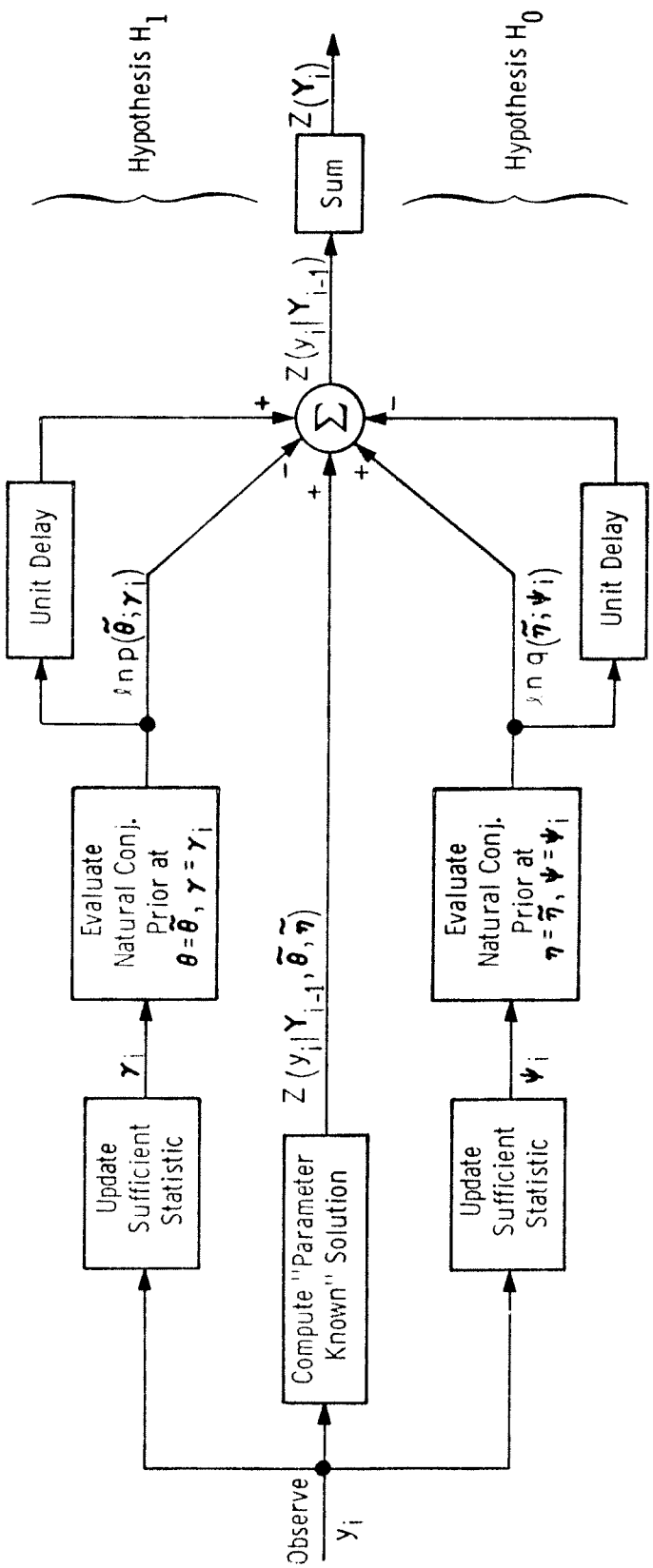

Fig. 3.1 The Primary Detection Processor

$\ell(y_i | \underline{Y}_{i-1}, \theta, \eta)$; one may as well perform the simplification of (3.18), in which case the "one-shot" likelihood ratio becomes

$$\ell(\underline{Y}_k) = \frac{K_p(\gamma_0)}{K_p(\gamma_k)} \frac{K_q(\psi_k)}{K_q(\psi_0)} \frac{G(\underline{Y}_k | H_1)}{G(\underline{Y}_k | H_0)} \qquad (3.26)$$

$K_p(\cdot)$ and $K_q(\cdot)$ are normalizing constants of the corresponding natural conjugate densities, see (3.14). Equation (3.26) depends on $\underline{Y}_k$ through the functions $G(\cdot)$, see (3.1); it can be intractable unless these are themselves of fixed dimension. In most applications, this will be the case.

The result can be made sequential by noting from (1.40) that

$$\ell(y_i | Y_{i-1}) = \ell(\underline{Y}_i) / \ell(\underline{Y}_{i-1}) \qquad (3.27)$$

and using (3.26) for the numerator and denominator; this is of little use unless the resulting ratios are simpler to evaluate than the functions themselves. In a sense, though, (3.26) is already sequential: Provided that the functions $G(\cdot | H)$ are tractable, it depends on $\underline{Y}_k$ only through statistics of fixed dimension which are "updated" in sequential fashion.

3.2.2 Other Reproducing Densities. Suppose that natural conjugate prior densities are not appropriate, but that the actual a priori p.d.f.'s can be written as in (3.19) (see Table 3.1). Then (3.24) holds under both hypotheses, and

$$\ell(\underline{Y}_i | f_0) = \ell(\underline{Y}_i | p_0, q_0) \frac{\int_{\Theta} r_\theta(\theta) p(\theta;\gamma_i) d\theta}{\int_{\mathcal{N}} r_\eta(\eta) q(\eta;\psi_i) d\eta} \tag{3.28}$$

where $\ell(\cdot | p_0, q_0)$ is the likelihood ratio assuming that $p(\theta;\gamma_0)$ and $q(\eta;\psi_0)$ are the a priori densities; this is computed by the

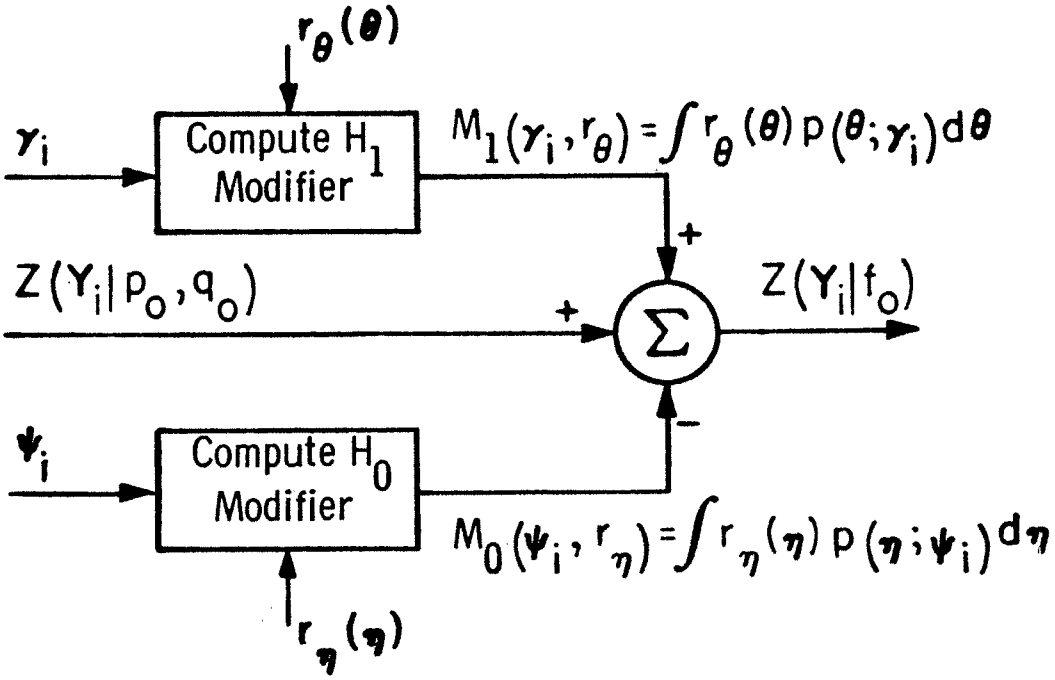

Fig. 3.2 The Secondary Processor

"primary processor" given by (3.25), Fig. 3.1, or (3.26). A "secondary processor" (see Fig. 3.2) then computes the "modifier gains" in (3.28) and computes the actual likelihood ratio. Again, the secondary processing can be minimal if the integrals of (3.28) exist in closed form; if they must be evaluated numerically, then the designer must trade the freedom of choosing a priori p.d.f.'s against the cost of the processing involved.

3.3 Continuous Observations

Suppose that, as discussed in Sections 1.1 and 2.2.1, discrete observations are generated by sampling $\{y_t, t\in[0, T]\}$; one wishes to infer how the continuous observation should be processed by studying the preceding results as the samples grow dense. Much of the necessary groundwork has been laid in Section 2.2; the development here will be less rigorous and will deal primarily with p.d.f.'s and the convergence of their ratios; the conditions necessary for convergence have already been established.

Under hypothesis H_1, $\{y_t, t\in[0, T]\}$ is defined on a probability space (sample space) $(\mathcal{Y}, \mathcal{A}, \mathcal{P}_\theta)$; $\mathcal{M}=\{\mathcal{P}_\theta;\theta\in\Theta\}$ is a family of probability measures and $\mathcal{P}_{\theta_0}$ dominates $\mathcal{M}$, i.e., $\mathcal{P}_{\theta_0} << \mathcal{P}_\theta$ for all θ. Under H_0, the process is defined on $(\mathcal{Y}, \mathcal{A}, \mathcal{P}_\eta)$ and a probability measure $\mathcal{P}_{\eta_0}$ dominates $\{\mathcal{P}_\eta\}$; assume that $\mathcal{P}_{\theta_0} << \mathcal{P}_{\eta_0}$. It is not necessary,

but will often be the case, that $\mathcal{P}_{\theta_0}$ and $\mathcal{P}_{\eta_0}$ are members of the families they dominate.

Assume further that $t_1(\cdot) : (\mathcal{Y}, \mathcal{A}) \to (R^r, \mathcal{R}^r)$ and $t_0(\cdot) : (\mathcal{Y}, \mathcal{A}) \to (R^s, \mathcal{R}^s)$ are sufficient statistics for θ and η respectively; then it has been shown that, as $k \to \infty$,

$$\Lambda_k^{\,1}(\underline{Y}_k; \theta) \to \lambda_1(y_t; \theta) = g_1[t_1(y_t); \theta]\, G_1(y_t) \qquad (3.29)$$

$$\Lambda_k^{\,0}(\underline{Y}_k; \eta) \to \lambda_0(y_t; \eta) = g_0[t_0(y_t); \eta]\, G_0(y_t) \qquad (3.30)$$

where y_t denotes a sample function and where, for example,

$$\Lambda_k^{\,1}(\underline{Y}_k; \theta) = \frac{f(\underline{Y}_k | \theta, H_1)}{f(\underline{Y}_k | \theta_0, H_1)} \qquad (3.31)$$

Again, the use of "t" for time and "t(·)" for sufficient statistics should cause no confusion.

Section 3.3.1 will treat the estimation problem by itself and as usual employ the notation of H_1. Section 3.3.2 then applies the results to the detection problem, and 3.3.3 discusses how continuous solutions on subintervals of $[0, T]$ can be treated sequentially. Throughout, the results will be seen to bear a striking resemblance to those obtained for the discrete case.

3.3.1 Estimation. Recall that, under the assumptions, an a posteriori p.d.f. for continuous observations can be defined as in (2.23). It will be assumed throughout that the integral in the denominator exists. Using the generalized factorizations given

above (i.e., using Theorem 2.3),

$$f(\theta \mid y_t) = \frac{g[\, t(y_t);\theta \,] f_0(\theta)}{\int_{\Theta} (\text{Numerator})\, d\theta} \tag{3.32}$$

Now, analogous to Theorem 3.1, put

$$p_c(\theta;\gamma) = \left. \frac{g[\, t(y_t);\theta \,]}{\int_{\Theta} g[\, t(y_t);\theta' \,] d\theta'} \right|_{t(y_t) = \gamma} \tag{3.33}$$

where the subscript "c" denotes "continuous." It is easy to verify that

$$\{ p_c(\theta;\gamma),\ \theta \in \Theta \}_{\gamma \in \Gamma} = t(\mathcal{Y})$$

is a reproducing class of p.d.f.'s on Θ under the generalized Bayes' rule (3.32); it will, again, be called the natural conjugate class and its members are indexed by the conjugate parameter γ. This is true because the form of (3.32) is the same as that of the finite-dimensional Bayes' rule, and that form alone was necessary

to prove Theorem 3. 1.[1] The relation by which the conjugate parameter is updated depends upon the parametrization of p_c, which in turn depends upon the choice of sufficient statistic (recall that this is not unique).

Suppose that, analogous to (3. 19), the a priori p. d. f. is not natural conjugate but can be written

$$f_0(\theta) = r(\theta)\, p_c(\theta ; \gamma_0) \quad , \quad \begin{matrix} \theta \in \Theta \\ \gamma_0 \in \Gamma \end{matrix} \tag{3.34}$$

Since the proof of Theorem 3. 2 also depended only on the functional form of Bayes' rule, the analog to (3. 23) is immediate:

$$f(\theta \mid y_t, f_0) = \frac{r(\theta) p_c(\theta ; \gamma_T)}{\int_\Theta \text{(Numerator)}\, d\theta} \tag{3.35}$$

---

[1] From an alternate point of view, the function $g[\, t(y_t) ; \theta \,]$ can be normalized and considered as a bona fide p. d. f. on the "sufficient space" $\mathcal{T} = t(\mathcal{Y})$; this space is by assumption finite-dimensional, so that all the results can be extended with only a slight notational modification.

where γ_T is computed assuming that the a priori p.d.f. is $p_c(\theta;\gamma_0)$.

3.3.2 Compound-Hypothesis Detection. For a finite number of samples, and assuming a simple-hypothesis solution known, the problem is solved by (1.38), i.e.,

$$\ell(\underline{Y}_k) = \frac{f_0(\theta)}{f(\theta \mid \underline{Y}_k, H_1)} \quad \frac{f(\eta \mid \underline{Y}_k, H_0)}{f_0(\eta)} \quad \ell(\underline{Y}_k \mid \theta, \eta) \tag{3.36}$$

Recall that this expression does not depend on θ or η; so long as it makes sense, it may be evaluated at arbitrary fixed values of those parameters.

Under the assumptions, each term in (3.36) converges and thus the decision statistic converges to

$$\ell(y_t) = \frac{f_0(\theta)}{f(\theta \mid y_t, H_1)} \quad \frac{f(\eta \mid y_t, H_0)}{f_0(\eta)} \quad \ell(y_t \mid \theta, \eta) \tag{3.37}$$

where the last term is the simple-hypothesis solution and is presumed to be known. If natural conjugate a priori p. d. f. 's are used then, in the previously introduced notation,

$$\ell(y_t) = \frac{p_c(\theta;\gamma_0)}{p_c(\theta;\gamma_T)} \frac{q_c(\eta;\psi_T)}{q_c(\eta;\psi_0)} \ell(y_t|\theta,\eta) \qquad (3.38)$$

If the actual priors are not natural conjugate but can be written as in (3. 19), (see Table 3. 1), then an expression analogous to (3. 28) can also be developed.

3. 3. 3 Sequential Processing of Continuous Observations. Suppose that the interval $[0, T]$ is partitioned into subintervals by defining a set of partitioning times T_i so that $0 < T_1 < T_2 < \ldots < T_{n-1} < T_n = T$. It is tempting to consider processing the continuous observation on the subintervals $(T_i, T_{i+1}]$, using the sequential results derived in Sections 3. 1 and 3. 2 to obtain analogous notions of a posteriori p. d. f. 's and likelihood ratios which "evolve" as further subintervals are observed. This is a very difficult problem to treat in general,[1] but can be quite

---

[1] See, e. g., Bahadur [3], Arts. 8-11.

tractable in specific cases.

Consider, for example, that $\{y_t, t\epsilon[0, T]\}$ is an M^{th}-order Markov process and that $k \geq M$ equally-spaced samples are taken from each subinterval (see Fig. 3.3).[1] Let y_{ij} denote the j^{th} sample from the i^{th} subinterval; analogous to (1.4), let

$$\underline{y}_{ij} = (y_{ij}, y_{i,j-1}, \cdots, y_{i,\ j-M+1})^*$$

and

$$\underline{Y}_{i,k} = (y_{i,1}, \cdots, y_{i,k})^*$$

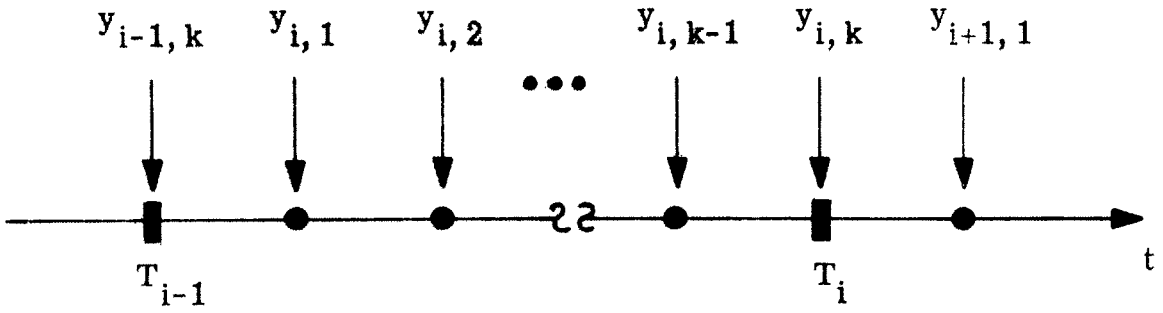

Fig. 3.3 Sampling Scheme

---

[1] See Section 1.1.

Note that all the samples of the vector $\underline{y}_{i,0}$ belong to the $(i-1)^{st}$ subinterval and are known if that interval has been processed; thus, using the p.d.f. $f(\underline{Y}_{i,k}|\underline{y}_{i,0},\theta)$ makes sense. Under suitable conditions, the likelihood ratio function can be formed, and its limit as $k \to \infty$ taken as outlined earlier; this in turn yields the sufficient statistic, natural conjugate density, and the updating expressions. The procedure will be illustrated for a specific problem in Chapter V.

3.4 Conclusions and Historical Sketch

3.4.1 Summary and Discussion.

It has been shown that the optimal detector solves two separate Bayesian estimation problems, one under each hypothesis, and then uses the a posteriori p.d.f.'s to modify a simple-hypothesis likelihood ratio; thus, detection and estimation occur simultaneously. Without further assumptions, this is not generally a tractable solution since all densities must be stored and manipulated as functions during the estimation.

If the observations admit a sufficient statistic (of fixed size) for the unknown parameters, then a class of natural conjugate reproducing densities exists under either hypothesis; each class is indexed by a parameter vector of the same dimension as the sufficient statistic. The ensuing simplifications are dramatic:

1. Sequential Bayesian estimation becomes tractable since a posteriori p.d.f.'s are completely characterized by the fixed-dimensional "conjugate" parameter vector; this parameter is updated through the use of fixed-dimensional sufficient statistics to reflect evolution of the p.d.f.'s. Both operations are inherently recursive.

2. Thus, if a simple-hypothesis detector is known, the compound hypothesis detection problem is tractably solved as stated above.

3. The optimal solution becomes independent of a priori distributions on the parameters in the following sense: "Primary" processing (detection and estimation) of the signal is done under the assumption that the a priori densities are natural conjugate; a secondary processor then modifies both the likelihood ratio and the a posteriori p.d.f.'s using the Radon-Nikodym derivatives of the actual a priori p.d.f. with respect to the assumed natural conjugate priors. This partitioning is illustrated in Fig. 3.4.

The primary processor is as developed in (3.25) and shown in Fig. 3.1; it can be permanently designed ahead of time using the "natural conjugate" prior densities. The secondary processor modifies the output of the primary processor according to (3.28); it can be reprogrammed for different a priori p.d.f.'s simply by supplying the necessary Radon-Nikodym derivatives. The user

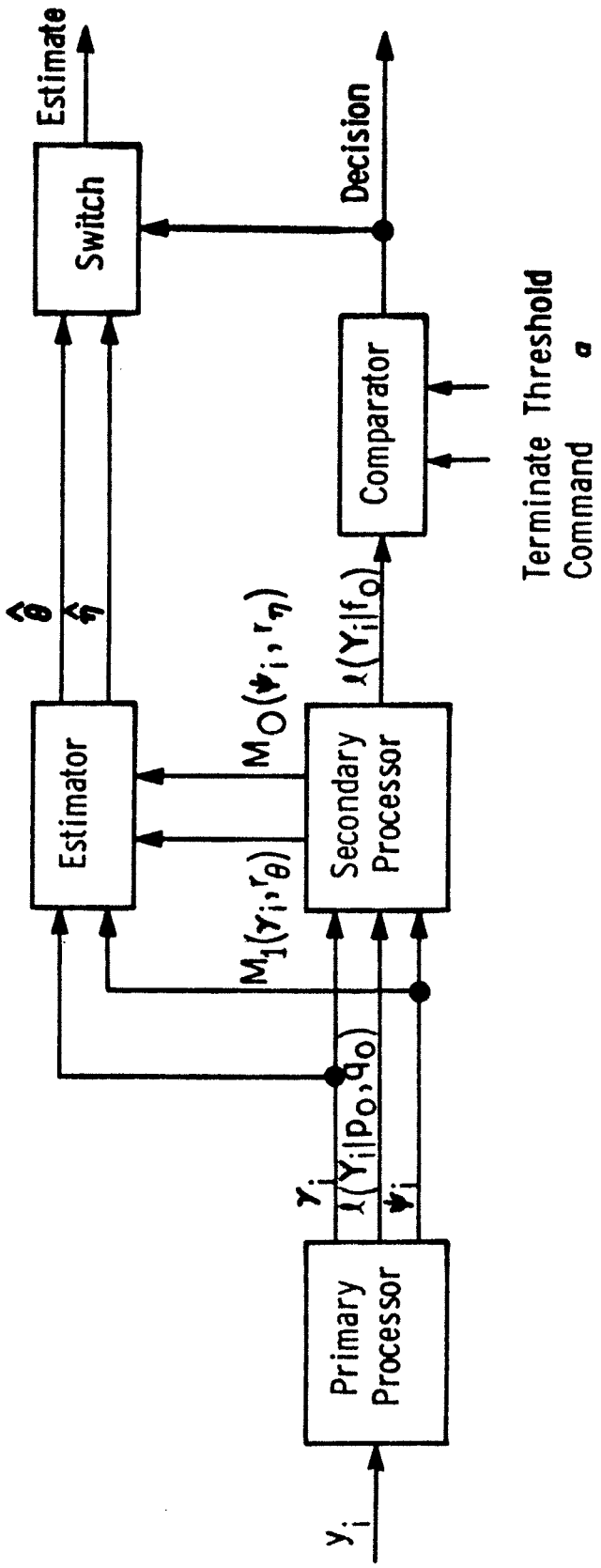

Fig. 3 4 Partitioning of the Estimator/Detector

can reoptimize the receiver "in the field" for any a priori p.d.f. which represents a distribution absolutely continuous w.r.t. the natural conjugate distribution. The estimation processor has a structure which depends on the cost functional; regardless of the specific cost functional, however, it uses the true a posteriori p.d.f., see (3.23), and hence requires the inputs shown.

The advantages of this partitioning are threefold. First, it eliminates the need for the designer to know exactly the a priori p.d.f.'s with respect to which he must optimize. Second, he can use the mathematical tractability of the natural conjugate prior p.d.f.'s in developing the primary processor. And third, the user obtains a capability for reoptimizing the receiver as his information (or opinion) about the a priori densities changes.

Under suitable restrictions, many of the results apply to continuous observations made on a fixed finite interval as well. The resulting expressions are remarkably similar to those obtained in the discrete case.

3.4.2 Historical Outline. It has long been known that the existence of sufficient statistics is a prerequisite for practical Bayesian estimation, especially if a large number of observations are to be processed; many of the techniques presented in Section 3.1 are implicit in the actions of the Bayesian statistician, even though he may not have bothered to explicitly formulate them or even be

aware of the fact that he is using them.

The concepts of sufficient statistics, natural conjugate densities, and Bayesian estimation appear to have first been explicitly connected in a decision-theoretic context by Raiffa and Schlaiffer [46], Chapters 2 and 3. Another text, which considerably clarifies the results and employs them in the same context, is that of De Groot [10], Chapter 9. Both references (as well as most other statistical works) treat only the case of discrete, independent observations and are thus rather limited in their applicability to communications or control problems.

It has for the past decade been apparent to researchers at The University of Michigan that the same fundamental connection between these concepts could be useful in detection theory. For example, Roberts [48] observed that if a uniform a priori p.d.f. was placed on phase in the "signal known exactly except for phase + Gaussian noise of known spectrum" detection problem, then the a posteriori p.d.f. had a form which remained unchanged as more information was processed; he called this property "closure of the distribution." Spooner [58] similarly found that a Gamma p.d.f. placed on the unknown intensity of white Gaussian noise would reproduce. The report of Birdsall [12] sought to summarize the knowledge about the phenomenon; in many ways, it is a departure point for the research presented here.

Simultaneously, the works of Spragins [59] and Grettenberg [20] were published; both struck close to the core of the subject as presented here and were instrumental in the formulation of this work.

The development of this research began as an attempt to extend and clarify the report by Birdsall [12]. The key steps turned out to be:

(a) The explicit factorization of the conditional density function as in, e.g., (2.10).

(b) Defining the natural conjugate density as in (3.8); previous works used a similar definition but employed the entire conditional p.d.f., which is correct but obscures what actually occurs.

(c) Separating the concepts of "parameters which index the natural conjugate densities" and "sufficient statistics"; previous works lumped these under a common symbol, obscuring the actual nature of the "updating" relationships.

The results of this early research are published in [7] and constitute Sections 3.1 and 3.2 of this dissertation.

The remainder of the results were obtained as an outgrowth of an attempt to solve the problem of detecting a known signal in 1-SAG noise of unknown parameters (Example 1.3, except α is known). It soon became apparent that a sampled version of the problem fit quite naturally into the theory developed above; the

results are in Section 4.4. Attempts to limit these results by letting the samples grow dense failed until the technique of using the likelihood ratio function (Section 3.3) was attempted; an application of a theorem due to Baxter (Theorem 5.3) gave the desired result, and provided a clue to the intimate connection between the singularity of measures and the use of Bayesian methods in infinite dimensional spaces. The papers of Halmos and Savage [24] and Bahadur [3] clarified what was happening, and the continuous-time results of Section 3.3 and Appendix B followed. Finally, a study of the subject of Gaussian measures set the stage for the work of Chapter V (and, incidentally, revealed that much of the 1-SAG solution had already been obtained by Striebel [60]).

CHAPTER IV

EXACTLY-KNOWN SIGNALS IN DISCRETE STATIONARY AUTOREGRESSIVE GAUSSIAN NOISE

This and the following chapter will seek to illustrate the theory of Chapters II and III by applying it to the problem of optimally estimating the spectral parameters of M-SAG noise while detecting a known, sure signal corrupted by that noise. The problem, though closely related to the classical Gaussian detection problem, is heretofore unsolved at this level of generality.

As stated in the introduction, no claim is made as to the practicality of the result. That is not the object here. Rather, "estimation" consists of explicit knowledge of the a posteriori p. d. f. ; no attempt at a detailed analysis of this density and estimates based on it will be made. Similarly, the detection statistic will turn out to be a very complicated function of the observation which, again, will be left as is rather than being approximated or analyzed further. The purpose of these chapters is primarily to illustrate the theory. However, it will become clear that all the elements upon which a practical analysis must be based are present here.

The problem addressed in this chapter can be summarized as follows:

$$H_0 : y_i = n_i$$

$$H_1 : y_i = n_i + s_i \; ; \; i = 1, 2, 3, \ldots \tag{4.1}$$

where $\{s_i\}$ is a known sequence of real numbers and where $\{n_i\}$ is discrete-parameter, M^{th}-order, stationary, autoregressive Gaussian (M-SAG) noise whose spectral parameters are not known. The object is to detect the presence or absence of the signal, simultaneously estimating the parameters of the noise.

Section 4. 1 will discuss the noise in some detail; two parametrizations will be considered. The first is the spectral parametrization mentioned above, while the second characterizes the noise in terms of the elements of the correlation matrix of $M + 1$ sequential samples; relations between the two sets of parameters will be derived. The ultimate interest is in discrete noise which results from sampling continuous-time M-SAG noise, but this question will be postponed until Chapter V.

Section 4. 2 treats estimation of the noise parameters by applying the concepts of Chapter III, especially Section 3. 1. Although the notation of H_0 will be used, it is clear from (4. 1) that this is possible under either hypothesis, so long as the hypothesis is specified.

Section 4. 3 then treats the related detection problem by

applying the results of Section 4.2; finally, the results are specialized to the case $M = 1$ in Section 4.4.

4.1 Noise Models and Parametrization

M-SAG noise is the stationary solution to an M^{th}-order autoregressive stochastic difference equation driven by a sequence of independent Gaussian random variables. As will be shown in Chapter V, a sequence of samples from the continuous M-SAG noise process considered in communications engineering can be modeled in this fashion. Alternate parametrizations will be investigated and the joint p.d.f. of k samples determined. This is necessary as a prelude to finding the sufficient statistics and natural conjugate class of densities for the unknown parameters of the autoregression.

4.1.1 The M^{th}-Order Discrete Autoregression. The process $\{n_k; k = 1, 2, 3\ldots\}$ of interest is the stationary solution to the stochastic difference equation

$$n_k + \beta_1 n_{k-1} + \cdots + \beta_M n_{k-M} = \alpha e_k \qquad (4.2)$$

where $\{e_k\}$ is a sequence of $N(0, 1)$ random variables which are statistically independent of each other and of the present and preceding noise samples. The parameters $\{\beta_1, \ldots, \beta_M, \alpha\}$ are real. For the equation to be stable, the β's must be such that all

the roots of

$$Q(z) = z^M + \beta_1 z^{M-1} + \ldots + \beta_M = 0 \qquad (4.3)$$

have modulus less than 1. The region where this is true will be called $\mathcal{S} \subset R^M$. The parameter α will be assumed nonzero.

There is a technical inconsistency in calling $\{n_k\}$ a stationary sequence and yet considering only $k > 0$. This is resolved below by judiciously choosing the initial conditions $\{n_{-M+1}, \ldots, n_0\}$ and then agreeing to not consider values of k which are not positive.

The power spectral density of $\{n_k\}$ is

$$f(\omega) = \frac{\alpha^2}{Q(e^{i\omega})Q(e^{-i\omega})}$$

$$= \frac{\alpha^2}{|e^{iM\omega} + \beta_1 e^{i(M-1)\omega} + \ldots + \beta_M|^2} \qquad (4.4)$$

and thus the parameters $\{\beta_1 \ldots \beta_M, \alpha\}$ will be called the <u>spectral parameters</u> of the noise.

To simplify the notation, define the following column vectors:

$$\begin{aligned} \underline{\beta} &= (\beta_1, \ldots, \beta_M)^* && \in R^M \\ \underline{n}_i &= (n_i, \ldots, n_{i-M+1})^* && \in R^M \\ \underline{N}_k &= (n_1, n_2, \ldots, n_k)^* && \in R^k \end{aligned} \tag{4.5}$$

Equation (4.2) can then be written

$$n_k + \underline{\beta}^* \underline{n}_{k-1} = \alpha\, e_k \quad , \quad \begin{matrix} \underline{\beta} \in \mathscr{S} \\ \alpha \neq 0 \end{matrix} \tag{4.6}$$

The sequence $\{n_k\}$ is to be stationary; since its mean is zero this requires that

$$r_0 \stackrel{d}{=} E(n_i^2) \tag{4.7}$$

not depend on i. In practice this can be achieved by means of a "quiet start," i.e., by choosing the initial conditions $\underline{n}_0 = (n_{-M+1}, \ldots, n_0)^*$ from an ensemble with the desired stationary statistics. It is of interest to see what these statistics are. Using

(4.6) in (4.7) yields

$$r_0 = \underline{\beta}^* R \underline{\beta} + \alpha^2 \tag{4.8}$$

where R is the covariance matrix

$$R = E \underline{n}_i \underline{n}_i^* \tag{4.9}$$

Since it is a covariance, R is symmetric and positive definite.[1] Since $\{n_i\}$ is stationary, it is also Toeplitz.[2] Thus, its ij^{th} entry is

$$E n_i n_j = r_{|i-j|} \tag{4.10}$$

and R is constant along its diagonals:

---

[1] Strict positivity follows from condition (4.3); see Doob [11], pp. 253-255.

[2] Grenander and Szego [19], p. 170.

$$R = \begin{bmatrix} r_0 & r_1 & r_2 & \cdots & r_{M-1} \\ r_1 & r_0 & r_1 & & r_{M-2} \\ r_2 & r_1 & r_0 & & r_{M-3} \\ \vdots & & & & \vdots \\ r_{M-1} & r_{M-2} & r_{M-3} & \cdots & r_0 \end{bmatrix} \tag{4.11}$$

The quantities $\{r_0, r_1, \ldots, r_M\}$ will henceforth be called the correlation parameters; it will be useful to find relations between them and the spectral parameters. To accomplish this, M equations are needed in addition to (4.8). These are found by defining the M-vector of cross-correlations

$$\underline{r} = (r_1, \ldots, r_M)^* \tag{4.12}$$

and noting that

$$\begin{aligned} \underline{r} &= E\, n_k \underline{n}_{k-1} \qquad \forall\, k > 0 \\ &= - R\, \underline{\beta} \end{aligned} \tag{4.13}$$

where (4. 13) follows from (4. 6). If the correlation parameters are given, the spectral parameters can thus be found:

$$\underline{\beta} = - R^{-1} \underline{r} \tag{4.14}$$

$$\alpha^2 = r_0 - \underline{r}^* R^{-1} \underline{r} \tag{4.15}$$

the latter of which follows by using the former in (4. 8). Writing these equations for $\{r_0 \ldots r_M\}$ in terms of $\{\alpha, \underline{\beta}\}$ is more complicated and will not be done for general M. In a specific problem it is necessary to know the correlation matrix of $\underline{y}_0$ to achieve a "quiet start" in simulation or to account for the conditioning on $\underline{y}_0$ in sequential theoretical work, and the equations must ultimately be inverted. This will, e. g., be done in Section 4. 4 for $M = 1$.

4. 1. 2 The Transition and Joint Densities. The joint conditioned p. d. f. of k observations $\underline{N}_k = (n_1, \ldots, n_k)^*$ will be found as outlined in Section 1. 1, Eq. (1. 3). This is most conveniently done in terms of the spectral parameters.

By inspection of (4. 6), the distribution of n_k given $\underline{n}_{k-1}$ is $N(-\underline{\beta}^* \underline{n}_{k-1}, \alpha^2)$; thus, the transition density is

$$f(n_i \mid \underline{n}_{i-1}, \underline{n}_0, \underline{\beta}, \alpha)$$

$$= (2\pi\alpha^2)^{-\frac{1}{2}} \exp\left[-\frac{1}{2\alpha^2}(n_i + \underline{\beta}^* \underline{n}_{i-1})^2\right] \quad (4.16)$$

From (1.5), the desired joint p.d.f. is a k-fold product of transition densities. Recall that it is conditioned on $\underline{n}_0$ to be tractable for sequential estimation. If the resulting sum in the exponent is expanded, one obtains

$$f(\underline{N}_k \mid \underline{n}_0, \underline{\beta}, \alpha)$$

$$= (2\pi\alpha^2)^{-\frac{k}{2}} \exp\left\{-\frac{1}{2\alpha^2}\left[\underline{\beta}^* T_M(\underline{N}_{k-1})\underline{\beta} + 2\underline{\beta}^* \underline{t}_M(\underline{N}_k) + t_0(\underline{N}_k)\right]\right\}, \quad \underline{\beta} \in \mathcal{S}, \; -\infty < \alpha < \infty$$

(4.17)

where

$$T_M(\underline{N}_{k-1}) = \sum_{i=1}^{k} \underline{n}_{i-1}\, \underline{n}^*_{i-1}$$

$$= \begin{bmatrix} \sum_1^k n^2_{i-1} & \sum n_{i-1}\, n_{i-2} & \cdots & \sum n_{i-1}\, n_{i-M} \\ \sum n_{i-2}\, n_{i-1} & \sum n^2_{i-2} & \cdots & \sum n_{i-2}\, n_{i-M} \\ \vdots & & & \vdots \\ \sum n_{i-M}\, n_{i-1} & \sum n_{i-M}\, n_{i-2} & \cdots & \sum n^2_{i-M} \end{bmatrix}$$

(4. 18)

$$\underline{t}_M(\underline{N}_k) = \sum_{i=1}^{k} n_i\, \underline{n}_{i-1}$$

$$= \left[\sum_1^k n_i\, n_{i-1} \quad \sum n_i\, n_{i-2} \quad \cdots \quad \sum n_i\, n_{i-M} \right]^*$$

(4. 19)

$$t_0(\underline{N}_k) = \sum_{i=1}^{k} n^2_i \qquad (4.20)$$

To obtain the joint or transition densities in terms of the correlation parameters, it is only necessary to use (4.14) and (4.15) in these expressions.

4.2 Sequential Estimation

Before proceeding with a solution of the estimation problem, it is necessary to resolve a notational matter which arises because all the uncertain parameters (say, α and $\underline{\beta}$) are common to both hypotheses: They are parameters of the noise, and under H_0 the observation is the noise, $n_i = y_i$, while under H_1 the noise may be reconstructed by subtracting the exactly-known signal, $n_i = y_i - s_i$. Strictly speaking, the parameters are different random variables under H_0 and H_1 since they carry different distributions and may even be assigned different <u>a priori</u> densities; for this reason, and because in general they need not be common to both hypotheses, they were assigned different symbols (θ for H_1 and η for H_0) in Chapter III.[1] That practice will be discontinued from here on, and the parameters will be explicitly referred to as $(\alpha, \underline{\beta})$; the hypothesis can always be deduced from the notation

---

[1] As discussed in Section 1.1, this ambiguity could have been avoided through use of a more rigorous notation, but only at the cost of considerable complexity.

being used (recall Table 3.1).

For obvious reasons, the estimation problem will be solved assuming H_0 true. For notational simplicity, no further mention of the hypothesis will be made in this section.

4.2.1 Sufficient Statistics and the Natural Conjugate Class. The sufficient statistics have already been found as Example 2.1, Eqs. (2.36) - (2.38). The factorization (see (2.10) or (3.1)) is trivial;

$$G(\underline{Y}_k) = (2\pi)^{-k/2} \tag{4.21}$$

It will be convenient henceforth to use

$$\rho \stackrel{d}{=} \alpha^{-2} > 0 \tag{4.22}$$

instead of α. From (4.17) and Theorem 3.1, the natural conjugate class is seen to be

$$q(\rho,\underline{\beta};\psi) = K(\Psi)\, \rho^{\psi_c} \exp \left\{ -\frac{\rho}{2} [\underline{\beta}^* \Psi_M \underline{\beta} + 2\underline{\beta}^* \underline{\psi}_M + \psi_0] \right\} \tag{4.23}$$

where Ψ_M is an $M \times M$, symmetric, positive definite matrix; $\underline{\psi}_M$ is an M-vector; $\psi_0 > 0$ is a scalar; and ψ_c is a scalar "counting parameter"(see comment "a." following Theorem 2.6). The symbol "ψ" without subscripts refers to the totality of all the conjugate parameters; $K(\psi)$ is a normalizing constant.

For fixed values of $\underline{\beta}$, (4.23) represents an unnormalized Gamma p.d.f. in ρ; for fixed ρ, it is an M-variate Gaussian

in $\underline{\beta}$, truncated to be zero on the complement of the region $\mathscr{S}$. The latter fact makes it very difficult to deal with this p.d.f. in general terms; unfortunately, $K(\psi)$ is explicitly required to solve the detection problem (see (3.25) and (3.26)), and

$$K^{-1}(\psi) = \int_0^\infty \int_{\mathscr{S}} \rho^{\psi_c} \exp \left\{ -\frac{\rho}{2} [\underline{\beta}^* \Psi_M \underline{\beta} + 2\underline{\beta}^* \psi_M + \psi_0] \right\} d\underline{\beta}\, d\rho \tag{4.24}$$

The outer integral is easily evaluated,[1] giving

$$K^{-1}(\psi) = \int_{\mathscr{S}} \frac{\Gamma(\psi_c+1)\, 2^{\psi_c+1}\, d\underline{\beta}}{[\underline{\beta}^* \Psi_M \underline{\beta} + 2\beta^* \psi_M + \beta_0]^{\psi_c+1}} \tag{4.25}$$

where the integrand is the marginal density of $\underline{\beta}$. The remaining integral is quite difficult, and no attempt to evaluate or approximate it will be made here.

---

[1] See, e.g., Hogg and Craig [28], p. 91.

4.2.2 Updating and Estimation. Suppose that $q(\rho, \beta; \psi^{(0)})$ is used as an a priori p.d.f., $\underline{Y}_k = (y_1 \ldots y_k)^*$ is observed, and $\underline{y}_0 = (y_{-M+1}, \cdots, y_0)$ is known. From (3.9), it is clear that the a posteriori p.d.f. is $q(\rho, \beta; \psi^{(k)})$, where the conjugate parameter has been updated using

$$\Psi_M^{(k)} = \Psi_M^{(0)} + T_M(\underline{Y}_{k-1}) \tag{4.26}$$

$$\underline{\psi}_M^{(k)} = \underline{\psi}_M^{(0)} + \underline{t}_M(\underline{Y}_k) \tag{4.27}$$

$$\psi_0^{(k)} = \psi_0^{(0)} + t_0(\underline{Y}_k) \tag{4.28}$$

$$\psi_c^{(k)} = \psi_c^{(0)} + k/2 \tag{4.29}$$

Because the a posteriori p.d.f. is explicitly known, the estimation problem is in principle solved. Practically speaking, it is not, since an estimate based on (4.23) may be quite difficult. For large k one expects the p.d.f. to have a well-defined mode near the true value of the parameters, and the "maximum a posteriori" estimates

$$\hat{\beta}_{MAP}^{(k)} = - \left[\Psi_M^{(k)} \right]^{-1} \underline{\Psi}_M^{(k)}$$

$$\hat{\rho}_{MAP}^{(k)} = \frac{2\psi_c^{(k)}}{\psi_0^{(k)} - \underline{\Psi}_M^{(k)*} \left[\Psi_M^{(k)} \right]^{-1} \underline{\Psi}_M^{(k)}} \qquad (4.30)$$

obtained by differentiating (4.23) may be reasonable. These relations can also be useful in choosing $\psi^{(0)}$ to model the a priori p.d.f.

Making suitable ergodicity assumptions and using a law of large numbers, the MAP estimates are seen to be consistent: For large k the initial values $\psi^{(0)}$ in (4.26) through (4.29) become negligible; if $\bar{\underline{\beta}}$ and $\bar{\alpha}$ are the spectral parameters of the process generating the observations, then using (4.18) through (4.20) it is clear that

$$\hat{\beta}_{MAP}^{(k)} \to - R^{-1} \underline{r}$$

$$\hat{\rho}_{MAP}^{(k)} \to \frac{1}{r_0 - \underline{r}^* R^{-1} \underline{r}} \qquad (4.31)$$

where r_0, $\underline{r}$, R are the correlation parameters corresponding to $\bar{\alpha}$ and $\bar{\underline{\beta}}$; using (4. 14) and (4. 15), one finds that

$$\begin{aligned} \hat{\underline{\beta}}_{MAP}^{(k)} &\rightarrow \bar{\underline{\beta}} \\ \hat{\rho}_{MAP}^{(k)} &\rightarrow (\bar{\alpha})^{-2} = \bar{\rho} \end{aligned} \tag{4.32}$$

This argument can be made rigorous.

If the a posteriori (or the desired a priori) density does not have a well-defined mode, then the estimation problem may involve computing moments of (4. 23) and can be quite complicated. As discussed, a detailed analysis of the p. d. f. is outside the scope of this dissertation and will not be attempted.

Should the desired (or known) a priori p. d. f. not belong to the natural conjugate class but be expressible as in (3. 19), then the technique of Section 3. 1. 2 can applied. It may be possible to choose classes of densities which simplify the computation of moments and the estimation problem (recall the closing comment of Section 3. 1. 2), but this again is a problem which will not be addressed.

4. 2. 3 The Conditioning on $\underline{y}_0$. Recall from Sections 1. 1 and 2. 3 that, as derived, the a posteriori p. d. f. is conditioned

on the initial observations $\underline{y}_0$; to obtain a joint p. d. f. of the desired form (see (1. 5) and Section 2. 3), the basic relation throughout was Bayes' rule conditioned on $\underline{y}_0$:

$$f(\rho, \underline{\beta} \mid \underline{Y}_k, \underline{y}_0) = \frac{f(\underline{Y}_k \mid \underline{y}_0, \rho, \underline{\beta}) f_0(\rho, \underline{\beta} \mid \underline{y}_0)}{\int \text{(Numerator)} \, d\rho \, d\underline{\beta}} \qquad (4.33)$$

with (1. 5) employed in the numerator. The "a priori" p. d. f. in this expression is actually conditioned on $\underline{y}_0$ and must be interpreted as a posteriori to the initial conditions; if it is natural conjugate, then so is the a posteriori p. d. f. defined above,

$$f(\rho, \underline{\beta} \mid \underline{Y}_k, \underline{y}_0) = q(\rho, \underline{\beta}; \psi^{(k)}) \qquad (4.34)$$

and the solution given by the use of natural conjugate techniques is thus indeed conditioned on $\underline{y}_0$. As previously remarked, this must be undone or accounted for before the solution can be considered complete. From (1. 3) it is clear that (4. 33) can be unconditioned as follows:

$$f(\rho, \underline{\beta} \mid \underline{Y}_k) = \frac{f(\underline{y}_0 \mid \rho, \underline{\beta}) \, f(\underline{Y}_k \mid \underline{y}_0, \rho, \underline{\beta}) \, \tilde{f}_0(\rho, \underline{\beta})}{\int \text{(Numerator)} \, d\rho \, d\underline{\beta}} \qquad (4.35)$$

where for the problem at hand,

$$f(\underline{y}_0|\rho, \underline{\beta}) = (2\pi)^{-\frac{M}{2}} |R|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2} \underline{y}_0^* R^{-1} \underline{y}_0 \right\} \tag{4.36}$$

and R is related to $\rho = \alpha^{-2}$ and $\underline{\beta}$ through (4.14) and (4.15). Examination of (4.35) suggests two equivalent ways to account for the conditioning:

a. Consider the first and last term in the numerator as a unit and compare with (4.34); clearly, the a priori p.d.f. $f_0(\rho, \underline{\beta})$ in the conditioned result is equivalent to the a posteriori (to $\underline{y}_0$) p.d.f. in the unconditioned result. To obtain a situation which reproduces unconditionally using the natural conjugate class, it is necessary to choose an a priori $[\underline{y}_0]$ density from the class

$$\tilde{f}_0(\rho, \underline{\beta}) = \frac{q(\rho, \underline{\beta}; \psi)}{f(\underline{y}_0|\rho, \underline{\beta})} \tag{4.37}$$

in which case the a posteriori $[\underline{y}_0]$ p.d.f. will be natural conjugate but not conditioned on $\underline{y}_0$.

b. Separate the first term of (4.35) and compare with

(3.19) or (3.20); once $\underline{y}_0$ is observed, $f(\underline{y}_0|\rho,\underline{\beta})$ has all the properties required of $r(\theta)$ (i.e., $r(\rho,\underline{\beta})$). Thus if a "modifying processor" to account for other-than-natural-conjugate a priori densities has been implemented, one may as well choose a natural conjugate prior as discussed following (3.24) and include $f(\underline{y}_0|\rho,\underline{\beta})$ in the modifying term.

Finally, one can ignore the problem; this amounts to nothing more than a refusal to process the information in $\underline{y}_0$, and may be entirely reasonable if a large number of observations ($k >> M$) are to be made.

4.3 Sequential Detection

4.3.1 The Signal Hypothesis.

Recall from (4.1) that

$$H_1 : n_i = y_i - s_i \tag{4.38}$$

Thus, all estimation results of the preceding section may be applied under the signal hypothesis if, instead of using the observation, one uses $y_i - s_i$. Specifically, a set of necessary and sufficient statistics is $t_1(\underline{Y}_k) = t(\underline{Y}_k - \underline{S}_k)$ as defined in Example 2.1, and the natural conjugate class of densities is

$$p(\rho, \underline{\beta}; \gamma) = K(\gamma)\, \rho^{\gamma_c} \exp \left\{ -\frac{\rho}{2} [\underline{\beta}^* \Gamma_M \underline{\beta} + 2\underline{\beta}^* \gamma_M + \gamma_0] \right\} \tag{4.39}$$

with the conjugate parameters γ defined in complete analogy with the parameters ψ of (4.23).

If the initial parameters are $\gamma^{(0)}$, updating takes place through

$$\Gamma_M^{(k)} = \Gamma_M^{(0)} + T_M(\underline{Y}_{k-1} - \underline{S}_{k-1}) \tag{4.40}$$

$$\underline{\gamma}_M^{(k)} = \underline{\gamma}_M^{(0)} + \underline{t}_M(\underline{Y}_k - \underline{S}_k) \tag{4.41}$$

$$\gamma_0^{(k)} = \gamma_0^{(0)} + t_0(\underline{Y}_k - \underline{S}_k) \tag{4.42}$$

$$\gamma_c^{(k)} = \gamma_c^{(0)} + k/2 \tag{4.43}$$

and the a posteriori p.d.f. may be analyzed as in the preceding section. Note that $\psi^{(0)}$ and $\gamma^{(0)}$ may be different; even though the same observation parameters are being estimated under different statistical hypotheses, there is no reason why the a priori densities must necessarily be the same.

4.3.2. Detection. Putting the preceding expressions into (3.25) or using (3.26) directly, one finds that the marginal likelihood ratio is

$$\ell(\underline{Y}_k) = \frac{K[\gamma^{(0)}]\ K[\psi^{(k)}]}{K[\gamma^{(k)}]\ K[\psi^{(0)}]} \tag{4.44}$$

where $K(\cdot)$ is defined by (4.24) or (4.25). Clearly, the natural-conjugate normalizing constants are explicitly required to solve the sequential detection problem. Recall that these involve integrating over $\mathscr{S}$; since detailed solution of the problem is not a primary goal of this dissertation, the matter will not be pursued further.

4.4 Gauss-Markov Noise: M = 1

4.4.1 The Autoregression.

The noise is generated by

$$n_k + \beta_1 n_{k-1} = \alpha e_k \tag{4.45}$$

where $|\beta| < 1$. It is desirable that the results correspond to samples from continuous-time lowpass (as opposed to bandpass) Gauss-Markov noise; this requires the further restriction that $-1 < \beta_1 < 0$, which will henceforth be made. One obtains the following correspondences with quantities defined in Section 4.1:

General M	M = 1
β	β_1
$\underline{n}_i$	n_i
$\underline{r}$	r_i
R	r_0

Equations (4. 14) and (4. 15) become

$$\begin{aligned} \beta_1 &= - r_1/r_0 \\ \alpha^2 &= r_0 - r_1^2/r_0 \end{aligned} \tag{4.46}$$

which may be inverted to find

$$\begin{aligned} r_0 &= \frac{\alpha^2}{1 - \beta_1^2} \\ r_1 &= \frac{-\beta_1 \alpha^2}{1 - \beta_1^2} \end{aligned} \tag{4.47}$$

and so y_0 has (or, in simulation, should be chosen from an ensemble which has) a $N\left(0, \frac{\alpha^2}{1-\beta_1^2}\right)$ distribution.

The joint conditional p.d.f. of k observations is given by (4. 17) as

$$f(\underset{\sim}{N}_k | n_0, \rho, \beta_1) = \left(\frac{\rho}{2\pi}\right)^{-\frac{k}{2}} \exp\left\{-\frac{\rho}{2}\left[\beta_1^2 \sum_{i=1}^{k} n^2_{i-1} + 2\beta_1 \sum_{i=1}^{k} n_i n_{i-1} + \sum_{i=1}^{k} n_i^2\right]\right\} \tag{4.48}$$

The statistics $T_M(\cdot)$, $\underset{\sim}{t}_M(\cdot)$, and $t_0(\cdot)$ of (4.18) through (4.20) are all scalars; they can here be combined into a single vector:[1]

$$\underset{\sim}{t}(\underset{\sim}{N}_k) = \begin{bmatrix} \sum_{i=1}^{k} n_{i-1}^2 \\ \sum n_i n_{i-1} \\ \sum n_i^2 \\ k/2 \end{bmatrix} \tag{4.49}$$

---

[1] For an explanation of the last component, recall again comment (a.) following Theorem 2.6.

4.4.2 Estimation. The notation of H_0 will be used; since all the vectors and matrices defined in Section 4.2 are one-dimensional for $M = 1$, much of the notation is superfluous and will be dropped.

To parallel the definition of $\underline{t}(\cdot)$ above, the natural conjugate density (4.23) is written

$$q(\rho, \beta_1; \underline{\psi}) = K(\psi)\, \rho^{\psi_4} \exp \left\{ -\frac{\rho}{2} [\beta_1^2 \psi_1 + 2\beta_1 \psi_2 + \psi_3] \right\} \tag{4.50}$$

where $\underline{\psi}$, the parameter indexing the densities, is a vector with the indicated components which takes values in the subset of R^4 defined by

$$\Psi = \{\underline{\psi} \in R^4 : \psi_1, \psi_3, \psi_4 \geq 0 \text{ and } \psi_2^2 - \psi_1 \psi_3 < 0\} \tag{4.51}$$

Analysis of the density can be carried somewhat further than for arbitrary M. The marginal p.d.f.'s of β_1 and ρ are, respectively,

$$q_\beta(\beta;\psi) = \frac{K(\psi)\,\Gamma(\psi_4+1)\,2^{\psi_4+1}}{(\beta^2\psi_1+2\beta\psi_2+\psi_3)^{\psi_4+1}} \quad , \quad -1 < \beta < 0 \tag{4.52}$$

and

$$q_\rho(\rho;\underline{\psi}) = K(\underline{\psi})\left(\frac{2\pi}{\psi_1}\right)^{\frac{1}{2}} \rho^{\psi_4-\frac{1}{2}} \cdot \left[\Phi\left(\frac{\psi_2}{\psi_1\sqrt{\rho\psi_1}}\right) - \Phi\left(\frac{\psi_2-\psi_1}{\psi_1\sqrt{\rho\psi_1}}\right)\right] \cdot \exp\left[-\frac{\rho}{2\psi_1}(\psi_1\psi_3-\psi_2^2)\right] \quad , \quad \rho > 0 \tag{4.53}$$

where $\Phi(\cdot)$ represents the unit Gaussian cumulative distribution function. If the initial value of ψ_4 is chosen an integral multiple of $1/2$, then it will remain so throughout and (4.52) can be integrated to obtain an explicit (though not closed) expression for the

normalizing constant:[1]

Let

$$n = \psi_4$$

$$m = \psi_4 - \tfrac{1}{2}$$

$$D = \psi_1\psi_3 - \psi_2^2$$

$$P = \psi_1 + 2\psi_2 + \psi_3$$

$$N_i = \frac{(i-1)!\,i!}{(2i)!}$$

$$M_i = \frac{(2i)!}{(i!)^2}$$

Then

$$K^{-1}(\underline{\psi}) = \frac{2(2n)!}{n!}\left(\frac{\psi_1}{2D}\right)^n \left\{ \frac{1}{2D}\left[(\psi_1 - \psi_2)\sum_{i=1}^{n} N_i \left(\frac{4D}{\psi_1 P}\right)^i + \psi_2 \sum_{i=1}^{n} N_i \left(\frac{4D}{\psi_1\psi_3}\right)^i \right] + \frac{1}{\sqrt{D}}\left[\tan^{-1}\left(\frac{\psi_1-\psi_2}{\sqrt{D}}\right) + \tan^{-1}\left(\frac{\psi_2}{\sqrt{D}}\right)\right]\right\} \tag{4.54}$$

---

[1] The necessary integrals may be found in the C. R. C. Tables [55], p. 404 (#113) and p. 414 (#241).

provided that n is an integer, or

$$K^{-1}(\underline{\Psi}) = \frac{(m+1)!\,m!\,2^{3(m+1)}\,\Gamma(m+3/2)\,\psi_1^{\,m}}{\sqrt{2}\,D^{m+1}(2m+2)!}$$

$$\left[\frac{(\psi_1-\psi_2)}{\sqrt{P}} \sum_{i=0}^{m} M_i \left(\frac{D}{4\psi_1 P}\right)^i + \frac{\psi_2}{\sqrt{\psi_3}} \sum_{i=0}^{m} M_i \left(\frac{D}{4\psi_1\psi_3}\right)^i \right] \tag{4.55}$$

if m is an integer. These expressions, though not impressive for the insight they provide, are certainly suitable for machine evaluation.

From (4.49) and (4.50), the sufficient statistic updates the parameter of the natural conjugate density through

$$\underline{\Psi}^{(k)} = \underline{\Psi}(0) + \underline{t}(\underline{Y}_k) \tag{4.56}$$

The mode of the a posteriori p.d.f. occurs at

$$\hat{\beta}_1^{(k)} = -\psi_2^{(k)} \Big/ \psi_1^{(k)}$$

$$\hat{\rho}^{(k)} = \frac{2\psi_1^{(k)}\psi_4^{(k)}}{\psi_1^{(k)}\psi_3^{(k)} - [\psi_2^{(k)}]^2} \qquad (4.57)$$

and these may provide reasonable estimates.

To account for the conditioning on y_0, recall that $y_0 \sim N(0, r_0)$ as given by (4.47). So one can follow Section 4.2.3(a.) and choose the a priori density from

$$\tilde{f}_0(\rho, \beta_1; \underline{\psi}^{(-)}) = \frac{\psi_4^{(-)}}{\sqrt{1-\beta_1^2}} \rho \exp\left\{-\frac{\rho}{2}[\beta_1^2 \psi_1^{(-)} + 2\beta_1 \psi_2^{(-)} + \psi_3^{(-)}]\right\} \qquad (4.58)$$

in which case the a posteriori $[y_0]$ density will be natural conjugate with

$$\underline{\psi}^{(0)} = \underline{\psi}^{(-)} + \begin{bmatrix} -y_0^2 \\ 0 \\ y_0^2 \\ 1/2 \end{bmatrix} \qquad (4.59)$$

Note that (4.58) has no limit as $\beta_1 \to -1$; this technique may be undesirable unless β_1 can be bounded away from -1 . Alternately, one can follow the technique of Section 4.2.3(b.) as outlined there.

4.4.3 Detection. Not much more can be said. Since $n_i = y_i - s_i$ under H_1 , all the results of the previous section may be applied as detailed in Section 4.3.1; the constants $K(\psi^{(k)})$ and $K(\chi^{(k)})$ can be evaluated as in (4.54) and (4.55), and the likelihood ratio found from (4.44). No significant simplification results when the ratio of constants, rather than the constants themselves, are considered.

Thus, the discrete solution is unreasonably complicated even for $M = 1$. Instead of pursuing it further, attention will be focused on the continuous solution. This is both theoretically more interesting and results in somewhat simpler expressions.

CHAPTER V

EXACTLY-KNOWN SIGNALS IN CONTINUOUS STATIONARY AUTOREGRESSIVE GAUSSIAN NOISE

The problem addressed here is

$$H_0 : y_t = n_t$$

$$H_1 : y_t = n_t + s(t) \quad , \quad t \in [0, T] \qquad (5.1)$$

where n_t is M-SAG noise with unknown spectral parameters (Section 5.1) and $s(t)$ is an exactly-known sure signal which is 2M times differentiable for $0 < t < T$. It is a continuous version of the problem solved in the preceding chapter.

Section 5.1 discusses and establishes parametrizations for the noise, and relates a sampled version of n_t to the process studied in Chapter IV. Section 5.2 presents pertinent results from the theory of Gaussian measures; these are needed to insure non-singularity of the estimation and detection problems, and are also useful for evaluating the required Radon-Nikodym (R-N) derivatives without the tedium of finding a limit for the likelihood ratio function. To verify that the resulting derivatives are the same, Appendix E evaluates such a limit for $M = 1$.

Section 5.3 solves the estimation and detection problem for

arbitrary M. The sufficient statistics are found and the natural conjugate density and updating equations are explicitly written.

The chapter concludes by giving more detailed solutions for $M = 1$ and $M = 2$. Even for these cases, it is found that the detection statistic is extremely complicated and probably not practical as written.

5.1 Continuous Stationary Autoregressive Gaussian Noise.

The noise process under consideration, denoted $\{n_t,\ t \in [0,\ T]\}$, is a finite segment of a zero mean, M^{th}-order, stationary autoregressive Gaussian (M-SAG) random process; its spectral density function is rational and of the form

$$S_n(f^2) = \frac{a^2}{\prod_{i=1}^{M}\left[(2\pi f)^2 + p_i^2\right]} \quad ;\ p_i \neq p_k \qquad (5.2)$$

It is well known that this process can be modeled as white Gaussian noise filtered by a causal, lumped parameter system with Fourier transfer function

$$H(j2\pi f) = \frac{a}{\prod_{i=1}^{M}\left[(j2\pi f + p_i)\right]} \quad ; \quad \begin{matrix} \mathrm{Re}(p_i) > 0 \\ p_i \neq p_k \\ a > 0 \end{matrix} \qquad (5.3)$$

An alternate parametrization for the filter of (5.3) results from expanding the denominator

$$H(j2\pi f) = \frac{a}{Q(j2\pi f)} \tag{5.4}$$

where

$$Q(z) = z^M + q_1 z^{M-1} + \ldots + q_M \tag{5.5}$$

is a polynomial with real coefficients and with distinct roots in the left-half z-plane. The coefficient q_k is the sum of all different combinations of the p_i taken k at a time. In terms of this parametrization, the spectral density can be re-written as

$$S_n(f^2) = \frac{a^2}{|Q(j2\pi f)|^2} \tag{5.6}$$

Occasionally, it will be necessary to use yet a third parametrization of $S_n(f^2)$, namely

$$S_n(f^2) = \frac{1}{\left| \sum_{i=0}^{M} \theta_{M-i} (j2\pi f)^i \right|^2} \tag{5.7}$$

where $\theta_0 = |a|^{-1}$ and $\theta_i = q_i |a|^{-1}$, $i = 1 \ldots M$.

From (5.2), the autocorrelation function of the noise is

$$R_n(\tau) = a^2 \sum_{i=1}^{M} \mathscr{R}_i \, e^{-p_i |\tau|} \tag{5.8}$$

where

$$\mathscr{R}_i = \left[2 p_i \prod_{\substack{k=1 \\ k \neq i}}^{M} (p_k^2 - p_i^2) \right] \tag{5.9}$$

The process n_t is well known to be M^{th}-order Markov.[1] Since it is Gaussian, the M+1 parameters $\{a^2, p_1, \ldots, p_M\}$ or $\{a^2, q_1, \ldots, q_M\}$ constitute a complete statistical description. It is easily verified that for $\delta > 0$ the parameters $\{R_n(0), R_n(\delta), \ldots, R_n(M\delta)\}$ provide an equivalent description; the two sets are uniquely related through (5.8).

Suppose n_t is sampled at $t = i\delta$, $i = -M+1, \ldots, 0, 1, 2, \ldots$, and the corresponding discrete process denoted $\{n_i\}$. It can be shown[2] that this discrete process is the solution to a related

---

[1] More precisely: If n_t is sampled at arbitrary $t_1 < t_2 < t_3 < \ldots$ and samples arranged into "state vectors"

$$\underline{n}_i = (n_{t_i}, n_{t_{i-1}}, \ldots, n_{t_{i-M+1}})^*$$

then the discrete vector process $\{\underline{n}_i ; i = 1, 2, 3, \ldots\}$ is a Markov process. See, e.g., Astrŏm [2], Arts. 3.3 and 3.10.

[2] Doob [11] or Astrŏm [2], p. 84.

autoregressive difference equation (see Section 4.1.1); clearly any (M+1) adjacent samples are a multivariate Gaussian random variable with zero mean and covariance matrix

$$R_{M+1} = \begin{bmatrix} r_0 & r_1 & r_2 & \cdots & r_M \\ r_1 & r_0 & r_1 & & \vdots \\ r_2 & r_1 & r_0 & \ddots & \\ \vdots & & \ddots & \ddots & r_1 \\ r_M & \cdots & & r_1 & r_0 \end{bmatrix} \tag{5.10}$$

which is positive definite, symmetric, and Toeplitz[1] so that $r_{ij} = r_{|i-j|} = R_n(|i-j|\delta)$. Using these facts and equations (4.11)-(4.15), the parameters of the related difference equation can be determined.

It is also well known that almost every sample function of n_t is uniformly continuous and is everywhere M-1 times continuously differentiable; the $(M-1)^{st}$ derivative will be a function of unbounded variation.[2]

---

[1] Grenander and Szego [19], p. 170.

[2] This follows from a theorem of Baxter [4] which will be stated later as Theorem 5.3; see, e.g., Wong [62] pp. 221-222.

Before studying the detection and estimation problem in M-SAG noise, it is necessary to investigate results concerning the singularity of, and R-N derivatives for, Gaussian measures.

5.2 Gaussian Measures

5.2.1 Equivalence and Singularity. Once again, consider the pre-probability space $(\mathcal{Y}, \mathcal{A})$ whose elements are sample functions of $\{y_t; t \in [0, T]\}$, where $\mathcal{A}$ is generated by the process. If $\mathcal{P}_1$ and $\mathcal{P}_0$ are any two probability measures on $(\mathcal{Y}, \mathcal{A})$, then one can combine the Radon-Nikodym theorem (Theorem B.1) and the Lebesgue decomposition theorem[1] and write

THEOREM 5.1

There exists $f \geq 0$ which is meas.$[\mathcal{A}]$, and a finite measure $\mu \perp \mathcal{P}_0$, such that $\forall A \in \mathcal{A}$

$$\mathcal{P}_1(A) = \int_A f \, d\mathcal{P}_0(y) + \mu(A) \quad \blacksquare \tag{5.11}$$

Clearly, $\mathcal{P}_1 << \mathcal{P}_0$ iff $\mu \equiv 0$ and then $f = d\mathcal{P}_1 / d\mathcal{P}_0$ as usual.

A measure $\mathcal{P}$ on $(\mathcal{Y}, \mathcal{A})$ is a Gaussian measure if $\{y_t, t \in [0, T]\}$ is a Gaussian random process with respect to $\mathcal{P}$.[2]

---

[1] Royden [50], p. 240.

[2] That is, the random variables $\{y_t, t \in T_k\}$, with T_k any finite parameter subset of $[0, T]$, are jointly Gaussian. See Wong [62], pp. 46-47.

Assume henceforth that the process is separable and continuous in probability. Any Gaussian $\mathscr{P}$ is uniquely described by the corresponding mean and covariance functions of $\{y_t,\ t \in [0,\ T]\}$, and the singularity (or R-N derivative) of two such measures can be studied in terms of these functions. An extensive literature exists on the subject.[1]

(a) The Basic Dichotomy. It is a basic result that if $\mathscr{P}_1$ and $\mathscr{P}_0$ are Gaussian measures on $(\mathscr{Y}, \mathscr{A})$ then either $\mathscr{P}_1 \equiv \mathscr{P}_0$ or $\mathscr{P}_1 \perp \mathscr{P}_0$. This is proven by simultaneously diagonalizing the covariance matrices of $\{y_t,\ t \in T_k\}$ under $\mathscr{P}_1$ and $\mathscr{P}_0$; as the finite index set T_k grows dense in $[0,\ T]$, $\mathscr{P}_1$ and $\mathscr{P}_0$ are obtained as infinite products of (independent) one-dimensional Gaussian measures, and the result follows from a theorem due to Kakutani.[2]

If under $\mathscr{P}_1$ and $\mathscr{P}_0$ the process has the same covariance function $R(t, s)$ but different mean-value functions (say, without loss of generality, $m_1(t) = \mu(t)$ and $m_0(t) = 0$), then only one "matrix" need be diagonalized; this is easily accomplished and the result can be stated quite simply:

---

[1] Especially useful are the papers by Root [49] and Yaglom [63], and Arts. 6.1-6.4 of Wong [62].

[2] See Root [49], p. 296 or Wong [62], p. 215.

THEOREM 5.2[1]

Let $\{\lambda_i, \phi_i(t)\}$ be the eigenvalues and functions of the kernel $R(t,s)$; let $S \subset L_2[0, T]$ be the span of $\{\phi_i(t)\}$, and expand y_t and $\mu(t)$ in terms of the ϕ_i. Then $\mathscr{P}_1 \equiv \mathscr{P}_0$ if $\mu(t) \in S$ and

$$\sum_{i=1}^{\infty} \frac{\mu_i^2}{\lambda_i} < \infty$$

in which case

$$\frac{d\mathscr{P}_1}{d\mathscr{P}_0} = \exp\left\{ \sum_{i=1}^{\infty} \frac{\mu_i y_i}{\lambda_i} - \tfrac{1}{2} \sum_{i=1}^{\infty} \frac{\mu_i^2}{\lambda_i} \right\} \qquad (5.12)$$

Otherwise, $\mathscr{P}_1 \perp \mathscr{P}_0$. ∎

This is exactly the simple hypothesis detection result stated in Section 1.3.1 [Eq. (1.18)]; indeed, the problems are seen to be same. The case $\mathscr{P}_1 \perp \mathscr{P}_0$ results in "singular detection," since one need only examine the separating set A (see the definition of singularity preceding Theorem B.1 in Appendix B) to decide $\mathscr{P}_0$ or $\mathscr{P}_1$ with probability 1. The possibility of singular detection raises serious questions regarding the Gaussian model; these have been discussed at length in the literature.[2]

---

[1] Much of this theorem is due to Grenander [18].

[2] For example, see Root [49] or Slepian [56].

The dichotomy theorem for the case that the mean and the covariance functions both are different under $\mathscr{P}_1$ and $\mathscr{P}_0$ is much more difficult; its general form was arrived at independently by Hajek [21] and Feldman [15]. More restrictive forms were known much earlier; a good survey is contained in the paper by Yaglom [63]. No benefit is derived from treating the general result here; instead, only the easier case where $\{y_t,\ t \epsilon [0,\ T]\}$ has rational spectral density under $\mathscr{P}_0$ and $\mathscr{P}_1$ will be presented.

Before turning to that result, the following theorem is presented for later reference.

THEOREM 5.3 (Baxter's Theorem)[1]

Let $\{y_t,\ t \epsilon [0,\ T]\}$ be Gaussian with mean $\mu(t)$ and covariance $R(t, s)$, where μ is of bounded variation and where

$$\frac{\partial^2}{\partial t\ \partial s} R(t,\ s) < \infty \tag{5.13}$$

except possibly for $t = s$. Put

$$f^2(t) = \lim_{s \uparrow t} \frac{R(t, t) - R(s, t)}{t - s} - \lim_{s \downarrow t} \frac{R(t, t) - R(s, t)}{t - s} \tag{5.14}$$

Let $\{T_k\}$ be a family of finite parameter subsets of $[0,\ T]$ which

[1] Baxter [4].

grow dense in [0, T] as $k \to \infty$. Then

$$\sum_{i=1}^{k} \left(y_{t_i} - y_{t_{i-1}} \right)^2 \longrightarrow \int_0^T f^2(t)\, dt \qquad (5.15)$$

a.s. as $k \to \infty$. ∎

This theorem is basic to many results on the singularity of Gaussian processes. If y_t is stationary then (5.14) and (5.15) say that, with probability one, the quadratic variation of a sample function is equal to the jump in the derivative of the autocorrelation function at the origin multiplied by the length of the observation interval. It is remarked that a continuous function with nonzero quadratic variation is necessarily of unbounded variation, and hence is not Riemann-Stieltjes integrable with respect to itself, or anywhere differentiable.

(b) Processes with Rational Spectral Density. Suppose that under $\mathscr{P}_0$ and $\mathscr{P}_1$, y_t is a zero-mean stationary Gaussian process with spectral density function

$$S_n(f^2) = \left| \frac{P_n(j2\pi f)}{Q_n(j2\pi f)} \right|^2 \quad ; \quad n = 0, 1 \qquad (5.16)$$

where $j = \sqrt{-1}$, $P_n(z)$ and $Q_n(z)$ are polynomials with no roots in the right-half z plane, and the order of $Q_n(\ )$ exceeds that of $P_n(\ )$. In engineering terms, y_t is the output of a stable realizable

filter with rational transfer function

$$H_n(s) = \frac{P_n(s)}{Q_n(s)} \qquad ; \; n = 0, 1 \tag{5.17}$$

which is driven by unit-density white Gaussian noise. The auto-correlation functions corresponding to (5.16) are

$$R_n(\tau) = \int_{-\infty}^{\infty} S_n(f^2) e^{j2\pi f\tau} \, df \tag{5.18}$$

The dichotomy theorem for processes of this type can be stated as follows:

THEOREM 5.4

$\mathscr{P}_1 \equiv \mathscr{P}_0$ iff

$$\lim_{f\to\infty} \frac{S_1(f^2)}{S_0(f^2)} = \lim_{f\to\infty} \left| \frac{P_1 Q_0}{P_0 Q_1} (j2\pi f) \right|^2 = 1 \tag{5.19}$$

Otherwise, $\mathscr{P}_1 \perp \mathscr{P}_0$. ∎

The necessity of (5.19) for equivalence was established by Slepian [56] ; the theorem in its entirety was established in the aforementioned works of Feldman and Hajek.[1] Slepians observation follows directly from Baxter's Theorem (Thm. 5.3): Let m+1

[1] See also Hajek [22], p. 439.

be the smaller of the differences between the order of the numerator and denominator polynomials in (5.16). Then y_t is m times differentiable and, unless (5.19) is satisfied, the test function of (5.15) when applied to the derivative $y_t^{(m)}$ will converge (with probability one and for arbitrarily small T) to a nonzero number or to zero according as whether $\mathcal{P}_1$ or $\mathcal{P}_0$ is the measure generating the process.

Note that (5.19) requires that the spectral densities $S_0(f^2)$ and $S_1(f^2)$ have the same high-frequency asymptote.

5.2.2 R-N Derivatives for Rational Spectrum Gaussian Processes. Recall that in the estimation problem the observed process is defined on $(\mathcal{Y}, \mathcal{A}, \mathcal{P}_\theta)$, $\mathcal{P}_\theta \in \mathcal{M}$.[1] The quantity of interest is the R-N derivative

$$\lambda(y_t ; \theta) = \frac{d\mathcal{P}_\theta}{d\mathcal{P}_{\theta_0}} (y_t) \tag{5.20}$$

where $\mathcal{P}_{\theta_0}$ dominates $\mathcal{M}$.[2] Theorem 2.3 showed that this quantity yields a sufficient statistic for θ and that a generalization of the factorization criterion applies. Under the assumptions of

---

(1) Again, the compound hypothesis detection problem will be left for a later section. As usual, the notation of H_1 will be used here.

(2) $\mathcal{P}_{\theta_0}$ may or may not be a member of $\mathcal{M}$.

Theorem 2.2, it can be found as the limit of a likelihood ratio. From (2.23) or (3.29), $\lambda(y_t ; \theta)$ may be employed in place of the conditional density function in a generalization of Bayes' rule.

The preceding section gave necessary and sufficient conditions for the R-N derivative to exist (i.e., for $\mathscr{P}_{\theta_0}$ to dominate $\mathscr{M}$) in the case that $\mathscr{P}_{\theta_0}$ and all the $\mathscr{P}_\theta \in \mathscr{M}$ are given by rational spectral densities; if those spectral densities are denoted

$$S_0(f^2) = \left| \frac{P_0(j2\pi f)}{Q_0(j2\pi f)} \right|^2 \tag{5.21}$$

and

$$S_\theta(f^2) = \left| \frac{P_\theta(j2\pi f)}{Q_\theta(j2\pi f)} \right|^2 \tag{5.22}$$

respectively, then it is necessary and sufficient that

$$\lim_{f \to \infty} \left| \frac{P_\theta(j2\pi f)}{Q_\theta(j2\pi f)} \quad \frac{Q_0(j2\pi f)}{P_0(j2\pi f)} \right|^2 = 1 \quad \forall \theta \in \Theta \tag{5.23}$$

The subscript θ is used to denote the spectral parameters regardless which set [see (5.2)-(5.7)] is actually being employed.

Suppose (5.23) is satisfied; it is then necessary to evaluate the R-N derivative. Finding a limit for the likelihood ratio can be extremely difficult and tedious (this is illustrated in Appendix E for

1-SAG noise); luckily, the literature contains general expressions, derived in a variety of ways, for such R-N derivatives.

This chapter is concerned only with M-SAG noise; i.e., $P_\theta(j2\pi f) = a^2$, a positive constant, and the order of $Q_\theta(\cdot)$ is M. R-N derivatives for this type of process take a particularly simple form; the results stated here are due to Hajek.[1] Instead of finding the R-N derivative of $\mathscr{P}_\theta$ with respect to some measure $\mathscr{P}_{\theta_0}$ given by another rational spectral density, Hajek defines the following:[2]

DEFINITION 5.1

Let $\mathscr{P}^+_{M,\alpha}$ denote the Gaussian measure such that under $\mathscr{P}^+_{M,\alpha}$:

(i) The vector $(y_0, y_0', \ldots, y_0^{(M-1)})$ is distributed according to M-dimensional Lebesgue measure.

(ii) $y_t^{(M-1)}$ is a zero-mean Gaussian process of independent increments such that

$$E\,|d\,y_t^{(M-1)}|^2 = \alpha^{-1}\,dt$$

(iii) For all t, $[y_t^{(M-1)} - y_0^{(M-1)}]$ is independent of $y_0, y_0', \ldots, y_0^{(M-1)}$. ∎

---

[1] Hajek [22], Art. 7.

[2] Ibid, p. 433.

Let $\underline{\theta} = (\theta_1, \ldots, \theta_M)^*$ be the set of parameters defined by (5.7). These are equivalent to $(p_1 \ldots p_M)$ or $(q_1 \ldots q_M)$. Recall that $\theta_0 = |a|^{-1}$. Define the state vector

$$\underline{y}_t = (y_t, y_t', \ldots, y_t^{(M-1)})^* \tag{5.24}$$

Hajek shows that if $\{y_t, t\epsilon[0, T]\}$ is to be stationary, the auto-correlation matrix of the "initial condition" $\underline{y}_0$ must be[1]

$$[E\, y_0^{(j)} y_0^{(k)}]_{j,k=0}^{M-1} = [D_{jk}]^{-1} \tag{5.25}$$

where the elements D_{jk} are

$$D_{jk} = \begin{cases} 2 \sum_i (-1)^{j-i} \theta_{M-i} \theta_{M+i-j-k-1} & \text{for } j+k \text{ even} \\ 0 & \text{for } j+k \text{ odd} \end{cases} \tag{5.26}$$

and the sum runs over $\max(0, j+k+1-M) \leq i \leq \min(j, k)$. He then proves

THEOREM 5.5[2]

Let $\{y_t, t\epsilon[0, T]\}$ be a finite segment of a stationary Gaussian process with spectral density (5.7). Then $\mathscr{P}_\theta \equiv \mathscr{P}^+_{M, \theta_0}$

---

[1] Ibid, p. 421.

[2] Ibid, p. 433.

and

$$\frac{d\mathcal{P}_\theta}{d\mathcal{P}^+_{M,\theta_0}} = |D_{jk}|^{\frac{1}{2}} \exp\left\{ \frac{\theta_1 T}{2\theta_0} - \tfrac{1}{2}\sum_{k=0}^{M-1} (-1)^k A_k \int_0^T [y_t^{(k)}]^2\, dt \right.$$
$$\left. - \frac{1}{4} \sum_{\substack{j=0 \\ j+k \text{ even}}}^{M-1} \sum_{k=0}^{M-1} [y_T^{(j)} y_T^{(k)} + y_0^{(j)} y_0^{(k)}] D_{jk} \right\} \tag{5.27}$$

where A_k is the coefficient of $(j2\pi f)^{2k}$ in the denominator of (5.7); i.e., is given by

$$\sum_{m,n=0}^{M} \theta_n \theta_m z^{M-n} (-z)^{M-m} = \sum_{k=0}^{M} A_k z^{2k} \quad \blacksquare \tag{5.28}$$

An explicit expression for A_k is easily found:

$$A_k = \sum_i \theta_i \theta_{2(M-k)-i} (-1)^{M-i} \tag{5.29}$$

where $\max[0, M-2k] \leq i \leq \min[M, 2(M-k)]$.

The fact that the dominating measure is $\mathcal{P}^+_{M,\theta_0}$ rather than a more familiar measure may initially seem bothersome; however, it turns out to be irrelevant since the terms in $\lambda(y_t; \theta)$ or in

$$\frac{d\mathcal{P}_\theta}{d\mathcal{P}^+_{M,\theta_0}}$$

which are unique to the dominating measure will cancel when that R-N derivative is used in Bayes' rule. In fact, the dominating measure of Definition 5.1 is superior because it introduces no extraneous constants into the problem; this, again, will be illustrated for $M = 1$ in Section 5.4.

5.3 Estimation of Noise Parameters and Detection for Arbitrary M.

5.3.1 The Implications of Singularity. Consider first the estimation problem; the uncertain noise parameters will be denoted as θ. Recall from (2.23) that a generalization of Bayes' rule is

$$f(\theta|y_t) = \frac{\frac{d\mathscr{P}_\theta}{d\mathscr{P}_{\theta_0}}(y_t;\theta)\, f_0(\theta)}{\int (\text{Numerator})\, d\theta} \tag{5.30}$$

where $\mathscr{P}_{\theta_0}$ dominates $\mathscr{M}$ and $\mathscr{P}_\theta$ is induced by the M-SAG noise process with parameter θ (the subscript θ will be used regardless of which parameterization is actually employed). If a sufficient statistic exists, the R-N derivative may be factored and one obtains (3.32).

From Theorem 5.4, it is clear that the measure $\mathscr{P}_{\theta_0}$ induced by an N-SAG noise process is equivalent to $\mathscr{P}_\theta$ if and only if $M = N$ and the numerators a^2 of the corresponding spectral densities (5.2) or (5.6) are identical; this means that the

parameter a^2 may not be included in θ . If it is included, then one can always construct a test function which converges to a^2 with arbitrarily small error for arbitrarily small observation intervals $[0, T]$. Since a^2 characterizes the properties of $\{y_t, t \in [0, T]\}$ for very high frequencies, this is heuristically reasonable. Physically, error in estimation of a^2 is bounded below by the ability to sample y_t at arbitrarily high rates; the argument is the same one found in discussions of singular Gaussian noise-in-noise detection. Indeed, the problem is the same since the "noise-in-noise" detection statistic is precisely the R-N derivative of (5.30).

Suppose the dominating measure is $\mathscr{P}^+_{M, \theta_0}$ as in Hajek's Definition 5.1. From (5.7), $\theta_0 = |a|^{-1}$ so that again the constant a^2 cannot be estimated but, rather, must be known. Section 5.4 will illustrate the singular estimation of a^2 for the case $M = 1$ by using Baxter's Theorem (Theorem 5.3).

Now consider the detection problem. Since a simple-hypothesis (parameters-known) solution is available (see Section 1.3.1), equations (3.37) or (3.38) are the obvious candidates for a solution. Recall the comment made at the beginning of Section 4.2; the parameters under either hypothesis are $\theta = (\theta_1 \cdots \theta_M)$ [or equivalently, $(q_1 \cdots q_M)$ or $(p_1 \cdots p_M)$] and the notation which indicates that their distributions evolve differently under H_0 and H_1 will

be dropped. From (5.2), it is clear that for any $\theta \in \Theta$ the measure induced on $(\mathcal{Y}, \mathcal{A})$ by $\{y_t, t \in [0, T]\}$ under H_0 is identical to the measure induced by $\{y_t - s(t), t \in [0, T]\}$ under H_1.[1] The R-N derivative of the former with respect to $\mathcal{P}^+_{M, \theta_0}$ is given by (5.27), and the R-N derivative of the latter is the same expression with $y_t - s(t)$ replacing y_t.

Clearly, a^2 (or θ_0) cannot be estimated under either hypothesis. If this parameter is known and one makes the necessary additional assumptions on $s(t)$,[2] then (3.37) [with (5.30) used for the a posteriori p.d.f.'s] provides a well-defined detection statistic.

If the dominating measures used in (5.30) are not $\mathcal{P}^+_{M, \theta_0}$ but are some other rational spectrum Gaussian measures satisfying (5.23), then from Theorem 5.4 they are equivalent to $\mathcal{P}^+_{M, \theta_0}$ and the result is unchanged. This is so because then

$$\frac{d\mathcal{P}_\theta}{d\mathcal{P}_0} = \frac{d\mathcal{P}_\theta}{d\mathcal{P}^+_{M, \theta_0}} \frac{d\mathcal{P}_{M, \theta_0}}{d\mathcal{P}_0} \tag{5.31}$$

and the second term cancels out of (5.30). Section 5.4 will specifically illustrate this for $M = 1$.

---

[1] This assumes that $s(t)$ has at least the same smoothness properties as almost every sample function of $\{y_t\}$; this assumption is weaker than the assumptions necessary to guarantee a solution to the simple-hypothesis problem, and hence is of no concern.

[2] See the beginning of Appendix A; $s(t)$ must be (2M)-times differentiable on the interval $(0, T)$.

5.3.2 Estimation of Noise Parameters. The notation of H_0 will be used. As is by now clear, all that is said can be repeated for H_1 if y_t is replaced by $y_t - s(t)$.

Henceforth, the R-N derivative of (5.27) will be denoted $f(y_t|\theta)$. This is only a slight abuse of notation since this quantity is a direct generalization of the usual notion of a p.d.f. (which is the R-N derivative of a probability measure with respect to Lebesgue measure). The ultimate goal is to factor this R-N derivative as in (2.22), recognize the sufficient statistics, and determine the natural conjugate class of p.d.f.'s on Θ.[1] From previous results and discussions (e.g., Sections 1.1, 2.3, and 3.3.3) one conjectures that this class will be much simpler and more useful for sequential estimation if $f(y_t|\theta)$ is first divided by ("made conditional to") the p.d.f. of the initial state $\underline{y}_0 = (y_0, y_0', \ldots y_0^{(M-1)})^*$.[2] Recall that

---

[1] After the factorization of (3.31) the remaining expression $g[t(y_t); \theta]$ represents a bona-fide p.d.f. on the finite-dimensional space of the sufficient statistics.

[2] In the finite-dimensional case, $f(\underline{Y}_k|\theta)$ was divided by $f(\underline{y}_0|\theta)$ and $\underline{y}_0$ consisted of M discrete samples. It seems reasonable, though no attempt will be made to prove, that passing to the limit after such a procedure yields the same result as above (Section 5.4 will illustrate that this is true for M=1). The question is of little relevance, since the primary concern is with the functional form of the densities involved. If the above procedure is arbitrarily adopted, then all the results of Section 4.2.3 can be justified.

$$f(\underline{y}_0 | \theta) = [(2\pi)^{-M} |D_{jk}|]^{\frac{1}{2}} \exp\left\{-\tfrac{1}{2}\, \underline{y}_0^{*} [D_{jk}] \underline{y}_0\right\} \qquad (5.32)$$

The quotient of (5.27) by (5.32) is the desired "conditional density"

$$f(y_t | \theta, \underline{y}_0) = (2\pi)^{\frac{M}{2}} \exp\left\{ \frac{T\theta_1}{2\theta_0} - \tfrac{1}{2} \sum_{k=0}^{M-1} (-1)^k A_k \int_0^T [y_t^{(k)}]^2 dt \right. \qquad (5.33)$$

$$\left. - \frac{1}{4} \sum_{\substack{j=0 \\ j+k \text{ even}}}^{M-1} \sum_{k=0}^{M-1} [y_T^{(j)} y_T^{(k)} - y_0^{(j)} y_0^{(k)}] D_{jk} \right\}$$

The following facts are obvious by inspection:

(i) Sufficient statistics for the estimation of $\theta = (\theta_1, \ldots, \theta_M)$ in (5.7) are:

$$\left\{ \int_0^T [y_t^{(k)}]^2 dt \;,\; y_0^{(k)} \;,\; y_T^{(k)} \right\} \; ; k=0, 1 \ldots M-1 \qquad (5.34)$$

(ii) With the results conditioned on $\underline{y}_0$, the natural conjugate class of densities for $\theta = (\theta_1 \ldots \theta_M) = (q_1/|a|, \ldots, q_M/|a|)$ is an M-variate Gaussian p.d.f. (Note from (5.26) and (5.28) that all terms in the exponent of (5.33) are either linear or quadratic in θ.) This density is truncated to be zero on that region

of R^M where the roots of

$$\theta_0 z^M + \theta_1 z^{M-1} + \cdots + \theta_M = 0$$

have positive real parts; the complement of that region is denoted Θ .[1]

It is shown in Section D.4 of Appendix D that (5.33) can be re-written as follows: Define the parameter vector

$$\underline{\theta} = (\theta_1 \cdots \theta_M)^* \tag{5.35}$$

and the functions of the observation

$$J^{(k)}(y_t) = \int_0^T [y_t^{(k)}]^2 \, dt \tag{5.36}$$

$$E^{(j,k)}(y) = y_T^{(j)} y_T^{(k)} - y_0^{(j)} y_0^{(k)} \tag{5.37}$$

Use these sufficient statistics to write an MxM , positive definite, symmetric matrix $T(y_t)$ whose elements are

---

[1] Note that the natural conjugate density written in terms of $(p_1 \cdots p_M)$ is much more complicated, see (5.5) ff., but is defined on a much simpler region. (Namely, $Re(p_i) < 0$ for all i .)

(5.38)

$$t_{ij}(y_t) = \begin{cases} (-1)^{\frac{3i+j}{2}} J^{(M - \frac{i+j}{2})}(y_t) & ; \ i+j \text{ even} \\ \frac{1}{2} \sum_{k}' (-1)^{M-k-1-\min(i,j)} \cdot E^{(2M-i-j-k-1,\ k)}(y) & ; \ i+j \text{ odd} \end{cases}$$

where $\max(0, M-i-j) \leq k \leq \min(M-1, 2M-i-j-1)$. Also, define the M-vector $\underline{t}(y_t)$ whose elements are:[1]

$$t_j(y_t) = \begin{cases} 2\theta_0 t_{01}(y_t) - \frac{T}{2\theta_0} & ; \ j = 1 \\ 2\theta_0 t_{0j}(y_t) & ; \ j = 2 \ldots M \end{cases} \tag{5.39}$$

then the "conditional density" for use in Bayes' rule is

$$f(y_t | \underline{\theta}, \underline{y}_0) = (2\pi)^{\frac{M}{2}} \exp\left\{ -\tfrac{1}{2} [\underline{\theta}^* T(y_t) \underline{\theta} + \underline{t}^*(y_t) \underline{\theta}] \right\} \tag{5.40}$$

Clearly, the natural conjugate density has the same form but involves an MxM conjugate parameter matrix Ψ and a conjugate parameter M-vector $\underline{\psi}$; these are then updated through

---

[1] Again, the use of t for "time" and $t(\cdot)$ for a sufficient statistic, and of T for observation period and $T(\cdot)$ for a sufficient statistic, will hopefully cause no confusion.

$$\Psi_T = \Psi_0 + T(y_t)$$
$$\underline{\Psi}_T = \underline{\Psi}_0 + \underline{t}(y_t) \tag{5.41}$$

Their components are functionally independent only to the extent that those of $T(y_t)$ and $\underline{t}(y_t)$ are.

Consider now the sequential processing of finite subintervals as discussed in Section 3.3.3. Define

$$\underline{y}_i \stackrel{d}{=} \underline{y}_{T_i}$$
$$y_{(i-1,\, i]} \stackrel{d}{=} \{y_t;\ T_{i-1} < t \leq T_i\} \tag{5.42}$$

and similar notation for the derivatives of y_t. Since sample functions of y_t and their first M-1 derivatives are a.s. uniformly continuous, it is irrelevant whether one considers open or half-closed intervals. If (5.36) is modified to run over the interval $(T_{i-1},\ T_i)$, the endpoints in (5.37) changed accordingly, the numerator of the first term of (5.39) changed from "T" to "$T_i - T_{i-1}$", and appropriate notational changes are made, then the density of (5.40) can be written

$$f[y_{(i-1,\, i)} | \underline{\theta},\ \underline{y}_{i-1}] = (2\pi)^{\frac{M}{2}} \exp\left\{ -\tfrac{1}{2}\left[\underline{\theta}^* T(y_{(i-1,\, i)})\, \underline{\theta} + \underline{t}^*(y_{(i-1,\, i)})\, \underline{\theta}\right]\right\} \tag{5.43}$$

The unconditioned density of (5.33) could be similarly modified; from either one, it is easy to verify the relations

$$f[y_{(0,k)}|\underline{\theta}, \underline{y}_0] = \prod_{i=k}^{k} f[y_{(i-1,i)}|\underline{\theta}, \underline{y}_{i-1}] \qquad (5.44)$$

By definition (see (5.33)),

$$f[y_{(0,k)}|\underline{\theta}] = f[y_{(0,k)}|\underline{\theta}, \underline{y}_0]\, f(\underline{y}_0|\underline{\theta}) \qquad (5.45)$$

The form of all these expressions is precisely the same as in the finite dimensional Markov case and so the formal manipulations (including forming the natural conjugate density, updating its parameter by using the sufficient statistic, and accounting for the conditioning on $\underline{y}_0$) may proceed in exactly the same fashion.

5.3.3 Detection in Noise of Unknown Parameters. Recall the discussion concerning the detection problem begun in Section 5.3.1. If one considers (3.37) and uses (5.30) with the appropriate numerator[1] for the a posteriori p.d.f.'s on θ , one obtains the non-sequential (not conditioned on $\underline{y}_0$) detection statistic. The a priori densities are seen to cancel, and there remains

---

[1] i.e., $f(y_t|\theta) f_0(\theta)$ for H_0 and $f(y_t - s(t)|\theta) f_1(\theta)$ for H_1 , where $f(\cdot|\theta)$ is the density defined by (5.27).

$$\ell(y_t) = \frac{\int_\Theta f(y_t - s(t)|\lambda)\, f_1(\lambda)\, d\lambda}{\int_\Theta f(y_t|\lambda)\, f_0(\lambda)\, d\lambda} \cdot \frac{f(y_t|\theta)}{f(y_t - s(t)|\theta)}\, \ell(y_t|\theta) \tag{5.46}$$

As has been previously remarked (e.g., following (1.38) or (3.25)), the parameters θ must cancel out of this expression; alternately, it may be evaluated for any admissible fixed value of θ. Owing to the complexity of the terms (recall that $\ell(y_t|\theta)$ is the simple-hypothesis solution as given by the Metzger model or by other "classical" techniques), this cancellation will not be verified for arbitrary values of M. The following sections will do so for $M=1$ and $M=2$; they will also present results equivalent to (5.46) but stated in terms of the natural conjugate densities and thus useful for sequential processing.

5.4 The Ornstein - Uhlenbeck Process: M = 1

Consider the 1-SAG noise process, i.e., the zero-mean stationary Gaussian process $\{n_t,\ t \in [0,\ T]\}$ with spectral density function

$$S_n(f^2) = \frac{a^2}{(2\pi f)^2 + p_1^2} \tag{5.47}$$

$$= \left| \frac{a}{(j2\pi f) + q_1} \right|^2 \tag{5.48}$$

where, in this case, $p_1 = q_1$ so that there is no difference between the parametrizations. (The parameter will be called q_1 from here on.) The corresponding autocorrelation function is

$$R_n(\tau) = \frac{a^2}{2q_1} e^{-q_1|\tau|} \quad ; \; q_1 > 0 \tag{5.49}$$

To illustrate the previously claimed singularity in the estimation of a^2, consider that n_t is observed on $[0, T]$ and is sampled as usual. Note that

$$R_n'(0^-) - R_n'(0^+) = a^2 \tag{5.50}$$

where the indicated derivatives are limits from the left and right respectively. From Theorem 5.3 (Baxter's Theorem),

$$\lim_{k \to \infty} \sum_{i=1}^{k} (n_i - n_{i-1})^2 = a^2 T \qquad \text{a.s.} \tag{5.51}$$

where n_i is the noise sample at $t = iT/k$. Since the convergence is strong (i.e., holds for almost every sample function), the discrete test function of (5.51) can be used to estimate a^2 with arbitrarily small error for arbitrarily small T. (Physically, of course, one is limited by the ability to sample n_t at very high rates). It should be noted that since a.e. sample function of n_t is uniformly continuous on $[0, T]$, (5.51) implies that a.e. sample function is of unbounded variation.[1]

By direct substitution into Hajek's result (5.27) one finds the desired Radon-Nikodym derivative

$$f(n_t \mid \theta) = \frac{d\mathscr{P}_\theta}{d\mathscr{P}^+_{1,\theta_0}}$$

$$= (2\theta_1\theta_0)^{\frac{1}{2}} \exp\left\{ \frac{T\theta_1}{2\theta_0} - \frac{\theta_1^2}{2} \int_0^T n_t^2\, dt \right.$$

$$\left. - \frac{1}{2}\,\theta_1\theta_0(y_T^2 + y_0^2) \right\} \qquad (5.52)$$

where θ_0, θ_1 are parameters of the spectral density written as in (5.7); recall $\theta_0 = |a|^{-1}$ is not estimable and must be known.

---

[1] See Appendix D, Section D.5.

In terms of the parameter $q_1 = \theta_1 |a|^{-1}$, this can be written

$$f(n_t | q_1) = \frac{\sqrt{2q_1}}{a} \exp\left\{ \frac{q_1 T}{2} - \frac{1}{2a^2} \left[q_1^2 \int_0^T y_t^2 dt + q_1 (y_T^2 + y_0^2) \right] \right\} \tag{5.53}$$

This can be considered as a R-N derivative or it can be normalized and viewed as a bona-fide p.d.f. on the "sufficient space" with coordinates $\{ \int n_t^2 dt \ , \ n_0 \ , \ n_T \}$.

It was claimed at the end of Section 5.3.1 that the same result is obtained if one evaluates the R-N derivative with respect to an equivalent measure (given by another rational spectral density) as the limit of a likelihood ratio function (Section 2.2.2).[1] This is verified in Appendix E using the following procedure: Let $\mathscr{P}_*$ be the measure induced by a 1-SAG process with fixed parameter $q^* > 0$ instead of q_1 ; $\mathscr{P}_*$ is equivalent to $\mathscr{P}_\theta$. Sample the process n_t as usual; the sequence of samples is the solution to a first order autoregression whose parameters may be found in terms of q_1 or q^* and the sampling interval. Thus the likelihood ratio function (conditional to n_0) may be evaluated using (4.48); its limit can be found, but it is necessary to explicitly employ Baxter's

[1] This was first done rigorously by Striebel [60] .

result (5.51) in the process. This limit is $\lambda(n_t | n_0 ; q_1)$ as defined in Theorem 2.3 but conditional to n_0 (i.e., its factorization yields the density for sequential processing). Multiplication by $f(n_0|q_1)$ gives the R-N derivative $\lambda(n_t ; q_1)$, which still contains the "nuisance parameter" q^* ; this cancels out in Bayes' rule (5.30), and the result is identical to Hajek's density (5.53).

5.4.1 Estimation. Suppose the observation is noise, i.e., H_0 is true. To obtain the sequential natural conjugate density on q_1, (5.53) must be conditioned on y_0. From (5.32), or by noting that

$$R_n(0) = \frac{a^2}{2q_1} \tag{5.54}$$

one finds the p.d.f. of the initial sample

$$f(y_0|q_1) = \left(\frac{q_1}{\pi a^2}\right)^{\frac{1}{2}} \exp\left[-\frac{q_1}{a^2} y_0^2\right] \tag{5.55}$$

Dividing (5.53) by (5.55), or merely putting $M = 1$ in (5.33), yields

$$f(y_t|q_1, y_0) = \sqrt{2\pi} \exp\left\{-\frac{1}{2a^2}\left[q_1^2 \int_0^T y_t^2\, dt - q_1\left(a^2 T + y_0^2 - y_T^2\right)\right]\right\} \tag{5.56}$$

By inspection (recall Theorem 3.1), the conjugate density on q_1 is Gaussian,[1]

$$p(q_1;\ \underline{\psi}) = K(\underline{\psi}) \exp\left\{ -\frac{1}{2a^2}\left[q_1^2 \psi_1 - q_1 \psi_2 \right] \right\},\quad q_1 > 0 \tag{5.57}$$

where $\psi_1 > 0$. Its parameter is updated through

$$\underline{\psi}^{(T)} = \underline{\psi}^{(0)} + \begin{bmatrix} \int_0^T y_t^2\, dt \\ a^2 T + y_0^2 - y_T^2 \end{bmatrix} \tag{5.58}$$

If observations are made on sequential subintervals, (5.56) may be suitably modified and the conjugate parameter updating equation becomes, in the previous notation,

$$\underline{\psi}^{(T_i)} = \underline{\psi}^{(T_{i-1})} + \begin{bmatrix} \int_{T_{i-1}}^{T_i} y_t^2\, dt \\ a^2 (T_i - T_{i-1}) + y_{i-1}^2 - y_i^2 \end{bmatrix} \tag{5.59}$$

---

[1] To be totally consistent with Chapter III, this density would have to be denoted $q(\cdot\ ; \underline{\psi})$ since H_0 is under consideration. However, it is clear that the natural conjugate class is Gaussian under either hypothesis and only the parameters ($\underline{\psi}$ for H_0, $\underline{\gamma}$ for H_1) differ.

Analysis of the natural conjugate density $p(q_1 ; \underline{\psi})$ is straightforward; it is a truncated Gaussian p.d.f. with mean and variance prior to truncation of

$$\mu(\underline{\psi}) = \frac{\psi_2}{2\psi_1} \quad , \quad \sigma^2(\underline{\psi}) = \frac{a^2}{\psi_1} \tag{5.60}$$

respectively. Once it is truncated to the positive half-line, its normalizing constant is

$$K(\underline{\psi}) = \frac{\sqrt{\psi}_1}{a} \; \Omega\left(\frac{\psi_2}{2a\sqrt{\psi}_1}\right) \tag{5.61}$$

where the function $\Omega(\cdot)$ is the logarithmic derivative of the Gaussian c.d.f. and is defined by[1]

$$\Omega(x) = \frac{\phi(x)}{\Phi(x)} \quad , \quad -\infty < x < \infty$$

where $\phi(x)$ is the unit Gaussian p.d.f. and $\Phi(x)$ the corresponding cumulative distribution function. The mean and variance of $p(q_1 ; \underline{\psi})$ are given by:

---

[1] The utility of this function in certain detection problems was first noticed by Roberts [47], p. 68.

$$E(q_1|\underline{\psi}) = \frac{\psi_2}{2\psi_1} + \frac{a}{\sqrt{\psi_1}}\,\Omega\left(\frac{\psi_2}{2a\sqrt{\psi_1}}\right) \tag{5.62}$$

$$E(q_1^2|\underline{\psi}) = \frac{a^2\psi_2}{2\psi_1^2} + \left(\frac{a^2}{\psi_1}\right)^{\frac{3}{2}}\left(1 + \frac{\psi_2^2}{4a^2\psi_1}\right)\Omega^{-1}\left(\frac{\psi_2}{2a\sqrt{\psi_1}}\right) \tag{5.63}$$

$$\mathrm{var}(q_1|\underline{\psi}) = E(q_1^2|\underline{\psi}) - E^2(q_1|\underline{\psi}) \tag{5.64}$$

and are thus quite complicated; under suitable conditions, the mode $\mu(\underline{\psi})$ of the untruncated p.d.f. may be a reasonable estimate.

Any reasonable estimate may be shown to be consistent; using $\underline{\psi}^{(T)}$ as given by (5.58) in (5.60) shows that the a posteriori p.d.f. on q_1 is Gaussian with mean and variance prior to truncation of

$$\mu_T = \frac{\psi_2^{(0)} + a^2T + y_0^2 - y_T^2}{2\psi_1^{(0)} + 2\int_0^T y_t^2\,dt}$$

$$\sigma_T^2 = a^2\left[\psi_1^{(0)} + \int_0^T y_t^2\,dt\right]^{-1} \tag{5.65}$$

Let $T \to \infty$ and recall that

$$\underset{T\to\infty}{\text{l.i.m.}}\; T^{-1} \int_0^T y_t^2\, dt = R_n(0) = \frac{a^2}{2q_1^*} \tag{5.66}$$

where q_1^* is the "true" value of the parameter for the process generating the observation; clearly

$$\mu_T \to q_1^* \quad ; \quad \sigma_T^2 \to 0 \quad \text{as} \quad T \to \infty$$

and the a posteriori p.d.f. tends to a "delta function" at q_1^* .

If desired, the conditioning on y_0 can be undone by any of the methods mentioned in Section 4.2.3.

5.4.2 Detection. It is again clear that under H_1 all the above results hold; the parameter of the natural conjugate p.d.f. is then $\underline{\gamma}$, its a priori value may be different from $\underline{\psi}^{(0)}$, and it is updated through

$$\underline{\gamma}^{(T)} = \underline{\gamma}^{(0)} + \begin{bmatrix} \int_0^T [y_t - s(t)]^2\, dt \\ a^2 T + [y_0 - s(0)]^2 - [y_T - s(T)]^2 \end{bmatrix} \tag{5.67}$$

The R-N derivatives of the observation may be written as in (5.53) or (5.56), with y_t replaced by $y_t - s(t)$.

The detection statistic is given by (3.37) or (3.38); if one wishes to detect sequentially and use the natural conjugate class, the conditioning on y_0 must be accounted for since the simple-hypothesis statistic $\ell(y_t \mid q_1)$ given in (1.25) was not conditioned on y_0. One way to do this is, as discussed in Section 4.3.2 (a), to choose an a priori (to y_0) p.d.f. from the class

$$p_0(q_1 ; \underline{\psi}^-) = K_0(\underline{\psi}^-) q_1^{-\frac{1}{2}} \exp\left\{ - \frac{1}{2a^2}\left[q_1^2 \psi_1^- - q_1 \psi_2^- \right] \right\} \tag{5.68}$$

in which case the a posteriori (to the discrete observation y_0) p.d.f. will be natural conjugate and hence the a posteriori (to y_t, $t \in [0, T]$) p.d.f. is also natural conjugate but is no longer conditioned on y_0; its parameter is found using (5.58) or (5.67).

Appendix E gives the calculations involved in simplifying the detection statistic; it is shown that the parameter q_1 does indeed cancel, and one finds

$$\ell(y_t) = \frac{K_0(\underline{\gamma}^-)}{K_0(\underline{\psi}^-)} \frac{K(\underline{\psi}^{(T)})}{K(\underline{\gamma}^{(T)})} \exp\left\{ - a^{-2}\left[\int_0^T s''(t) y_t \, dt + \tfrac{1}{2}\int_0^T [s'(t)]^2 \, dt - s'(0) y_0 - s'(T) y_T \right] \right\} \tag{5.69}$$

where $K(\cdot)$ is the normalizing constant in (5.57) and $K_0(\cdot)$ normalizes (5.68).

The second ratio of constants involves y_t and thus represents processing of the observation. Eq. (E.26) explicitly shows the resulting terms; even for the case $M=1$, the processing necessary to obtain the optimal detection statistic is extremely complicated.

If one does not want to use the a priori p.d.f. of (5.68), then (5.46) may be used directly; equivalently, the natural conjugate densities may be employed and the results modified as in Section 3.2.2. In either case, the results are similar to the above.

5.5 Estimation and Detection: 2-SAG Noise

The case $M = 2$, though not of great interest for its own sake, is more indicative of the general case than $M = 1$. The various parameterizations of the spectral density (see (5.2), (5.4), and (5.7)) are related by

$$\begin{aligned} q_1 &= p_1 + p_2 \\ q_2 &= p_1 p_2 \end{aligned} \tag{5.70}$$

and

$$\begin{aligned} \theta_i &= q_i/|a| \quad ; \quad i = 1, 2 \\ \theta_0 &= 1/|a| \end{aligned} \tag{5.71}$$

The autocorrelation function is written in (A.22). As usual, let $\mathscr{P}_\theta$ represent the measure induced by the process regardless of which parametrization is being employed.

It is by now clear that instead of evaluating $d\mathscr{P}_\theta / d\mathscr{P}_{\theta_0}$ as the limit of a likelihood ratio, it is more convenient to employ $d\mathscr{P}_\theta / d\mathscr{P}^+_{2,\theta_0}$ directly.[1] From (5.27) and (5.33), the desired R-N derivatives (densities of the sufficient statistics) are, in the usual notation

$$f(y_t|\underline{q}) = \frac{2q_1\sqrt{q_2}}{a^2} \exp\left\{\frac{q_1 T}{2} - \frac{1}{2a^2}\left[q_2^2 \int_0^T y_t^2\,dt + (q_1^2 - 2q_2)\int_0^T (y_t')^2\,dt + q_1 q_2 (y_T^2 + y_0^2) + q_1 (y_T'^2 + y_0'^2)\right]\right\} \tag{5.72}$$

and

$$f(y_t|\underline{y}_0, \underline{q}) = 2\pi \exp\left\{\frac{q_1 T}{2} - \frac{1}{2a^2}\left[q_2^2 \int_0^T y_t^2\,dt + (q_1^2 - 2q_2)\int_0^T (y_t')^2\,dt + q_1 q_2 (y_T^2 - y_0^2) + q_1 (y_T'^2 - y_0'^2)\right]\right\} \tag{5.73}$$

---

[1] There is some ambiguity here since the first case "θ_0" represents any arbitrary Gaussian measure equivalent to $\mathscr{P}_\theta$, while in the second case θ_0 refers to the number $|a|^{-1}$. This will not present cause for further confusion.

The latter equation can be re-written, by inspection or using (5.36)-(5.40), as

$$f(y_t | \underline{y}_0, \underline{q}) = 2\pi \exp\left\{ - \frac{1}{2a^2} \left[\underline{q}^* T(y_t) \underline{q} + \underline{q}^* \underline{t}(y_t) \right] \right\} \tag{5.74}$$

where

$$\underline{q} = (q_1, q_2)^* \tag{5.75}$$

$$\underline{y}_0 = (y_0, y_0')^* \tag{5.76}$$

$$T(y_t) = \begin{bmatrix} \int_0^T (y_t')^2 \, dt & \frac{y_T^2 - y_0^2}{2} \\ \frac{y_T^2 - y_0^2}{2} & \int y_t^2 \, dt \end{bmatrix} \tag{5.77}$$

$$\underline{t}(y_t) = \begin{bmatrix} y_T'^2 - y_0'^2 - a^2 T \\ -2 \int_0^T (y_t')^2 \, dt \end{bmatrix} \tag{5.78}$$

The natural conjugate class of densities is

$$p(\underline{q} ; \psi) = K(\psi) \exp\left\{ - \frac{1}{2a^2} \left[\underline{q}^* \Psi \, \underline{q} + \underline{q}^* \underline{\psi} \right] \right\}$$

$$, \ \underline{q} \in \Theta \tag{5.79}$$

where Θ is the set on which the roots of

$$z^2 + q_1 z + q_2 = 0$$

have negative real parts; this is illustrated in Fig. 5.1. $K(\psi)$ involves integrating over Θ and will not be evaluated.

From (5.77) and (5.78), it is clear that for a minimal parameterization one must put

$$\Psi = \begin{bmatrix} \psi_1 & \psi_3 \\ \psi_3 & \psi_2 \end{bmatrix} \quad , \quad \underline{\psi} = \begin{bmatrix} \psi_4 \\ -2\psi_1 \end{bmatrix} \tag{5.80}$$

which contains only four independent parameters; or, as commented in Theorem 3.1, one can let the vector $\underline{\psi}$ of conjugate parameters be arbitrary and still remain within the class as observations are processed. In either case, the parameters are updated as in (5.41).

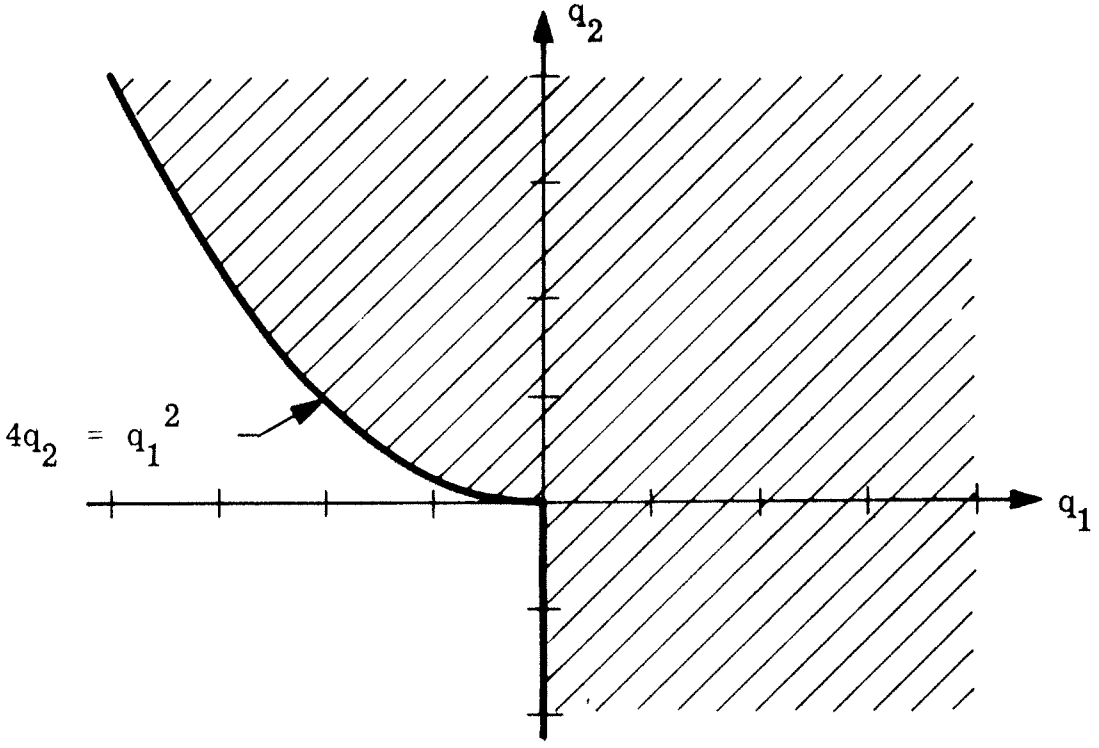

Fig. 5.1. Parameter Space for 2-SAG Noise

By completing the square, the natural conjugate density (5.79) can be re-written

$$p(\underline{q}\,;\psi) = K(\psi) \exp\left\{ \frac{1}{2a^2} \underline{\mu}^* \Psi \underline{\mu} \right\} \exp\left\{ -\frac{1}{2a^2} (\underline{q} - \underline{\mu})^* \Psi (\underline{q} - \underline{\mu}) \right\}, \quad \underline{q} \in \Theta \tag{5.81}$$

where

$$\underline{\mu} = -\tfrac{1}{2} \Psi^{-1} \underline{\psi} \tag{5.82}$$

is the mode of the density provided that $\underline{\mu} \in \Theta$. This may provide a reasonable estimate of $\underline{q}$,

$$\hat{\underline{q}}_{MAP}^{(T)} = -\tfrac{1}{2} \left\{ \Psi^{(T)^{-1}} \underline{\psi}^{(T)} \right\} \quad \text{if } \hat{\underline{q}}_{MAP} \in \Theta \tag{5.83}$$

and is analogous to (5.65).

The detection problem is solved in complete analogy with the work of Section 5.4.2. Once again, the required cancellation with the simple-hypothesis solution as given by (A.51) [recall that $z(y) = \ell n\, \ell(y_t | \underline{q})$] occurs. If $f_1(\underline{q})$ and $f_0(\underline{q})$ are arbitrary *a priori* densities under H_1 and H_0 , then using (5.46) one finds that

$$\ell(y_t) = \frac{K_1(y_t)}{K_0(y_t)} \exp\left(S(y_t)\right) \tag{5.84}$$

where $K_i(y_t)$, $i = 0, 1,$ are normalizing constants for the a posteriori p.d.f., and where

$$S(y_t) = -a^{-2}\left\{ \int_0^T y_t' \, s^{(3)}(t)dt + \tfrac{1}{2} \int_0^T [s''(t)]^2 \, dt + y_0' \, s''(0) - y_T' \, s''(T) \right\} \tag{5.85}$$

If one chooses the a priori p.d.f. from the class

$$p_0(\underline{q}; \psi^-) = K_0(\psi^-) \frac{1}{q_1\sqrt{q_2}} \exp\left\{ -\frac{1}{2a^2}[\underline{q}^* \Psi \underline{q} + \underline{\psi}^* \underline{q}] \right\} \tag{5.86}$$

and treats the discrete observation $\underline{y}_0$ as usual to obtain a natural conjugate situation which is not conditioned on $\underline{y}_0$, then the detection statistic is

$$\ell(\underline{y}_t) = \frac{K_0(\gamma^-)}{K_0(\psi^-)} \frac{K(\psi^{(T)})}{K(\gamma^{(T)})} \exp\left(S(y_t)\right) \tag{5.87}$$

In both (5.84) and (5.87) the ratios of normalizing constants represent complicated signal processing, so that the expressions are

not as simple as they first appear (recall (E.26) for the simpler case $M = 1$).

CHAPTER VI

SUMMARY AND CONCLUSIONS

6.1 Narrative Summary[1]

6.1.1 Problem Statement: General Solution. The problem statement postulates two mutually exclusive and exhaustive hypotheses (#1.2), under each of which a statistical description of the observation is known except for a finite-dimensional parameter. These parameters, which index families of probability distributions on the observation space, are considered random variables; they may or may not have common components.

If the hypothesis is specified, one is left with the problem of estimating the corresponding parameter based upon the observation(s) (#1.2.2). For the present purpose, this estimation problem is considered solved when the a posteriori p.d.f. based on Bayes' rule (1.27) is known. Suppose neither hypothesis is specified and one is interested in deciding which one is true. For a large class of problems (#1.3.3) the likelihood ratio of marginal observation p.d.f.'s (1.36) is an optimal detection statistic. Using

---

[1] Throughout this chapter, references to section numbers are placed in parentheses and preceded by #, and references to equations are merely placed in parentheses.

Bayes' rule, the marginal density for either hypothesis can be found as the product of the ratio of a priori to a posteriori parameter p.d.f. and the conditional observation p.d.f. (1.29). The expression can be evaluated at arbitrary, fixed values of the parameter. Thus, the marginal likelihood ratio is given by a similar expression (1.38) which also does not depend on the parameters, but which involves the related simple-hypothesis likelihood ratio. One finds that the solutions to the two related estimation problems (one under each hypothesis) serve quite naturally to solve the detection problem, provided that the simple-hypothesis result is known.

With no further assumptions, all the conditional observation densities and a priori and a posteriori parameter densities must be evaluated, stored, and manipulated as functions (or as discrete approximations to functions); furthermore, all the observations must be saved. This clearly makes the procedure intractable, especially for recursive processing. It becomes tractable, however, if the observation distributions admit finite and fixed-dimensional sufficient statistics for the unknown parameters. Before pursuing the subject, Chapter II was used to investigate the concept of sufficient statistics.

6.1.2 Necessary and Sufficient Statistics. Heuristically, the concept is best understood by considering sets of observations such that knowing the set into which an observation falls is equivalent

(for the purpose of finding the a posteriori p.d.f. on the parameters) to knowing the observation itself. A class of such sets is called sufficient for estimating the parameter; the coarsest such class is called necessary and sufficient. A mapping constant on the sets of such a class is called a sufficient or a necessary and sufficient statistic. This approach serves well in the trivial case of finite spaces (#2.1.1) and in the difficult case of infinite-dimensional probability spaces (Appendix B). For the intermediate case considered by classical probability and statistics, the concept is formulated differently but equivalently.

If the observation space is finite dimensional and the conditional probability distribution of the observation admits a density w.r.t. Lebesgue measure (as will be assumed), then a sufficient statistic can (if it exists) be found by factoring that density (2.10). Equivalently, one can factor the ratio of that density to another probability density defined on the same space; if this second density happens to be a member of the family of conditional p.d.f.'s being investigated, then that ratio is called the likelihood ratio function (2.12) and its factorization always yields a necessary and sufficient statistic (Theorem 2.1). Bayes' rule is unchanged if one replaces the conditional observation density with the likelihood ratio function (1.30).

To establish results needed in Chapters IV and V, the existence of sufficient statistics when sequential observations possess an M^{th}-order Markov dependence was studied in detail (#2.3). If all p.d.f.'s are conditioned on M initial samples $\underset{\sim}{y}_0$, then the joint conditional p.d.f. of k samples is a k-fold product of transition densities (1.5) and is very similar in form to that of independent samples (1.7); results for the latter case are readily available in the literature. The similarity was pursued. Under suitable restrictions (2.24 ff.), only those processes possessing exponential class transition densities (2.33) admit sufficient statistics of fixed dimension; these statistics may be found by inspection of the transition density and can be updated recursively based on the M-dimensional observation "state vector." It must be borne in mind that all the results are conditioned on $\underset{\sim}{y}_0$ and that this must ultimately be undone or justified.

If sufficient statistics exist, the observations themselves need not be saved and the memory necessary for storage of the observation has fixed size regardless of the number of samples processed.

Statistics which are sufficient for estimation may or may not be sufficient for detection; if a simple-hypothesis solution is known and one can use the procedure of the preceding section, then they are. If not, the conditional p.d.f. of the observation

must be more carefully investigated.

6.1.3 Continuous Observations; General Solution and Sufficient Statistics. This problem was approached by sampling the observation, using the preceding results, and then letting the samples grow dense (#1.1); the form of the results is quite similar to the discrete case. Suppose that the hypothesis is specified and the corresponding family of measures on the observation space (indexed by the unknown parameter) is dominated. The R-N derivative with respect to the dominating measure, which is a function of the parameter, may be employed as the "conditional observation density" in a generalization of Bayes' rule (2.23). Under suitable conditions, this derivative is the limit of the likelihood ratio function of the samples (2.20). If the dominating measure belongs to the family being studied, the R-N derivative may be factored as usual to find a necessary and sufficient statistic (Theorem 2.3); if not, then only sufficiency can be guaranteed. If the a posteriori p.d.f.'s under either hypothesis can be found as above and the continuous simple-hypothesis result is known, then the detection statistic can also be found by a procedure analogous to the discrete result.

In any case, one must be extremely careful to investigate absolute continuity of the measures involved. The mathematical state of the art pretty well dictates that the continuous results are

practically useful only for Gaussian problems.

6.1.4 Reproducing Densities (#3.1). Existence of a sufficient statistic implies that there exists a "natural conjugate" class of p.d.f.'s on the unknown parameter which reproduces in the sense of Def. 3.1. This class is indexed by a parameter (called the conjugate parameter) whose dimension is the same as that of the sufficient statistic, and the functional form of its members can be deduced by inspection of the conditional observation density (3.8). Suppose one chooses a natural conjugate a priori p.d.f.; it is then possible to determine explicit relations to "update" the conjugate parameter based upon the sufficient statistics of the observation such that the "updated" parameter indexes the a posteriori p.d.f. Explicit consideration of Bayes' rule becomes unnecessary. Further, the dimensionality of the p.d.f.'s on the unknown parameter becomes finite since they are indexed by the conjugate parameter.

Since knowledge of the a posteriori conjugate parameter solves the estimation problem, that problem is therefore collapsed to a very tractable procedure. The detection problem may or may not behave similarly, depending on the tractability of the simple-hypothesis solution (#3.2.1).

In many problems, the results found as above are inherently sequential and easily lend themselves to recursive processing of the observation, [(3.12) - (3.18)].

Suppose the a priori p.d.f. is not itself natural conjugate but may be written as the product of a natural conjugate density and a known function of the parameter (3.19); i.e., it represents a measure absolutely continuous w.r.t. the conjugate class. One may then assume that the a priori p.d.f. is natural conjugate, proceed as above, and forestall modification of the results to account for the actual a priori density until the very end of the procedure. The necessary modification is given in (3.23) and (3.24), and may be considered to take place in a "secondary" processor (Fig. 3.4); it can be quite tractable if the integrals involved are available in closed form.

Again, all the results extend in an obvious way to solve the detection problem; the resultant receivers are derived in #3.2. The partitioning into "primary" and "secondary" processors illustrates that much of the receiver structure is independent of the a priori densities. The advantages of this approach to the problem are discussed in #3.4.1.

All these results can be applied to continuous observations if, instead of considering the conditional observation density in finding sufficient statistics and the natural conjugate class, one uses the R-N derivative of the observation measures with respect to a dominating probability measure. This derivative may be found as the limit of a likelihood-ratio function, or any such derivative

available in the literature may be employed.

6.1.5 Discrete M-SAG Noise. To illustrate application of the discrete results, the problem of detecting an exactly-known signal in (and simultaneously estimating the parameters of) discrete M-SAG noise was solved. The noise is considered to be the stationary solution to an M^{th}-order autoregressive difference equation driven by a white Gaussian random sequence [(4.2), (4.6)]. All coefficients of the equation, including the intensity of the driving sequence, are treated as unknown parameters.

The joint p.d.f. of k sequential samples (conditioned on the parameters and M "initial samples" $\underline{y}_0$) is easily written (4.17). By inspection, the sufficient statistics are found to be vectors and matrices whose elements are the sample auto-and cross-correlations up to order M, [(4.18) - (4.20)]. The natural conjugate density is seen to be a composite p.d.f. which is an M-variate Gaussian on the parameters of the autoregression and a Gamma density on the intensity of the driving sequence, (4.23). The normalizing constant and moments of this p.d.f. are extremely difficult to compute because the Gaussian portion is truncated to be zero on that region of R^M which would yield an unstable difference equation, (4.3); no attempt was made to compute moments or thoroughly analyze the conjugate class. The conjugate parameter updating relations take a very simple additive form [4.26) -(4.29)],

however, and a modal (MAP) estimate is easily written (4.30) and is shown to be consistent.

Solution of the estimation problem is similar under either hypothesis, since the uncertain parameters are the same and the signal is simply additive. The solution given is for H_0, and may be used for H_1 if the known signal is first subtracted from the observations [(4.38) - (4.43)]. Several techniques for eliminating the conditioning on $\underset{\sim}{y}_0$ were discussed in detail (#4.3.2).

Since in the discrete problem there is no advantage to explicitly retaining the simple-hypothesis solution as part of the overall detection statistic, that statistic was simplified. It turned out to consist merely of the product of ratios of a priori to a posteriori natural conjugate normalizing constants (4.44). As stated above, these were not evaluated further.

The case $M = 1$ was done in more detail (#4.4); here, it was possible to explicitly find the normalizing constant and write the detection statistic. These were, however, found to be extremely complicated expressions [(4.43), (4.55), (4.44)].

6.1.6 Continuous M-SAG Noise. This problem is essentially the same as the preceding one except that the noise is a continuous-parameter process. It may be modeled as the solution to a stochastic differential equation, in which case its parameters are the coefficients of the equation. Alternately, it may be

modeled as the output of a linear, rational transfer-function filter [(5.2) - (5.4)] which is driven by white Gaussian noise, in which case its parameters are the coefficients of the denominator polynomial of the transfer function when the leading coefficient is unity (5.5). To obtain the required continuity of measures, it is necessary that the parameter which represents the high-frequency asymptote of the filter transfer function be known (#5.3.1).

Rather than finding a limit for the likelihood ratio function, a R-N derivative found by Hajek [60] was employed, (5.27). The sufficient statistics were found to be the quadratic content of the observation and its first $M - 1$ derivatives, as well as the value of these quantities at the endpoints of the observation interval (5.34). By "conditioning" the R-N derivative on y_0 (which now consists of the observation and its first $M - 1$ derivatives at $t = 0$), the R-N derivative was put into a form suitable for sequential processing (5.33); see Section 6.1.2 above. The natural conjugate density for this form was a truncated M-variate Gaussian p.d.f. [(5.40), (5.41)]; the region on which it is truncated is again very complicated (5.5), and the normalizing constant was not found for arbitrary M. Estimation under either hypothesis is accomplished using similar techniques, exactly as discussed in the preceding section. The detection statistic was written in general form (5.46), but was not simplified because the normalizing constants were not

available. The simple-hypothesis term was explicitly retained since that result is available from classical detection theory.

The cases $M = 1$ and $M = 2$ were again done in more detail. MAP parameter estimates were given [(5.65), (5.83)] and shown to be consistent. For $M = 1$, the limit of the likelihood ratio function was evaluated (Appendix E) and the result shown to be the same as Hajek's R-N derivative. The required cancellations with the classical simple-hypothesis result were shown to occur [Appendix E, (5.69), and (5.84)] and the resulting detection statistic, though optimal, was found to be an extremely complicated function of the observation; (5.69), (5.84), and (E.26).

6.2 Contributions of this Work: Discussion

Speaking very generally, the main contributions of this work lie in the notational unification of known results of mathematical probability and statistics and in their application to the detection and estimation problem. This was accomplished both by recasting the statistical results in the language of communications theory and by generalizing the communications problem enough so that its relation to the statistical theory became clear. It is significant that the results include the case of continuous observations, and especially that this is possible without the use of the stochastic calculus which has recently become popular in communication and

control problems. All derivatives and integrals of the observation in Chapter V, for example, are ordinary derivatives or integrals of sample functions.

No novelty is claimed for the statistical results concerning sufficient statistics or natural conjugate reproducing densities in Chapters II and III; to the Bayesian statistician concerned with time series analysis or decision theory, these would appear somewhat less than startling. The only thing which is possibly unique about the material is the unified presentation which makes it clear that sufficient statistics, natural conjugate densities, and Bayes' rule all arise from and involve the same quantity. Classically, this quantity is considered to be the joint conditional p.d.f. of k observations (preferably written as a k-fold product of densities (1.7), (1.5) to make sequential processing tractable). Here, it is noted that the same results are obtained if that quantity is the likelihood ratio function, also preferably written as a product. Its use has the added advantage that the results concerning sufficient statistics are more immediately apparent, and that all results extend readily to the (infinite-dimensional) case of continuous observations. The concept of employing a Bayesian approach and using natural conjugate densities to process continuous observations on sequential finite intervals is, to the author's knowledge, original.

Further, the application of all these results to the simultaneous estimation and detection problem is believed original. Occasionally, a previous work has touched on these concepts in a specific example (#3.4.2); however, the general application of the theory and the resulting parametrizations and partitioning of the problem are unique to this work and, in some degree, to its predecessors [6] and [7].

Chapters IV and V illustrated the theory by treating the detection of a known signal in, and the simultaneous estimation of the parameters of, M^{th}-order stationary autoregressive Gaussian (Gauss-Markov) noise. Both the discrete and continuous versions of this problem were heretofore considered unsolved in communications theory. To be sure, the discrete estimation problem of Chapter IV bears a close resemblance to the well-known statistical problem of maximum-likelihood estimation of the parameters of an autoregressive stationary time series. This is natural if one recalls the close relation between maximum likelihood and Bayesian methods (see the remark following (1.34)). The explicit use of the two estimation results to solve the detection problem (#1.3.3), (#3.2) is believed original, as are most of the corresponding techniques for continuous observations as presented in (#3.3) and Chapter V.

6.3 Areas for Future Research

As is often the case, the work presented has raised at least as many interesting questions as it answered. Some of these are listed and discussed below:

a. An obvious omission has been the lack of consideration or evaluation of the performance of detectors derived here. A study of the subject might include the usual approximations to the normal receiver operating characterististic (ROC) curves and detectability index d, especially for small observation times (when the a priori parameters are very significant) and for large observations (when they may be neglected).

b. As observations are processed, the conjugate parameters trace some type of curve in the spaces $\mathcal{T}$ in which the sufficient statistics take values. As the parameters are thus learned, these trajectories presumably tend to some subspace or point which represents complete knowledge of the parameter (i.e., which corresponds to the a posteriori p.d.f. being a delta function). The concept is intuitively appealing; perhaps a metric on $\mathcal{T}$ can be defined in such a way that the "distance" from the target subspaces is easily evaluated and gives an indication of estimator and detector performance.

c. A common artifice in communications theory is to consider that a small amount of "white Gaussian noise" is added to

the "colored" noise being considered.[1] This eliminates some singularities and simplifies the solution of many problems. An interesting exercise might be to attempt an analog to Chapter V using this artifice.

d. In Chapters IV and V, the order M of the noise must be known. It is of much current interest to estimate M based on the observations. Very few sound results are available in this area.

e. Because of the difficulty of evaluating moments for or even normalizing the natural conjugate densities in Chapters IV and V, those results cannot be considered practical. It would be of interest to carry the solutions further, perhaps arriving at practically usable approximations. Also, one might be able to determine reproducing classes in the sense of #3.1.2 which are more tractable. A natural departure point would be the 1-SAG noise solution, which is at least given in closed form.

f. As samples grow dense, estimation of the parameter "a^2" is singular in Chapter V. If the discrete results of Chapter IV are examined one finds that under the same conditions the conjugate parameter ψ_c (also called ψ_4 in #4.4) tends to infinity.

---

[1] See, e.g., Van Trees [61], p. 288.

It seems clear that this has the effect of making the (discrete case) natural conjugate density singular in the subspace which corresponds to the parameter a^2 of the continuous process. The subject, especially the rate at which singular convergence occurs as samples grow dense, bears further investigation.

g. Suppose that the discrete solution of Chapter IV is given the same information as the continuous solution of Chapter V, i.e., the value of the discrete parameters which correspond to a^2 (these can be related; see (5.10) ff., and (E.5) for $M = 1$) are fixed. It would be of interest to compare the resulting discrete solution with a corresponding finite approximation to the continuous solution.

APPENDIX A

THE METZGER MODEL

Consider the detection problem

$$H_0 : y_t = n_t$$

$$H_1 : y_t = s(t) + n_t \quad , \quad t \in [0, T]$$

where n_t is zero-mean, stationary Gaussian noise with known rational spectral density $N(f^2)$; the numerator and denominator of $N(f^2)$ are polynomials in f^2 , and the order of the denominator polynomial exceeds that of the numerator by $p \geq 1$. Let the autocorrelation function of n_t be

$$R_n(\tau) = \int_{-\infty}^{\infty} N(f^2)\, e^{j2\pi f\tau}\, df \qquad (A.1)$$

Suppose $s(t)$ is an exactly-known signal which is $2p$ times continuously differentiable for $0 < t < T$, so that the first $(2p-1)$ derivatives are continuous from inside the interval at $t = 0$ and $t = T$.

The classical solution is as follows [9], [16], [6] : Let $z(y)$ denote the natural logarithm of the likelihood ratio, and d the "detectability index," a performance (probability of detection measure which, in the Gaussian case, is given by

$$d = E_1[z(y)] - E_0[z(y)] \tag{A.2}$$

where $E_i[\cdot]$ is the expectation under H_i, $i = 0, 1$. Then $z(y)$ is an optimum decision axis, and

$$z(y) = \int_0^T y_t \, s_2(t) \, dt - \frac{d}{2} \tag{A.3}$$

$$d = \int_0^T s_1^2(t) \, dt \tag{A.4}$$

The functions $s_1(\cdot)$ and $s_2(\cdot)$ are defined as follows: Let $\{\lambda_k, \phi_k(\cdot)\}$ be the eigenvalue-eigenfunction pairs of the L_2 kernel $R_n(\lambda)$; $\{\phi_k(\cdot)\}$ are a complete orthonormal set of functions in L_2. Expand $s(t)$ and y_t in terms of these functions so that, e.g.,

$$s_k = \int_0^T s(t) \, \phi_k(t) \, dt$$

Assume that

$$\sum_{k=1}^{\infty} \frac{s_k^2}{\lambda_k^2} < \infty$$

Then

$$s_1(t) = \sum_{i=1}^{\infty} \frac{s_i}{\sqrt{\lambda_i}} \, \phi_i(t)$$

$$s_2(t) = \sum_{i=1}^{\infty} \frac{s_i}{\lambda_i} \phi_i(t) \qquad \text{(A. 5)}$$

and $s_2(t)$ is the solution to a Fredholm integral equation of the first kind,

$$\int_0^T s_2(\mu)\, R_n(t - \mu)\, d\mu = s(t) \qquad \text{(A. 6)}$$

$$\text{for } 0 \leq t \leq T$$

A useful mnemonic device which simultaneously models the "generation" of the observed signal and permits evaluation of the function $s_2(\cdot)$ and the quadratic content of $s_1(\cdot)$ is the Metzger model,[1] illustrated in Fig. A. 1. The following discussion refers to that figure.

First, normalize the autocorrelation so that

$$R_n(\tau) = \frac{N_0}{2} \tilde{R}_n(\tau) \qquad \text{(A. 7)}$$

where $\tilde{R}_n(0) = 1$. Hence,

$$N(f^2) = \frac{N_0}{2} \int_{-\infty}^{\infty} \tilde{R}_n(\tau)\, e^{-j2\pi f\tau}\, d\tau \qquad \text{(A. 8)}$$

Factor $N(f^2)$ as follows:

$$N(f^2) = \frac{N_0}{2} H(j2\pi f)\, H^*(j2\pi f) \qquad \text{(A. 9)}$$

---

[1] Based on a personal correspondence from K. Metzger, [38].

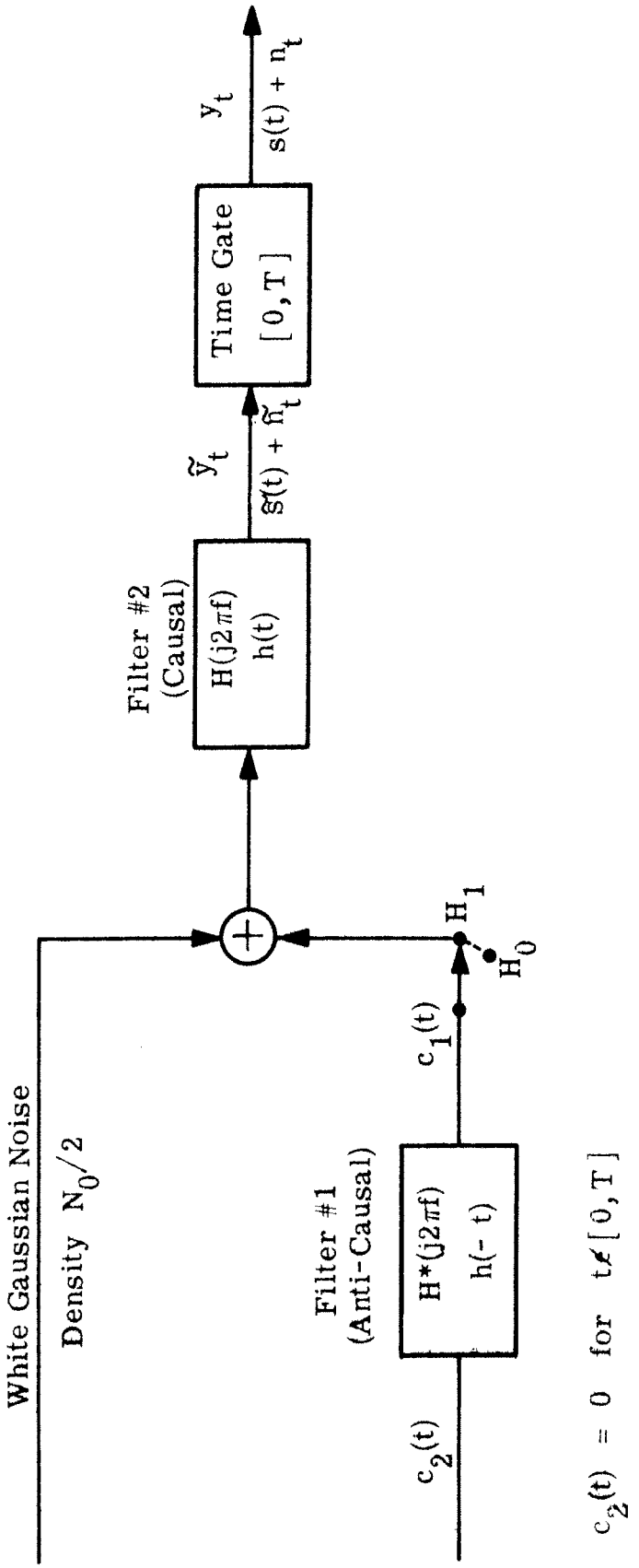

Fig. A. 1. The Metzger Model

where $H(\cdot)$ is a rational, causal transfer function and * denotes complex conjugate; hence $H^*(\cdot)$ represents an anti-causal filter. Let the inverse Fourier transforms of $H(\cdot)$ and $H^*(\cdot)$ be $h(t)$ and $h(-t)$ respectively.

Obviously, the output of Filter #2 is n_t when that filter is driven by white Gaussian noise of density $N_0/2$.

The signal $c_2(t)$ is proportional to $s_2(t)$ as follows. Let $C_2(f)$ be the Fourier transform of $c_2(t)$ and $\tilde{S}(f)$ the transform of $\tilde{s}(t)$, the signal component of $\tilde{y}_t$. From Fig. A. 1,

$$\tilde{S}(f) = H(j2\pi f)\, H^*(j2\pi f)\, C_2(f)$$

and thus, using the convolution theorem,

$$\tilde{s}(t) = \frac{2}{N_0} \int_{-\infty}^{\infty} R_n(t - \lambda)\, c_2(\lambda)\, d\lambda \quad , \quad -\infty < t < \infty$$

Since $c_2(t)$ is defined to be zero outside the interval $[0, T]$ and $s(t)$ is $\tilde{s}(t)$ truncated, this implies

$$s(t) = \frac{2}{N_0} \int_{0}^{T} R_n(t - \lambda)\, c_2(\lambda)\, d\lambda \quad , \quad 0 \leq t \leq T \tag{A. 10}$$

Comparison with (A. 6) shows that

$$s_2(t) = \frac{2}{N_0}\, c_2(t) \tag{A. 11}$$

The detectability, as defined in (A. 2) and given by (A. 4), is proportional to the quadratic content of $c_1(t)$:

$$d = \frac{2}{N_0} \int_{-\infty}^{T} c_1^2(t)\, dt \qquad \text{(A. 12)}$$

This is proved as follows. First note from (A. 1) and (A. 9) that

$$\begin{aligned} R_n(t - \mu) &= \frac{N_0}{2}\, h(t - \mu) \circledast h(\mu - t) \\ &= \frac{N_0}{2} \int_{-\infty}^{\infty} h(\mu - \lambda)\, h(t - \lambda)\, d\lambda \end{aligned} \qquad \text{(A. 13)}$$

where $\circledast$ denotes convolution. Now put (A. 4) into (A. 3) and take its expectation under both hypotheses:

$$E_0[\, z(y)\,] = -\frac{1}{2} \int_0^T s_1^2(t)\, dt$$

$$E_1[\, z(y)\,] = \int_0^T s(t)\, s_2(t)\, dt - \frac{1}{2} \int_0^T s_1^2(t)\, dt$$

Hence, from (A. 2), an alternate expression for d is

$$d = \int_0^T s(t)\, s_2(t)\, dt \qquad \text{(A. 14)}$$

$$= \int_{t=0}^{T} \int_{\mu=0}^{T} \frac{2}{N_0}\, R_n(t - \mu)\, c_2(\mu)\, s_2(t)\, d\mu\, dt$$

$$= \int_{t=0}^{T} \int_{\mu=0}^{T} \frac{2}{N_0} \int_{\lambda=-\infty}^{\infty} \frac{N_0}{2}\, h(\mu - \lambda)\, h(t - \lambda)\, c_2(\mu)\, s_2(t)\, d\lambda\, d\mu\, dt$$

where the second equality follows from (A. 6) and (A. 11), the third from (A. 13). Recombining, changing the order of integration, and using (A. 11) yields

$$d = \frac{2}{N_0} \int_{-\infty}^{\infty} \left[\int_0^T c_2(t)\, h(t - \lambda)\, dt \right] \left[\int_0^T c_2(\mu)\, h(\mu - \lambda)\, d\mu \right] d\lambda$$

Now $c_2(t) = 0$ for $t \notin [0, T]$ so the inner integrals are convolutions and, by inspection of Fig. A. 1, are precisely the output of filter #1.

$$d = \frac{2}{N_0} \int_{-\infty}^{\infty} c_1^2(\lambda)\, d\lambda$$

But $h(-t) = 0$ for $t > 0$, so $c_1(t) = 0$ for $t > T$ and thus (A. 12) is proven.

In terms of the Metzger Model signals, (A. 3) and (A. 4) can be written as

$$z(y) = \frac{2}{N_0} \int_0^T y_t\, c_2(t)\, dt - \frac{1}{N_0} \int_{-\infty}^{T} c_1^2(t)\, dt \qquad \text{(A. 15)}$$

The Metzger Model provides no new theoretical results; rather, it permits solution of the integral equation (A. 6) using techniques familiar to the engineer and thus provides much more

insight into the problem than would a straighforward solution of that equation. As a side benefit it provides an alternate method (Eq. (A. 12)) of evaluating the detectability index.

To illustrate its application, the model will be used to solve the problem of detecting a known signal in M^{th}-Order stationary autoregressive Gaussian (M-SAG) noise (see Section 5. 1), i. e., in zero-mean Gaussian noise with spectral density

$$N(f^2) = \frac{a^2}{\prod_{i=1}^{M} [(2\pi f)^2 + p_i^2]} \quad , \quad \begin{matrix} \mathrm{Re}(p_i) > 0 \\ p_i \neq p_j \end{matrix} \qquad \text{(A. 16)}$$

The solution will be derived in detail for $M = 2$, and will be stated for $M = 1$ and for arbitrary M .

Consider Fig. A. 1; recall that $s(t)$ is assumed 2M-times differentiable in $(0, T)$ so that $s^{(2M-1)}(0)$ and $s^{(2M-1)}(T)$ exist as limits from inside the interval. Thus, $\tilde{s}(t)$ and its first (2M- 1) derivatives must be continuous from inside the interval at the endpoints $t = 0$ and $t = T$. Now $c_2(t)$ is zero outside of $[0, T]$; it is easily verified that the most general form for $c_2(t)$ is [1], [2]

---

[1] This is, of course, the same as the classical result of Zadeh and Ragazzini [64]. In fact, the Metzger Model is merely a system which models the differential equation and boundary conditions associated with (A. 6).

[2] If the numerator polynomial is nontrivial, $c_2(t)$ must also contain exponential terms in t which are related to the zeros

$$c_2(t) = g_2(t)\ [u(t) - u(t - T)] + \sum_{i=0}^{M-1} (-1)^i \alpha_i \delta^{(i)}(t) + \sum_{j=1}^{M-1} (-1)^j \beta_j \delta^{(j)}(t) \qquad \text{(A. 17)}$$

where $g_2(t)$ is piecewise continuous, $u(t)$ is the unit step function, and $\delta^{(i)}(t)$ is the i^{th} derivative of the Dirac delta function. Recall

$$\int f(t)\, \delta^{(i)}(t)\, dt = (-1)^i f^{(i)}(0) \qquad \text{(A. 18)}$$

This in turn implies that $c_1(t)$ contains no singularities but may be discontinuous at $t = 0$ and $t = T$, and that $\tilde{s}(t)$ and its first $(2M - 1)$ derivatives are not only continuous from inside $[0, T]$ but are continuous.

To use the Metzger Model, one assumes a $c_2(t)$ as in (A. 17) and solves the model in the "forward" direction for $\tilde{s}(t)$. This is somewhat tedious but the procedure is straightforward and will be illustrated in the example which follows. Since the first $(2M - 1)$ derivatives of $\tilde{s}(t)$ are continuous at $t = 0$ and $t = T$, and since they are known from inside the interval, "matching up" the boundaries yields a set of equations which may be

---

of the numerator; these represent the homogeneous solution to the associated differential equation. See Helstrom, [26] p. 441 and [25].

solved for the constants α_i and β_i ; the function $g_2(t)$ is just the "infinite interval" solution and may be found using, for example, standard transform techniques. The procedure will now be illustrated for $M = 2$:

Example A. 1

$$N(f^2) = \frac{a^2}{[(2\pi f)^2 + p_1^2][(2\pi f)^2 + p_2^2]} \qquad \begin{matrix} \mathrm{Re}(p_1, p_2) > 0 \\ p_1 \neq p_2 \end{matrix} \tag{A. 19}$$

$$= \frac{a^2}{|(j2\pi f)^2 + q_1(j2\pi f) + q_2|^2} \tag{A. 20}$$

where

$$\begin{aligned} q_1 &= p_2 + p_1 \\ q_2 &= p_2 p_1 \end{aligned} \tag{A. 21}$$

The alternate parameters introduced in (A. 20) will be convenient. From (A. 19) the autocorrelation function is

$$R_n(\tau) = \frac{a^2}{2p_1p_2(p_2^2 - p_1^2)} \left[p_2 e^{-p_1|\tau|} - p_1 e^{-p_2|\tau|} \right] \tag{A. 22}$$

Normalize this as in (A. 7); define the constant

$$K = \sqrt{2q_1q_2} = \sqrt{2p_1p_2(p_1+p_2)} \tag{A. 23}$$

Then one obtains the following quantities for the Metzger Model:

$$\frac{N_0}{2} = \frac{a^2}{K^2} \tag{A. 24}$$

$$\tilde{R}_n(\tau) = \frac{1}{p_2-p_1}\left[p_2e^{-p_1|\tau|} - p_1e^{-p_2|\tau|}\right] \tag{A. 25}$$

$$H(j2\pi f) = \frac{K}{(j2\pi f+p_1)(j2\pi f+p_2)} \tag{A. 26}$$

$$= \frac{K}{(j2\pi f)^2+q_1(j2\pi f)+q_2} \tag{A. 27}$$

and the impulse response is

$$h(t) = K\frac{e^{-p_1t} - e^{-p_2t}}{p_2-p_1}\,u(t) \tag{A. 28}$$

$H^*(j2\pi f)$ is found by inspection of (A. 25) or (A. 27); its impulse response is

$$h_*(t) = h(-t) \tag{A. 29}$$

Since the truncation of the "Time Gate" does not affect the linearity of the system for $0 < t < T$, it is easy to verify by use of Fourier transforms that, inside the interval, the signals of the Metzger Model are related by

$$c_1(t) = K^{-1} \left[\tilde{s}''(t) + q_1 \tilde{s}'(t) + q_2 \tilde{s}(t) \right] \qquad \text{(A. 30)}$$

$$c_2(t) = K^{-1} \left[c_1''(t) - q_1 c_1'(t) + q_2 c_1(t) \right] \qquad \text{(A. 31)}$$

$$= K^{-2} \left[\tilde{s}^{(4)}(t) + (2q_2 - q_1^2) \tilde{s}''(t) + q_2^2 \tilde{s}(t) \right] \qquad \text{(A. 32)}$$

and clearly $\tilde{s}(t) = s(t)$, $0 < t < T$. For clarity, the solution procedure will be divided into steps.

Step 1:

Assume that $c_2(t)$ is known:

$$c_2(t) = K^{-2} \{ g_2(t)[u(t) - u(t-T)] + \alpha_0 \delta(t) - \alpha_1 \delta'(t) + \beta_0 \delta(t - T) - \beta_1 \delta'(t - T) \} \qquad \text{(A. 33)}$$

Find $c_1(t)$ by convolution,

$$c_1(t) = c_2(t) \circledast h_*(t)$$

After some manipulation,

$$c_1(t) = K^{-1} \left\{ \left[\int_{\max(0,t)}^{T} g_2(\tau) \frac{e^{p_1(t-\tau)} - e^{p_2(t-\tau)}}{p_2 - p_1} d\tau \right] u(T - t) \right.$$

$$+ \left[\alpha_0 \frac{e^{p_1 t} - e^{p_2 t}}{p_2 - p_1} + \alpha_1 \frac{p_2 e^{p_2 t} - p_1 e^{p_1 t}}{p_2 - p_1} \right] u(-t)$$

$$\left. + \left[\beta_0 \frac{e^{p_1(t-T)} - e^{p_2(t-T)}}{p_2 - p_1} + \beta_1 \frac{p_2 e^{p_2(t-T)} - p_1 e^{p_1(t-T)}}{p_2 - p_1} \right] u(T - t) \right\} \tag{A.34}$$

It is easily verified that the discontinuities in (A.34) are

$$\begin{aligned} c_1(0^+) - c_1(0^-) &= -K^{-1} \alpha_1 \\ c_1(T^+) - c_1(T^-) &= -K^{-1} \beta_1 \end{aligned} \tag{A.35}$$

and that

$$\begin{aligned} c_1'(0^+) - c_1'(0^-) &= K^{-1} [\alpha_0 - \alpha_1(p_2+p_1)] \\ c_1'(T^+) - c_1'(T^-) &= K^{-1} [\beta_0 - \beta_1(p_2+p_1)] \end{aligned} \tag{A.36}$$

Step 2:

The signal $c_1(t)$ as found in (A.34) has the form

$$c_1(t) = \begin{cases} K^{-1} \dfrac{\gamma_1 e^{p_1 t} - \gamma_2 e^{p_2 t}}{p_2 - p_1} & , \quad t < 0 \\ K^{-1} \, g_1(t) & , \quad 0 < t < T \\ 0 & , \quad t > T \end{cases} \tag{A.37}$$

Again using convolution

$$\tilde{s}(t) = c_1(t) \circledast h(t) \tag{A.38}$$

one can find $\tilde{s}(t)$. For $t < 0$, the result is

$$\tilde{s}(t) = [\, 2p_1p_2(p_2^2 - p_1^2) \,]^{-1} (p_2\gamma_1 e^{p_1 t} - p_1\gamma_2 e^{p_2 t}) \tag{A.39}$$

So that

$$\tilde{s}(0^-) = \frac{p_2\gamma_1 - p_1\gamma_2}{2p_1p_2(p_2^2 - p_1^2)} \tag{A.40}$$

$$\tilde{s}\,'(0^-) = \frac{\gamma_1 - \gamma_2}{2(p_2^2 - p_1^2)} \tag{A.41}$$

One can evaluate $\tilde{s}(t)$ for $t > 0$ and verify that $\tilde{s}$ and $\tilde{s}'$ are continuous, but that serves no purpose here.

Step 3:

Assume now that $s(t)$ is given; using the derived relations, one can evaluate the constants α_i and β_i of (A. 33) in terms of the known $s(t)$: Recall first that $\tilde{s}(0^-) = s(0)$ and $\tilde{s}'(0^-) = s'(0)$. Thus, from (A. 40) and (A. 41) ,

$$\begin{bmatrix} \gamma_1 \\ \\ \gamma_2 \end{bmatrix} = 2p_1p_2(p_1 + p_2) \begin{bmatrix} s(0) - \dfrac{s'(0)}{p_2} \\ \\ s(0) - \dfrac{s'(0)}{p_1} \end{bmatrix} \qquad \text{(A. 42)}$$

Finally, using this intermediate result and the differential equations (A. 30) and (A. 31) in (A. 35) and (A. 36), one obtains

$$\alpha_1 = - s''(0) + q_1 s'(0) - q_2 s(0) \qquad \text{(A. 43)}$$

$$\alpha_0 = s^{(3)}(0) + (q_2 - q_1^2) s'(0) + q_1q_2 s(0) \qquad \text{(A. 44)}$$

$$\beta_1 = s''(T) + q_1 s'(T) + q_2 s(T) \qquad \text{(A. 45)}$$

$$\beta_0 = s^{(3)}(T) - (q_2 - q_1^2) s'(T) + q_1q_2 s(T) \qquad \text{(A. 46)}$$

Using (A. 31) one thus finds that

$$c_2(t) = K^{-2} \left[s^{(4)}(t) - (q_1^2 - 2q_2)\, s''(t) + q_2^2\, s(t) \right] \left[u(t) - u(t-T) \right]$$

$$+ \alpha_0\, \delta(t) - \alpha_1\, \delta'(t) + \beta_0\, \delta(t-T) - \beta_1\, \delta'(t-T) \qquad \text{(A. 47)}$$

Step 5:

The detectability index d may be found using either (A. 12),

$$\frac{2}{N_0} d = \int_{-\infty}^{T} c_1^2(t)\, dt \qquad \text{(A. 48)}$$

or (A. 14),

$$\frac{2}{N_0} d = \int_{0}^{T} s(t)\, c_2(t)\, dt \qquad \text{(A. 49)}$$

The equality of these integrals provides an excellent check on the solutions obtained for the model. For this example, either method will (after some integration by parts and use of (A. 24)) yield

$$d = a^{-2} \left\{ \int_0^T \left[s''(t) \right]^2 dt + (q_1^2 - 2q_2) \int_0^T \left[s'(t) \right]^2 dt + q_2^2 \int_0^T s^2(t)\, dt \right.$$

$$+ q_1 q_2 \left[s^2(T) + s^2(0) \right] + q_1 \left\{ \left[s'(T) \right]^2 + \left[s'(0) \right]^2 \right\}$$

$$\left. + 2q_2 \left[s(T)\, s'(T) - s(0)\, s'(0) \right] \right\} \qquad \text{(A. 50)}$$

Step 6:

Utilizing (A. 15), all these results may be pieced back together to write the solution to the detection problem:

$$z(y) = a^{-2} \left\{ \int_0^T y_t [\, s^{(4)}(t) - (q_1^2 - 2q_2)\, s''(t) + q_2^2\, s(t)\,]\, dt \right.$$
$$+ y_0 [\, s^{(3)}(0) + (q_2 - q_1^2)\, s'(0) + q_1 q_2\, s(0)\,]$$
$$- y_T [\, s^{(3)}(T) + (q_2 - q_1^2)\, s'(T) - q_1 q_2\, s(T)\,]$$
$$- y_0' [\, s''(0) - q_1\, s'(0) + q_2\, s(0)\,]$$
$$\left. + y_T' [\, s''(T) + q_1\, s'(T) + q_2\, s(T)\,] \right\} - \frac{d}{2} \qquad \text{(A. 51)}$$

where d is given by (A. 50).

Example A. 2

Using the same technique as above one finds that if n_t is 1-SAG noise (also known as the Ornstein-Uhlenbeck process), with spectral density function

$$N(f^2) = \frac{a^2}{(2\pi f)^2 + p_1^2} \qquad \text{(A. 52)}$$

and autocorrelation

$$R_n(\tau) = \frac{a^2}{2p_1} \exp(-p_1|\tau|) \tag{A.53}$$

Then

$$s_2(t) = \frac{2}{N_0} c_2(t)$$

$$= a^{-2} \left\{ -s''(t) + p_1^2 s(t) + [p_1 s(0) - s'(0)]\, \delta(t) + [p_1 s(T) + s'(T)]\, \delta(t-T) \right\} \tag{A.54}$$

and

$$d = \frac{2}{N_0} \int_{-\infty}^{T} c_1^2(t)\, dt$$

$$d = a^{-2} \left\{ \int_0^T [(s'(t))^2 + p_1^2 s^2(t)]\, dt + p_1 [s^2(0) + s^2(T)] \right\} \tag{A.55}$$

As usual, the solution to the simple-hypothesis detection problem can then be written using (A.3).

Example A.3

For M-SAG noise, the spectral density function is as in (A.16); this can alternately be written

$$N(f^2) = \frac{a^2}{Q(j2\pi f)Q(-j2\pi f)} \tag{A.57}$$

where $Q(\cdot)$ is an M^{th}-order polynomial

$$Q(z) = z^M + q_1 z^{M-1} + \ldots + q_M \tag{A.58}$$

with no zeros in the right-hand z-plane. The coefficient q_i is the sum of all different combinations of the p_k taken i at a time. For this general case, it can be shown that[1]

$$\begin{aligned} s_2(t) = a^{-2} \Bigg\{ & \left[Q\left(\frac{d}{dt}\right) Q\left(-\frac{d}{dt}\right) \right] \{s(t)\} \\ & + \sum_{\ell=0}^{M-1} \delta^{(\ell)}(t) \sum_{m=0}^{2M-1-\ell} s^{(m)}(0) \left[\sum_k q_{M-k} q_{M-m-\ell-1+k} (-1)^k \right] \\ & + \sum_{\ell=0}^{M-1} \delta^{(\ell)}(t-T) \sum_{m=0}^{2M-1-\ell} s^{(m)}(T) \left[\sum_k q_{M-k} q_{M-m-\ell-1+k} (-1)^{m+\ell-k} \right] \Bigg\} \end{aligned} \tag{A.59}$$

---

[1] Pisarenko [44], p. 59.

where $q_0 = 1$ and where the limits on the sums over k are $\max(0,\ m + \ell + 1 - M) \leq k \leq \min(m,\ M)$. Using (A. 14) and integrating by parts, the detectability index is found to be

$$d = a^{-2} \left\{ \sum_{k=0}^{M} (-1)^k Q_{2k} \int_0^T [s^{(k)}(t)]^2 \, dt + E_1 + E_2 \right\}$$

where E_1 arises from the delta-functions in (A. 59) and is given by

$$E_1 = \sum_{i,j=0}^{M-1} \left[(-1)^i s^{(i)}(T)\, s^{(j)}(T) + (-1)^j s^{(i)}(0)\, s^{(j)}(0) \right] \cdot \sum (-1)^k \quad q_{M-k} q_{M-i-j-1+k}$$

and the limits on k are: $\max(0,\ i + j + 1 - M) \leq k \leq i$;

the term E_2 arises from the integration by parts and is

$$E_2 = \sum_{k=1}^{M} Q_{2k} \sum_{i=0}^{k-1} (-1)^i \left[s^{(i)}(T)\, s^{(2k-i)}(T) - s^{(i)}(0)\, s^{(2k-i)}(0) \right]$$

The coefficient Q_{2k} is the coefficient of z^{2k} in the expansion of $Q(z)Q(-z)$, where $Q(z)$ is given by (A. 58).

APPENDIX B

MEASURE AND PROBABILITY THEORY; SUFFICIENT STATISTICS ON MEASURE SPACES

The purpose of this appendix is two-fold: to establish the measure-theoretic background necessary for portions of the main body of this dissertation, and to present an exposition of some modern results concerning necessary and sufficient statistics. The treatment of measure and probability theory is necessarily terse and incomplete. The aim is mostly to establish a notation; the scheme follows three primary references: Chapter 1 of Wong [62], Articles 2.1 to 2.4 of Lehmann [35], and the author's notes from a course taught by Professor W. L. Root in 1972.

The treatment of necessary and sufficient statistics follows an elegant formulation of that subject in terms of sub-σ-algebras as first presented by Bahadur [3]; basic to that work is a classic paper due to Halmos and Savage [24]. An attempt is made throughout to relate the results to the more traditional concepts of statistics and conditional probability.

1. Basic Measure Theory. Consider an abstract space $\mathcal{Y}$ whose points are denoted y; often $\mathcal{Y}$ will represent the totality of all outcomes y of a random experiment and will be called the sample space. Aggregates of points (outcomes) are subsets of $\mathcal{Y}$ and

denoted, e.g., $A \subset \mathcal{Y}$. If $\mathcal{A}$ is a class of such sets which is closed under complementation and countable union, it is called a σ-algebra. A is said to be an $\mathcal{A}$-measurable set if $A \in \mathcal{A}$. An arbitrary class of sets $\mathcal{C}$ is $\mathcal{A}$-measurable if each $C \in \mathcal{C}$ is; given such a class, there exists a σ-algebra generated by $\mathcal{C}$, $\mathcal{A}(\mathcal{C})$, which is the minimal σ-algebra with respect to which $\mathcal{C}$ is measurable. A σ-algebra $\mathcal{A}_0$ is a sub-algebra of $\mathcal{A}$, $\mathcal{A}_0 \subset \mathcal{A}$, if $A \in \mathcal{A}_0 \Longrightarrow A \in \mathcal{A}$.

The couple $(\mathcal{Y}, \mathcal{A})$ is called a measurable space or a pre-probability space. Let μ be a nonnegative set function on $\mathcal{A}$ which is a σ-additive: If $\{A_n\} \in \mathcal{A}$ are pairwise disjoint, then

$$\mu\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mu(A_n) \qquad \text{(B.1)}$$

μ is called a measure on $\mathcal{A}$. If there exists a sequence of sets satisfying $\bigcup_n A_n = \mathcal{Y}$ and if each $\mu(A_n) < \infty$, then μ is a σ-finite measure. If $\mu(\mathcal{Y}) < \infty$, then μ is a finite measure. If $\mu(\mathcal{Y}) = 1$, then μ is a probability measure and is denoted $\mathcal{P}$.

The triple $(\mathcal{Y}, \mathcal{A}, \mu)$ is called a measure space. Sets of μ-measure zero and their subsets are called null sets, and $(\mathcal{Y}, \mathcal{A}, \mu)$ is said to be complete if all null sets are $\mathcal{A}$-measurable. Every measure space can be uniquely completed. Any relation which holds for all $y \in \mathcal{Y}$ except on a null set is said to hold almost surely

(a.s[μ]).[1]

If $\mathcal{Y}$ is R^n, then the smallest σ-algebra which contains all rectangles, i.e., all n-fold products of half-open intervals

$$B = \{\underline{x} \in R^n : a_i < x_i \leq b_i,\ i = 1 \ldots n\} \qquad (B.2)$$

is called the Borel σ-algebra $\mathcal{R}^n$, and its members are the Borel sets. Measures on $(R^n, \mathcal{R}^n)$ are called Borel measures. The unique Borel measure which assigns to each rectangle the product of the lengths of the intervals which comprise it is called the Lebesgue measure on $(R^n, \mathcal{R}^n)$ and is denoted $\mathcal{L}$. The completion of $(R^n, \mathcal{R}^n, \mathcal{L})$ defines a σ-algebra which contains the Lebesgue-measurable sets, but which will not be distinguished notationally.

Let $T : \mathcal{Y} \to \mathcal{T}$ be a mapping, and $\mathcal{B}$ a σ-algebra of sets in $\mathcal{T}$. T is a measurable mapping (meas.[$\mathcal{A}$]) if $T^{-1}(B) \in \mathcal{A}$ for all $B \in \mathcal{B}$. Writing $T : (\mathcal{Y}, \mathcal{A}) \to (\mathcal{T}, \mathcal{B})$ will imply T meas.[$\mathcal{A}$]. If a σ-algebra $\mathcal{B}$ is not determined by other considerations, then T can be considered to generate on its range the σ-algebra $\mathcal{B}'$ consisting of all sets B such that

$$T^{-1}(B) = \{y \in \mathcal{Y} : Ty \in B\} \in \mathcal{A} \qquad (B.3)$$

---

[1] Throughout, square brackets should be read as "modulo" or "with respect to" when they occur in the text.

Given a measure μ on $(\mathcal{Y}, \mathcal{A})$, T induces a measure μ^T on $(\mathcal{T}, \mathcal{B})$ which satisfies

$$\mu^T(B) = \mu[T^{-1}(B)] \quad \forall B \in \mathcal{B} \tag{B.4}$$

and often one writes $\mu^T = \mu T^{-1}$.

2. Basic Probability Theory. A measure space $(\mathcal{Y}, \mathcal{A}, \mathcal{P})$ is called a probability space and often a sample space; a probability measure satisfies all the basic axioms normally attributed to the concept of probability.

In applications, probabilities over the sample space $(\mathcal{Y}, \mathcal{A})$ refer to random experiments whose outcomes are the points $y \in \mathcal{Y}$. Denote these observations as Y, and let the probability that Y falls in $A \subset \mathcal{Y}$ be $\mathcal{P}^Y(A) = \mathcal{P}^Y\{y : y \in A \in \mathcal{A}\}$. Considered as a variable whose value is determined by the observation, Y is called a random variable[1] over $\mathcal{Y}$ and the measure $\mathcal{P}^Y$ is called the probability distribution of Y. If $T : (\mathcal{Y}, \mathcal{A}) \to (\mathcal{T}, \mathcal{B})$, then $\mathcal{P}^Y$ induces a probability measure $\mathcal{P}^T$ on the sets of $\mathcal{B}$ as in (B.4). The values taken on by T(Y) can be considered outcomes of a related random experiment, and so

---

[1] This differs from the definition usually given in treatments on probability theory, but does not conflict with them. The approach here follows Lehmann [35], and is consistent with the classical concept of a random variable. Mathematically, a random variable is nothing more than the carrier of its distribution.

$T = T(Y)$ is a random variable with probability distribution $\mathscr{P}^T$. A measurable mapping T defined on the sample space is called a statistic; if $\mathscr{T}$ is a Euclidean space, then $\mathscr{B}$ will always be considered the Borel sets. It is clear that any statistic is also a random variable.

Let $(R^n, \mathscr{R}^n, \mathscr{P})$ be a Borel probability space; then the point function $P : R^n \to R^1$,

$$P(a_1 \ldots a_n) = \mathscr{P}\{A : \underline{x} \epsilon A \Longrightarrow -\infty < x_i < a_i \; ; \; i = 1 \ldots n\} \tag{B.5}$$

is called the cumulative distribution function (c.d.f.) of $\mathscr{P}$ and has all the usual properties of such a function;[1] conversely, every c.d.f. uniquely determines a Borel probability measure.

3. Integration and Expectation. Let $f: (\mathscr{Y}, \mathscr{A}) \to (R^1, \mathscr{R}^1)$; f is the indicator function of a set A, $f(y) = I_A(y)$, if $f(y) = 1$ for $y \epsilon A$ and zero elsewhere. Clearly, $A \epsilon \mathscr{A} \longleftrightarrow I_A(y)$ meas. $[\mathscr{A}]$. f is a simple function if, for some sets $A_i \epsilon \mathscr{A}$ and some constants $a_i < \infty$,

$$f = \sum_{i=1}^{N} a_i I_{A_i} \tag{B.6}$$

---

[1] Wong [62], pp. 6-7.

Let μ be a measure on $(\mathcal{Y}, \mathcal{A})$; if f is simple, define

$$\int_{y} f(y)\, d\mu(y) \stackrel{d}{=} \sum_{i=1}^{N} a_i \mu(A_i) \qquad (B.7)$$

Any measurable f can be approximated by a sequence of simple functions[1] and the definition of (B.7) can be unambiguously extended. A function f is said to be integrable with respect to μ, (f is integ.$[\mathcal{A}, \mu]$), if f is $\mathcal{A}$-measurable and $\int_y |f| d\mu < \infty$.

If $(\mathcal{Y}, \mathcal{A}, \mathcal{P})$ is a sample space and T a real-valued statistic, the expectation of T is defined as

$$\begin{aligned} E(T) &= \int_{\mathcal{Y}} T(y)\, d\mathcal{P}(y) \\ &= \int_{\mathcal{T}} t\, d\mathcal{P}^T(t) \end{aligned} \qquad (B.8)$$

where $\mathcal{P}^T = \mathcal{P}_{T^{-1}}$. The equivalence of these expressions will be established by Lemma B-2 below.

Let μ and ν be two measures on $(\mathcal{Y}, \mathcal{A})$. ν is said to be absolutely continuous with respect to μ ($\nu << \mu$) if $\mu(A) = 0 \Rightarrow \nu(A) = 0$. μ and ν are equivalent ($\nu \equiv \mu$) if they are mutually absolutely continuous. ν and μ are (mutually) singular, $\nu \perp \mu$, if there exists a set A such that $\mu(A) = 0$ and $\nu(A^c) = 0$,

---

(1) Ibid. pp. 15-17.

where "c" denotes complement. A basic theorem is:

THEOREM B.1 (Radon-Nikodym)[1]

Let μ,ν be σ-finite measures on $(\mathcal{Y},\mathcal{A})$.

Then $\nu << \mu$ iff there exists $f : (\mathcal{Y},\mathcal{A}) \to (R^1,\mathcal{R}^1)$, $f \geq 0$, such that

$$\nu(A) = \int_A f\,d\mu \qquad \forall A \in \mathcal{A} \tag{B.9}$$

f is unique (a.s.$[\mu]$). ∎

f is called the Radon-Nikodym derivative of ν with respect μ, $f = \frac{d\nu}{d\mu}$, and often one writes $d\nu = f\,d\mu$. If ν is a probability measure, f is also called the probability density function of ν with respect to μ(p.d.f.$[\mu]$). If $(\mathcal{Y},\mathcal{A},\mu) = (R^n,\mathcal{R}^n,\mathcal{L})$, then f is just called the probability density function (p.d.f.) of ν.

4. Families of Measures. Given $(\mathcal{Y},\mathcal{A})$, let $\mathcal{M} = \{\mu_\theta : \theta \in \Theta\}$ be a family of measures on $\mathcal{A}$. A set is $\mathcal{M}$-null if it is a null set $[\mu_\theta]\ \forall\theta$. A function is $\mathcal{M}$-integrable, integ.$[\mathcal{A},\mathcal{M}]$, if it is integ.$[\mathcal{A},\mu_\theta]\forall\theta$. $\mathcal{M}$ is said to be a dominated family of measures ($\mathcal{M} << \lambda$) if there exists a σ-finite measure λ on $(\mathcal{Y},\mathcal{A})$, not necessarily a member of $\mathcal{M}$, such that $\mu_\theta << \lambda$

[1] See, e.g., Royden [50], p. 238.

for all $\theta \in \Theta$. Domination by a σ-finite measure is equivalent to domination by a finite or a probability measure.[1] $\mathcal{M}$ is an equivalent family if the μ_θ are pairwise equivalent.

5. Sub-Algebras, Statistics, and Conditional Expectation. Let $\mathcal{A}_0 \subset \mathcal{A}$ be a sub-algebra. A measure μ on $\mathcal{A}$ is also a measure on $\mathcal{A}_0$. Let μ, ν be measures on $\mathcal{A}$; then $\nu << \mu$ on $\mathcal{A} \Rightarrow \nu << \mu$ on $\mathcal{A}_0$, and thus $\mu \equiv \nu$ on $\mathcal{A} \Rightarrow \mu \equiv \nu$ on $\mathcal{A}_0$.

Let $(\mathcal{Y}, \mathcal{A}, \mathcal{P})$ be the sample space and $T : (\mathcal{Y}, \mathcal{A}) \to (\mathcal{T}, \mathcal{B})$ a statistic. T induces on $\mathcal{Y}$ a sub-algebra

$$\mathcal{A}_0 = T^{-1}(\mathcal{B}) = \{A_0 \in \mathcal{A} : A_0 = T^{-1}(B),\ B \in \mathcal{B}\} \tag{B.10}$$

If the events $\{A^i\}$ are all in the same set $A_0 \in \mathcal{A}_0$, they cannot be distinguished by observation of T(Y) . The correspondence of sets (B. 10) establishes a similar correspondence for measurable functions:

Lemma B.1[2]

Let T be as above. Then $f : (\mathcal{Y}, \mathcal{A}) \to (R^1, \mathcal{R}^1)$ is

---

[1] Halmos and Savage [24], pp. 232; see also Theorem B. 2 below.

[2] Ibid., p. 223.

meas.$[\mathscr{A}_0]$ iff there exists $g : (\mathscr{T}, \mathscr{B}) \to (R^1, \mathscr{R}^1)$ such that

$$f(y) = g[T(y)] \quad \forall y \in \mathscr{Y} \blacksquare \tag{B.11}$$

Integrals of these functions can also be related:

Lemma B.2[1]

Let μ be a σ-finite measure on $(\mathscr{Y}, \mathscr{A})$, μ^T be the measure on $(\mathscr{T}, \mathscr{B})$ induced by T, and the functions f, g be as above. Then for any $B \in \mathscr{B}$

$$\int_{T^{-1}(B)} g[T(y)]\, d\mu(y) = \int_B g(t)\, d\mu^T(t) \blacksquare \tag{B.12}$$

Suppose $\mathscr{A}_0$ is an arbitrary sub-algebra of $\mathscr{A}$, and $f : (\mathscr{Y}, \mathscr{A}) \to (R^1, \mathscr{R}^1)$ is nonnegative and integ.$[\mathscr{A}, \mathscr{P}]$. The conditional expectation (c. exp.) of f given $\mathscr{A}_0$, $E^{\mathscr{A}_0} f$, is the unique[2] nonnegative, real-valued function which satisfies:

---

[1] Ibid., pp. 228-229.

[2] Existence and uniqueness follow from the Radon-Nikodym Theorem B.1.

a. $E^{\mathscr{A}_0} f$ is meas$[\mathscr{A}_0]$

b. $\int_{A_0} [E^{\mathscr{A}_0} f](y)\, d\mathscr{P}(y) = \int_{A_0} f(y)\, d\mathscr{P}(y)$

for all $A_0 \in \mathscr{A}_0$

(B.13)

$E^{\mathscr{A}_0} f$ yields the same average as f on the "coarser" sets of $\mathscr{A}_0$, but is meas.$[\mathscr{A}_0]$ which f is not. Some properties of c.exp., most of which follow directly from (B.13) are:[1]

Lemma B.3

Suppose $\mathscr{A}_0 \subset \mathscr{A}$; f_i, i = 1, 2 ... are integ.$[\mathscr{A}, \mathscr{P}]$; and h is integ.$[\mathscr{A}_0, \mathscr{P}]$. Then, a.s. $[\mathscr{A}, \mathscr{P}]$,

a. $\int_{\mathscr{Y}} E^{\mathscr{A}_0} f(x)\, d\mathscr{P}(x) = \int_{\mathscr{Y}} f(x)\, d\mathscr{P}(x) = E f$

b. $f_1(x) \leq f_2(x) \Longrightarrow E^{\mathscr{A}_0} f_1 \leq E^{\mathscr{A}_0} f_2$

and the same holds for equalities.

c. $E^{\mathscr{A}_0}(\alpha f_1 + \beta f_2) = \alpha E^{\mathscr{A}_0} f_1 + \beta E^{\mathscr{A}_0} f_2$

---

[1] Wong [62], pp. 29-32.

d. $E^{\mathcal{A}_0}[\,h(x)\,f(x)] = h(x)\,E^{\mathcal{A}_0}f(x)$

provided the terms make sense. In particular, $E^{\mathcal{A}_0}h(x) = h(x)$.

e. If $\mathcal{A}_1 \subset \mathcal{A}_0$ is a sub-algebra, then $E^{\mathcal{A}_1}f = E^{\mathcal{A}_1}\{E^{\mathcal{A}_0}f\}$

To relate this to the more familiar notion of c.exp. "given a value of a statistic," suppose that $\mathcal{A}_0$ was generated by $T : (\mathcal{Y}, \mathcal{A}) \to (\mathcal{T}, \mathcal{B})$. From Lemma B.1, $[E^{\mathcal{A}_0}f](y)$ depends on y only through $T(y)$, since there must exist a function $g : (\mathcal{T}, \mathcal{B}) \to (R^1, \mathcal{R}^1)$ such that

$$[E^{\mathcal{A}_0}f](y) = g[T(y)] \tag{B.14}$$

Traditionally, this is written as

$$E[f(Y) \mid T = t] = g(t) \tag{B.15}$$

It appears that the formulation using sub-algebras is more basic and elegant. Suppose that $S : (\mathcal{Y}, \mathcal{A}) \to (\mathcal{S}, \mathcal{C})$ also generates $\mathcal{A}_0$; by Lemma B.1 one can write the c.exp.

$$[E^{\mathcal{A}_0}f](y) = g[T(y)] = h[S(y)]$$

In fact, any statistic which generates $\mathcal{A}_0$ gives a similar determination of $E^{\mathcal{A}_0}f$ which depends only on the statistic. Many

notions traditionally expressed in terms of specific statistics are much simpler and more elegantly expressed in terms of sub-algebras; this is true in particular of the notion of "necessary and sufficient statistics": One speaks of necessary and sufficient sub-algebras, and then any statistics generating these sub-algebras are necessary and sufficient statistics.[1]

The conditional probability of a set is the conditional expectation of its indicator function

$$P\{A \mid \mathcal{A}_0\} = E^{\mathcal{A}_0} I_A , \quad A \in \mathcal{A} \tag{B.16}$$

or, in terms of a specific statistic,

$$P\{A \mid t\} = E[I_A(Y) \mid t] \tag{B.17}$$

These relations are defined and must be considered as functions of t for fixed A. Under suitable restrictions however (namely that $\mathcal{Y}$ be a Euclidean space), there exist determinations of $P\{A \mid t\}$ which make it a probability measure on $(\mathcal{Y}, \mathcal{A})$ for each fixed $t \in \mathcal{T}$.[2]

---

[1] This approach is due, among others, to Bahadur [3].

[2] Lehmann [35], Art. 2.5.

6. Necessary and Sufficient Statistics. Let $\mathcal{M} = \{ \mathcal{P}_\theta , \theta \in \Theta \}$ be a family of probability measures on $(\mathcal{Y}, \mathcal{A})$, and $T : (\mathcal{Y}, \mathcal{A}) \rightarrow (\mathcal{T}, \mathcal{B})$. Let the induced measures on $(\mathcal{T}, \mathcal{B})$ be denoted $\mathcal{N} = \{ \mathcal{P}_\theta T^{-1} , \theta \in \Theta \}$

If $\mathcal{Y}$ is a finite-dimensional Euclidean space, then the conditional probabilities discussed in the preceding section constitute a well-defined conditional probability measure $\mathcal{P}_\theta(A|t)$, on $(\mathcal{Y}, \mathcal{A})$. Traditionally, T is said to be a sufficient statistic for $\mathcal{M}$ (i.e., for θ) if this conditional distribution does not depend on θ,

$$\mathcal{P}_\theta(A|t) = \mathcal{P}(A|t) \qquad \forall \theta \in \Theta \tag{B.18}$$

Such a definition is heuristically justified in Section 2.1.1. Early works attempted to extend this concept to abstract probability spaces; for example, Halmos and Savage [24] defined: T is a sufficient statistic for $\mathcal{M}$ if, for every $A \in \mathcal{A}$, there exists $p = p(A|t)$ which is real-valued and meas.$[\mathcal{A}]$ on $\mathcal{Y}$ such that

$$P_\theta[A|t] = E_\theta[I_A |t] = p(A|t) \quad \text{a.s.}[\mathcal{P}_\theta T^{-1}] \tag{B.19}$$

Lehmann [35] (pp. 47-48) uses the same definition; in an early section of his work, Bahadur [3] (p. 429) rephrases the concept:

T is a sufficient statistic for $\mathcal{M}$ if, for each $A \in \mathcal{A}$, there exists a $[\mathcal{B}, \mathcal{N}]$-integrable function $\phi_A(t)$ such that

$$\int_{A \cap T^{-1}(B)} d\mathcal{P}_\theta = \int_B \phi_A(t)\, d\mathcal{P}_\theta T^{-1}(t) \qquad (B.20)$$

for all $B \in \mathcal{B}$, $\theta \in \Theta$.

These approaches are successful but, due to the explicit use of statistics, are unnecessarily cumbersome. As previously discussed, a more elegant formulation uses the concept of sub-algebras. (The development here will follow that of Bahadur [3].) A rigorous justification for eliminating T from explicit consideration and working only with $\mathcal{A}_0$ lies in the fact that the measure spaces $(\mathcal{Y}, \mathcal{A}_0, \mathcal{P})$ and $(\mathcal{T}, \mathcal{B}, \mathcal{P}T^{-1})$ are isomorphic and the isomorphism is independent of $\mathcal{P}$.[1] Thus, explicit consideration of $(\mathcal{T}, \mathcal{B})$, of the values t of T, and of the distributions $\mathcal{N} = \{\mathcal{P}_\theta T^{-1}\}$ of T is not essential to a study of T. One can equivalently study the distributions $\mathcal{M} = \{\mathcal{P}_\theta\}$ in the reduced sample space $(\mathcal{Y}, \mathcal{A}_0)$; for example, $\mathcal{N}$ is dominated on $\mathcal{B}$ iff $\mathcal{M}$ is dominated on $\mathcal{A}_0$, etc.[2]

---

[1] Halmos [23], p. 167.

[2] Bahadur [3], p. 430.

Accordingly, let $\mathscr{A}_0$ be an arbitrary sub-algebra of $\mathscr{A}$ in $(\mathscr{Y}, \mathscr{A}, \mathscr{P}_\theta)$, $\mathscr{P}_\theta \in \mathscr{M}$. Generalizing (B.19), one obtains[1]

DEFINITION B.1

$\mathscr{A}_0$ is said to be a sufficient sub-algebra for $\mathscr{M}$ if, for each $A \in \mathscr{A}$, there exists ϕ_A which is meas.$[\mathscr{A}_0]$ and does not depend on θ, such that

$$\phi_A(y) = E_\theta^{\mathscr{A}_0}[I_A(y)] \quad \text{a.s.}[\mathscr{A}, \mathscr{M}] \quad \blacksquare \tag{B.21}$$

This is readily seen to be equivalent to (B.20) also. In accordance with the heuristic concept of a necessary statistic (Section 2.1.1), a necessary and sufficient sub-algebra should be the "coarsest" of all sufficient sub-algebras; accordingly one makes

DEFINITION B.2

A sub-algebra $\mathscr{A}_* \subset \mathscr{A}$ is necessary for $\mathscr{M}$ (on $\mathscr{A}$) if $\mathscr{A}_* \subset \mathscr{A}_0$ a.s.$[\mathscr{A}, \mathscr{M}]$ for every sub-algebra $\mathscr{A}_0$ which is sufficient for $\mathscr{M}$. $\blacksquare$

---

[1] Bahadur [3], p. 430.

The following facts are clear intuitively; all are proven in Bahadur [3]:

Lemma B.4.

If T is a sufficient statistic for $\mathcal{M}$, then there exists $\mathcal{B}_0 \subset \mathcal{B}$ such that $\mathcal{A}_0 = T^{-1}(\mathcal{B}_0)$ is a necessary and sufficient sub-algebra for $\mathcal{M}$. ∎

Lemma B.5.

(i) If $\mathcal{A}_0 \subset \mathcal{A}' \subset \mathcal{A}$ and $\mathcal{A}_0$ is sufficient for $\mathcal{M}$, then $\mathcal{A}'$ is sufficient for $\mathcal{M}$.

Suppose $\mathcal{A}^* \subset \mathcal{A}$ is necessary and sufficient for $\mathcal{M}$.

Then:

(ii) $\mathcal{A}'$ is necessary for $\mathcal{M}$ iff $\mathcal{A}' \subset \mathcal{A}^*$.

(iii) $\mathcal{A}'$ is sufficient for $\mathcal{M}$ iff $\mathcal{A}^* \subset \mathcal{A}'$.

(iv) $\mathcal{A}'$ is necessary and sufficient for $\mathcal{M}$ iff $\mathcal{A}' = \mathcal{A}^*$.

(v) The elementary sufficient sub-algebra is $\mathcal{A}$.

(vi) The elementary necessary sub-algebra is $\{\mathcal{Y}, \Phi\}$ where Φ is the empty set. ∎

Assume from now on that $\mathcal{M}$ is dominated by a σ-finite measure λ. Then there exists a countable subset $\mathcal{M}_0 = \{ \mathcal{P}_{\theta_1}, \mathcal{P}_{\theta_2} \ldots \}$ of $\mathcal{M}$ such that $\mathcal{P}_\theta \equiv \mathcal{P}_{\theta_i}$ for

$i = 1, 2, \ldots$ and all $\theta \in \Theta$.[1] Choose a sequence $\{c_i\}$, $\sum_i c_i = 1$, and put

$$\lambda_0(A) = \sum_i c_i \mathscr{P}_{\theta_i}(A) , \quad A \in \mathscr{A} \tag{B.22}$$

THEOREM B.2[2]

As defined, $\lambda_0 \equiv \mathscr{M}$. The sub-algebra $\mathscr{A}_0$ is sufficient for $\mathscr{M}$ on $\mathscr{A}$ if and only if, for each $\mathscr{P}_\theta$, there exists a non-negative meas.$[\mathscr{A}_0]$ function g_θ such that

$$d\mathscr{P}_\theta = g_\theta(y)\, d\lambda_0 \quad \text{a.s.}[\mathscr{A}] \ \blacksquare \tag{B.23}$$

In other words, $\mathscr{A}_0$ is a sufficient sub-algebra iff each Radon-Nikodym derivative $\frac{d\mathscr{P}_\theta}{d\lambda_0}$ is meas.$[\mathscr{A}_0]$. As an easy corollary, one obtains a generalization of the factorization theorem. This will be stated in terms of a statistic:

---

[1] See Halmos and Savage [24] p. 232, or Lehmann [35] p. 354.

[2] This theorem was first proven by Halmos and Savage [24], p. 233, and was restated in the present form by Bahadur [3], p. 437.

Corollary[1]

If $\mathcal{M} << \lambda$, λ is σ-finite on $\mathcal{A}$,

Then $T : (\mathcal{Y}, \mathcal{A}) \to (\mathcal{T}, \mathcal{B})$ is a sufficient statistic for $\mathcal{M}$ iff there exists a nonnegative function $h : (\mathcal{Y}, \mathcal{A}) \to (R^1, \mathcal{R}^1)$ and a family of nonnegative functions $\{g_\theta ; \theta \in \Theta\}$, $g_\theta : (\mathcal{T}, \mathcal{B}) \to (R^1, \mathcal{R}^1)$, such that

$$d\mathcal{P}_\theta = h(y)\, g_\theta[T(y)]\, d\lambda \quad \text{a.s.}[\mathcal{A}] \quad \blacksquare \qquad (B.24)$$

Note that here it is the p.d.f. $[\lambda]$ of $\mathcal{P}_\theta$ which is factored. Recall from Lemma B.1 that the composition $g_\theta[T(y)]$ is meas.$[\mathcal{A}_0]$.

Now recall the measure λ_0 of the theorem. Since $\lambda_0 \equiv \mathcal{M}$, there corresponds to each $\theta \in \Theta$ a nonnegative, meas.$[\mathcal{A}]$ function $f_\theta(y)$ such that

$$d\mathcal{P}_\theta = f_\theta(y)\, d\lambda_0 \qquad (B.25)$$

Let $A_\theta(r) \subset \mathcal{Y}$ be the set

---

[1] Strictly speaking, this corollary only follows from Theorem B.2 if λ_0 is used in place of λ; the fact that it is true as stated is easily proven (e.g., Lehmann [35], p. 49). The generalization makes the result much more useful; in applications, one almost always has the case where λ is Lebesgue measure or where $\lambda = \lambda_0 \in \mathcal{M}$.

$$A_\theta(r) = \{y : f_\theta(y) < r\} \quad , \quad 0 < r < \infty \qquad \text{(B.26)}$$

This is the inverse image of a Borel set and hence is meas.$[\mathscr{A}]$.

Let $\mathscr{A}^* \subset \mathscr{A}$ be the σ-algebra generated by the $\{A_\theta(r); \theta \epsilon \Theta$, $r \epsilon (0, \infty)\}$.

THEOREM B.3[1]

$\mathscr{A}^*$ is a necessary and sufficient sub-algebra for $\mathscr{M}$. ∎

[1] Bahadur [3], p. 439.

APPENDIX C

QUASI-BAYESIAN: THE USE OF UTILITY MEASURES

In typical statistical treatments of Bayesian estimation and detection (see Section 1.3), the a priori p.d.f. on the parameters is interpreted strictly as a "relative frequency of occurrence" measure; under loose assumptions, its operational function is to enhance estimator performance in those regions of the parameter space where the "true" parameter, θ^*, is a priori considered most likely to lie. This is clear if one notes, from (1.27) for example, that the a posteriori p.d.f. is just the classical likelihood function multiplied by the a priori p.d.f. and normalized; this makes the Bayes' estimate closely related to the maximum likelihood estimate, but biases it toward those regions of Θ where the a priori p.d.f. places most of its mass.

Suppose, for example, that the conditions of the problem admit the a posteriori mean as an optimum estimate.[1] Since the likelihood function is objective, it is clear that the a priori p.d.f. serves to bias the estimate toward those values where θ^* is subjectively presumed most likely.

---

[1] See Van Trees [61], pp. 59-62.

More generally, note that (1.33) gives

$$\mathcal{R} = \int_{\Theta}\int_{\mathcal{Y}} J[\hat{\theta}(y),\ \theta^*]\ f_0(\theta^*) f(y|\theta^*)\, d\theta^* dy \qquad (C.1)$$

which is to be minimized by choice of $\hat{\theta}(y)$. The discussion above follows upon observing that the product of the first two terms in the integrand may be viewed as a new cost functional, say $\tilde{J}[\theta(y), \theta^*]$.

In engineering practice it is common to encounter system specifications which do not reflect a cost on the relationship between the estimate $\hat{\theta}$ and the true value θ^* , but which specify only that the equipment (estimator) is to perform better in some regions of Θ than in others. For instance, it might be desired to build a receiver which performs most efficiently for small signals even though large signals are known to be prevalent. From above, it is clear that such a specification can just as well be incorporated into the a priori p.d.f. (which must then be interpreted differently) as into the cost functional. In fact, such a specification can be translated into a nonnegative normalized weighting function, say $v(\theta)$, and used in place of an a priori p.d.f. even when no prior distribution is known and one is not willing to consider θ as a random variable. This can make Bayesian techniques acceptable when the philosophy is not.

1. Quasi-Bayesian Estimation. Assume (or construct) the cost

functional $J[\hat{\theta}(y), \theta^*]$ to have no factors which depend only on θ^* (usually, it will be $J[\hat{\theta} - \theta^*]$); incorporate all specifications of the type discussed above into a nonnegative, bounded, integrable weighting function $v(\theta)$ on Θ. Construct a utility density function $w(\theta)$ as follows:

$$w(\theta) = \begin{cases} f_0(\theta) & \text{if no } v(\theta) \text{ exists} \\ \dfrac{f_0(\theta)\, v(\theta)}{\int_\Theta f_0(\theta')\, v(\theta')\, d\theta} & \text{if both exists} \\ \dfrac{v(\theta)}{\int_\Theta v(\theta')\, d\theta'} & \text{if no } f_0(\theta) \text{ exists} \end{cases} \qquad (C.2)$$

The first case is the classical Bayesian one, and $w(\theta)$ represents a (possibly subjective) relative frequency of occurrence. In the last case, $w(\theta)$ represents a pure utility and θ need not even be considered random; the integral over $\mathcal{Y}$ in (C.1) can be viewed as an expectation, and that over Θ as an ordinary integral. The second case is, of course, a combination of the other two.

Once $w(\theta)$ is defined one proceeds with all the Bayesian techniques, using $w(\theta)$ in place of $f_0(\theta)$. Clearly, much advantage is gained if $w(\theta)$ can be chosen in the class of natural conjugate

densities defined in Section 3.1.1 (provided that sufficient statistic exist) or can be represented as in (3.19) using an $r(\theta)$ which makes the integral in (3.23) tractable.

2. Quasi-Bayesian Detection. Identical comments can be made for the compound hypotheses detection problem; the optimal decision statistic becomes the "marginal" likelihood ratio averaged over the utility densities. This is best illustrated with an example.

Example C.1

Suppose H_1 occurs with probability $P(H_1) = p$, and that $f(y|\theta, H_1)$ and an a priori p.d.f. $f_0(\theta)$ are known. Similarly, $P(H_0) = 1 - p$ and $f(y|\eta, H_0)$ and $f_0(\eta)$ are known. A "reward structure" is established as follows:

Decision \ Hypothesis	H_1	H_0	
D_1	$v_{11}(\theta)$	$v_{10}(\eta)$	, $v_{11}(\theta) > v_{01}(\theta)$
D_0	$v_{01}(\theta)$	$v_{00}(\eta)$	, $v_{00}(\eta) > v_{10}(\eta)$

Such a structure lends itself readily to emphasizing performance in certain regions of Θ and $\mathcal{N}$. The objective is to maximize the expectation of the reward,

$$V = P(H_1) \int_{\Theta} \int_{\mathcal{Y}} [v_{01}(\theta)\, P(D_0|y) + v_{11}(\theta)\, P(D_1|y)] \cdot f(y|\theta, H_1) f_0(\theta)\, dy\, d\theta + P(H_0) \int_{\mathcal{N}} \int_{\mathcal{Y}} [v_{00}(\eta)\, P(D_0|y) + v_{10}(\eta)\, P(D_1|y)] \cdot f(y|\eta, H_0) f_0(\eta)\, dy\, d\eta \tag{C.3}$$

Denote a randomized decision function as

$$g(y) = P(D_1|y) = 1 - P(D_0|y) \tag{C.4}$$

Use this in (C.3); after considerably rearranging, it is seen that maximizing V is equivalent to maximizing

$$\tilde{V} = \int_{\mathcal{Y}} g(y) \left\{ \frac{\int_{\Theta} f(y|\theta', H_1) [v_{11}(\theta') - v_{01}(\theta')] f_0(\theta')\, d\theta'}{\int_{\mathcal{N}} f(y|\eta', H_0) [v_{00}(\eta') - v_{10}(\eta')] f_0(\eta')\, d\eta'} - \frac{(1-p)}{p} \right\} \int_{\mathcal{N}} f(y|\eta, H_0) f_0(\eta)\, d\eta\, dy \tag{C.5}$$

Clearly, $\tilde{V}$ is maximized by putting $g(y) = 1$ when the term in braces is positive, $g(y) = 0$ otherwise. In the notation developed prior to this example, the result can be summarized:

Define the utility densities

$$w_1(\theta) = \frac{|v_{11}(\theta)-v_{01}(\theta)|\, f_0(\theta)}{\int_{\Theta} (\text{Numerator})\, d\theta} \tag{C.6}$$

$$w_0(\eta) = \frac{[v_{00}(\eta)-v_{10}(\eta)]\, f_0(\eta)}{\int_{\mathcal{N}} (\text{Numerator})\, d\eta} \tag{C.7}$$

Then all Bayesian results apply with these utility densities used in place of the a priori densities; the threshold for the "utility-averaged" likelihood ratio is

$$\alpha = \frac{(1-p)\int_{\mathcal{N}} f_0(\eta)[v_{00}(\eta)-v_{10}(\eta)]\, d\eta}{p\int_{\Theta} f_0(\theta)[v_{11}(\theta)-v_{01}(\theta)]\, d\theta} \tag{C.8}$$

All the results discussed here may be applied to sequential and continuous observations as well. For the sake of brevity, the body of this dissertation proceeds on the assumption that a priori p.d.f.'s are probabilities in the strict, classical sense.

APPENDIX D

PROOFS AND DERIVATIONS

D.1 Proof of Theorem 2.5

As discussed following the statement of the theorem, this proof closely parallels the proof of Dynkin's theorem and thus will only be outlined here.[1] It follows from three lemmas:

Lemma 1. For any $\varphi(\tilde{\underline{y}}) \in V_L$, the statistic of a sample of size k given by

$$t(\underline{Y}_k) = \sum_{j=1}^{k} \varphi(\tilde{\underline{y}}) \qquad \text{(D. 1)}$$

is a necessary statistic for θ. ∎

This is proved by noting that by the definition of V_L, φ can be written

$$\varphi(\tilde{\underline{y}}) = \sum_{q=1}^{r} c_q \, L(\theta_q; \tilde{\underline{y}}) + c_0 \qquad \text{(D. 2)}$$

Use this in (D. 1) to see that $t(\underline{Y}_k)$ is dependent on $L_k(\theta; \underline{Y}_k)$ which is a necessary statistic.

---

(1) See Dynkin [34], pp. 24-25.

Lemma 2. If $\{1, \varphi_1(\tilde{\underline{y}}), \ldots, \varphi_s(\tilde{\underline{y}})\}$, $s \geq r$, span V_L then the s-vector of statistics $\underline{t}_k(\underline{Y}_k)$, where

$$t_{ki}(\underline{Y}_k) = \sum_{j=1}^{k} \varphi_i(\tilde{y}_j) \; ; \; i = 1, 2, \ldots s \qquad (D.3)$$

is sufficient for θ . ∎

To prove Lemma 2, note that by assumption

$$L(\theta ; \tilde{\underline{y}}) = a_0(\theta) + \sum_{j=1}^{s} a_j(\theta) \varphi_j(\tilde{\underline{y}}) \qquad \forall \theta \in \Theta \qquad (D.4)$$

Use (D. 4) to rewrite (2.30) and then use (D. 3) to replace the resulting $\varphi_j(\tilde{\underline{y}})$ terms by $t_{kj}(\underline{Y}_k)$. This shows $L_k(\theta ; \underline{Y}_k)$, which is a sufficient statistic for θ , to be dependent on $\{t_{ki}(\underline{Y}_k), i = 1 \ldots s\}$.

Lemma 3. If $\{1, \varphi_1(\tilde{\underline{y}}), \ldots, \varphi_s(\tilde{\underline{y}})\}$, $s \leq r$, are linearly independent in V_L , then the s-vector of statistics defined in (D. 3) is functionally independent. ∎

This is proven by induction using a rather lengthy classical calculus argument (see [13] p. 26); it is true in essence because the $L(\theta ; \underline{Y}_k)$ are piecewise smooth.

Part b. of the theorem follows directly from these lemmas;

part a. requires a minor amount of further work using Lemma 3.

D.2 Proof of Theorem 3.1

The proof will be given in sequential notation, so that "a priori" refers to the state prior to observing y_{k+1}, and hence is posterior to $\underline{Y}_k$; "a posteriori" refers to $\underline{Y}_{k+1}$. Clearly, if $p(\theta;\gamma)$ reproduces under any y_{k+1}, then by induction it reproduces.

For an arbitrary a priori p.d.f. $f(\theta \mid \underline{Y}_k)$, the a posteriori density is given by (1.32)

$$f(\theta \mid \underline{Y}_{k+1}) = \frac{f(y_{k+1} \mid \underline{Y}_k, \theta)\, f(\theta \mid \underline{Y}_k)}{\int_{\Theta} (\text{Numerator})\, d\theta} \tag{D.5}$$

This will be used shortly; for now, consider the identity

$$f(\underline{Y}_{k+1} \mid \theta) = f(y_{k+1} \mid \underline{Y}_k, \theta)\, f(\underline{Y}_k \mid \theta) \tag{D.6}$$

Since a sufficient statistic of fixed dimension exists by assumption, both sides may be factored as in (2.10). Replace $t_k(\underline{Y}_k)$ by γ_k and normalize the $g[\gamma;\theta]$ functions on both sides. Using (3.8) and rearranging,

$$p(\theta\,;\gamma_{k+1}) = \left\{ \frac{G(\underline{Y}_k) \int g[\gamma_k, \theta\,']\, d\theta\,'}{G(\underline{Y}_{k+1}) \int g[\gamma_{k+1}, \theta\,'']\, d\theta\,''} \right\} \cdot f(y_{k+1} | \underline{Y}_k, \theta)\, p(\theta\,;\gamma_k) \tag{D.7}$$

The left side was constructed to be a probability density on Θ ; hence the term in braces on the right is a normalizing constant, and

$$p(\theta\,;\gamma_{k+1}) = \frac{f(y_{k+1} | \underline{Y}_k, \theta)\, p(\theta\,;\gamma_k)}{\int_\Theta (\text{Numerator})\, d\theta} \tag{D.8}$$

Compare with (D. 5), which holds for arbitrary a priori p. d. f. 's; clearly, $p(\theta\,;\gamma)$ reproduces.

D. 3 Proof of Theorem 3. 2

Using $f_n(\theta)$, n = 1, 2, as an a priori p. d. f. yields

$$f_n(\theta | \underline{Y}_k) = \frac{f_n(\theta)\, f(\underline{Y}_k | \theta)}{f_n(\underline{Y}_k)} \tag{D.9}$$

Note that $f(\underline{Y}_k | \theta)$ does not depend on n (i. e., on the prior). Put n = 2 and use (3.20),

$$f_2(\theta \mid \underline{Y}_k) = \frac{r_\theta(\theta) f_1(\theta) f(\underline{Y}_k \mid \theta)}{f_2(\underline{Y}_k)} \qquad \text{(D. 10)}$$

Now put $n = 1$ in (D. 9), and use the result to rewrite the last two terms of the numerator of (D. 10). Rearranging,

$$f_2(\theta \mid \underline{Y}_k) f_2(\underline{Y}_k) = f_1(\underline{Y}_k) r_\theta(\theta) f_1(\theta \mid \underline{Y}_k) \qquad \text{(D. 11)}$$

Integrate on Θ to verify (3.22). Then divide (D. 11) by (3.22) to verify (3.21).

D.4 The R-N Derivative and Natural Conjugate Density for Continuous M-SAG Noise

Consider the R-N derivative of (5.27) where the coefficients D_{jk} are defined by (5.2 6) and the A_k by

$$\sum_{i,\, j=0}^{M} \theta_i \theta_j z^{M-i} z^{M-j} (-1)^{M-i} = \sum_{k=0}^{M} A_k z^{2k} \qquad \text{(D.12)}$$

All terms for which $i+j$ is odd cancel on the left, so the expression makes sense. To obtain an explicit expression for A_k, fix k; this requires that

$$2M - (i+j) = 2k \qquad \text{(D.13)}$$

and relates the indices i and j, collapsing the double sum. Use (D.13) to eliminate the sum over j; recall that $0 \leq j \leq M$,

which restricts the limits on i :

$$A_k = \sum_{i=\max(0,\, M-2k)}^{\min[\, M,\, 2(M-k)]} \theta_i \theta_{2(M-k)-i} (-1)^{M-i} \qquad (D.14)$$

This verifies (5.29).

The Natural Conjugate density is found by re-writing (5.33) so that it has the form

$$f(y_t \mid \theta, \underline{y}_0) = K \exp \left\{ -\tfrac{1}{2} \sum_{i,\, j=0}^{M} t_{ij} \theta_i \theta_j \right\} \qquad (D.15)$$

Consider the last two terms in the exponent of (5.33), which have the form

$$-\tfrac{1}{2} \left\{ \sum_{k=0}^{M-1} (-1)^k A_k J^{(k)}(y) + \tfrac{1}{2} \sum_{\substack{k=0 \\ j+k \text{ even}}}^{M-1} \sum_{j=0}^{M-1} D_{jk} E^{(j,\,k)}(y) \right\} \qquad (D.16)$$

where

$$J^{(k)}(y) \triangleq \int_0^T [\, y_t^{(k)}]^2 \, dt \qquad (D.17)$$

$$E^{(j,\,k)}(y) \triangleq y_T^{(j)} y_T^{(k)} - y_0^{(j)} y_0^{(k)} \qquad (D.18)$$

and D_{jk} is given by (5.26). It is necessary to re-arrange the resulting double and triple sums to obtain explicitly the coefficients

of $\theta_i \theta_j$ as in (D.15). This is best done by inspection; consider the first term in (D.16), which may be re-written using (D.12) and (D.13):

$$\sum_{k=0}^{M-1} (-1)^k A_k J^{(k)}(y)$$

$$= \sum_{k=0}^{M-1} \sum_{i=0}^{M-1} \sum_{j=0}^{M-1} \theta_i \theta_j (-1)^{M-i} (-1)^k J^{(k)}(y) \qquad \text{(D.19)}$$

$$\underbrace{\qquad\qquad}_{2M-(i+j)=2k}$$

Only terms for which $(i+j)$ is even appear; each of these has a coefficient

$$t'_{ij} = (-1)^{2M-i-\frac{i+j}{2}} J^{(M-\frac{i+j}{2})}(y) \qquad \text{(D.20)}$$

where k has been eliminated through use of (D.13); $k \leq M+1$ means that $i+j$ must be strictly positive. The exponent on (-1) can be simplified without changing its effect; one finally obtains

$$\sum_{k=0}^{M-1} (-1)^k A_k J^{(k)}(y) = \sum_{\substack{i,j=1 \\ i+j \text{ even}}}^{M} \theta_i \theta_j \left[(-1)^{\frac{3i+j}{2}} J^{(M-\frac{i+j}{2})}(y) \right] \qquad \text{(D.21)}$$

Now consider the second term of (D.16). By inspection of (5.26) one concludes that, as j, k run over their allowable range, one obtains all products $\theta_p \theta_q$ such that $p > q$ and $p+q$ is odd.

Further, $D_{jk} = D_{kj}$ and both occur. Now use (5.26) in (D.16):

$$\tfrac{1}{2}\sum_{k=0}^{M-1}\sum_{j=0}^{M-1} D_{jk}\, E^{(j,k)}(y) \tag{D.22}$$

$$= \sum_{k=0}^{M-1}\sum_{j=0}^{M-1}\sum_{p} \theta_p \theta_{2M-p-j-k-1}\,(-1)^{p+j-M} E^{(j,k)}(y)$$

where the substitution $p = M-i$ has been made, the sum on p runs over all values for which the subscripts on θ make sense. To find the coefficient of an arbitrary $\theta_p\theta_q$ satisfying the above restrictions, fix p and call the second subscript q ; use that relation to eliminate j , changing that sum to one over q . Fix q :

$$t'_{pq} = \sum_{k}(-1)^{M-q-k-1} E^{(2M-p-q-k-1,\,k)}(y) \tag{D.23}$$

and the sum runs over all values $0 \leq k \leq M-1$ for which it makes sense, i.e., for which $0 \leq 2M-p-q-k-1 \leq M-1$. Thus,

$$\tfrac{1}{2}\sum_{\substack{k=0\\ j+k \text{ even}}}^{M-1}\sum_{j=0}^{M-1} D_{jk}\, E^{(j,k)}(y)$$

$$= \sum_{p=0}^{M-1}\sum_{\substack{q=p+1\\ p+q \text{ odd}}}^{M-1} t'_{pq}\,\theta_p\,\theta_q \tag{D.24}$$

Combine all this to write (D.16) as follows:

$$(D.16) = -\tfrac{1}{2}\sum_{i=0}^{M}\sum_{j=0}^{M} t'_{ij}\theta_i\theta_j \qquad (D.25)$$

where

$$t'_{ij} = \begin{cases} (-1)^{\frac{3i+j}{2}} J^{(M-\frac{i+j}{2})}(y) \; ; \; i+j > 0, \text{ even} \\ \tfrac{1}{2}\sum_k (-1)^{M-k-1-\min(i,j)} E^{(2M-i-j-k-1,\,k)}(y) \; ; \; i+j \text{ odd} \end{cases} \qquad (D.26)$$

and k runs over $\max(0, M-i-j) \leq k \leq \min(M-1, 2M-i-j-1)$.

Note that $t'_{ij} = t'_{ji}$, and $t'_{00} = 0$.

Since θ_0 is not estimable, it is desirable to separate it from the other parameters; since the sum is symmetric, (D.25) can be written

$$(D.16) = \tfrac{1}{2}\left\{ 2\theta_0 \sum_{j=1}^{M} t'_{0j}\theta_j + \sum_{i=1}^{M}\sum_{j=1}^{M} t_{ij}\theta_i\theta_j \right\} \qquad (D.27)$$

where $t_{ij} = t'_{ij}$ as in (D.26) and where

$$t'_{0j} = \begin{cases} (-1)^{j/2} J^{(M-j/2)}(y) & ; \; j \text{ even} \\ \tfrac{1}{2}\sum_k (-1)^{M-k-1} E^{(2M-j-k-1,\,k)}(y) & ; \; j \text{ odd} \end{cases} \qquad (D.28)$$

Define the parameter vector

$$\underline{\theta} = (\theta_1, \ldots, \theta_M)^* \tag{D.29}$$

Putting all these results together, the "density" of (5.33) can be written

$$f(y_t | \underline{\theta}, \underline{y}_0) = (2\pi)^{\frac{M}{2}} \exp\left\{-\tfrac{1}{2}\left[\underline{\theta}^* T(y_t)\underline{\theta} + t^*(y_t)\underline{\theta}\right]\right\} \tag{D.29}$$

where the elements of the sufficient statistic vector and matrix are as given in (5.38) and (5.39).

D.5 Lemma[1]

If $f(t)$, $0 \leq t \leq T$, is a real-valued continuous function with nonzero quadratic variation, then $f(t)$ is of unbounded variation on $[0, T]$.

Proof:

Let P_k be a partition of $[0, T]$ with modulus $|P_k|$. Then the quadratic variation of f is

---

[1] This lemma was taken from a set of lecture notes by Prof. W. L. Root.

$$Q_f[0, T] \stackrel{d}{=} \lim_{\substack{k \to \infty \\ |P_k| \to 0}} \sum_{i \leq k} |f(t_i) - f(t_{i-1})|^2$$

$$\leq \lim \left[\max_{i \leq k} |f(t_i) - f(t_{i-1})| \right] \left[\sum_{i \leq k} |f(t_i) - f(t_{i-1})| \right]$$

The first term $\to 0$ since f is continuous; the limit of the second is, by definition, the total variation $V_f[0, T]$.

So assume $f(t)$ is continuous with nonzero $Q_f[0, T]$, and $V_f[0, T] < \infty$. Then $Q_f[0, T] = 0$, a contradiction.

APPENDIX E

R-N DERIVATIVES FOR THE 1-SAG PROCESS

E.1 The R-N Derivative

The 1-SAG process with parameter q_1 has spectral density function

$$S_y(f^2) = \frac{a^2}{(2\pi f)^2 + q_1^2} \qquad \text{(E.1)}$$

and autocorrelation

$$R_y(\tau) = \frac{a^2}{2q_1} \exp(-q_1|\tau|) \quad , \quad q_1 > 0 \qquad \text{(E.2)}$$

Let $\{y_t ; t \in [0, T]\}$ be a finite segment of the process; let $\mathscr{P}_{q_1}$ be the measure on the space of sample functions $(\mathscr{Y}, \mathscr{A})$ which is induced by the process, and let $\mathscr{P}_{q^*}$ be the measure induced by a 1-SAG process with parameter $q^* > 0$. From Section 5.2, $\mathscr{P}_{q_1} \equiv \mathscr{P}_{q^*}$.

Suppose y_t is sampled as usual on $[0, T]$, see Section 1.1. The autocorrelation matrix of any two adjacent samples is, since the process is stationary,

$$R = \begin{bmatrix} r_0 & r_1 \\ r_1 & r_0 \end{bmatrix} \qquad \text{(E.3)}$$

where, from (E.2),

$$r_0 = \frac{a^2}{2q_1}$$

$$r_1 = \frac{a^2}{2q_1} e^{-q_1\delta} \qquad \text{(E.4)}$$

and $\delta = T/k$ is the sampling interval. From (4.46), it is clear that the samples may be considered as having been generated by a discrete autoregression with parameters

$$\beta_1 = -e^{-q_1\delta}$$

$$\alpha^2 = \frac{a^2}{2q_1}\left(1 - e^{-2q_1\delta}\right) \qquad \text{(E.5)}$$

and the joint density of the k samples, conditioned on y_0, is given by (4.48). If the parameter of the process was q^*, then of course all the above holds with q_1 replaced by q^*; it is desired to find

$$\lambda(y_t|y_0;q_1) = \lim_{k\to\infty} \frac{f(\underline{Y}_k|y_0,q_1)}{f(\underline{Y}_k|y_0,q^*)} \qquad \text{(E.6)}$$

Before passing to the limit, it is desirable to rearrange the exponent of (4.48) to obtain sums of the observation which will converge. These sums turn out to be:

$$\sum_{i=1}^{k} y_i^2 \delta \to \int_0^T y_t^2 \, dt \tag{E.7}$$

$$\sum_{i=1}^{k} (y_i - y_{i-1})^2 \to a^2 T \tag{E.8}$$

$$(y_0, y_k) = (y_0, y_T) \tag{E.9}$$

where $\delta = T/k$. Equation (E. 7) is true because y_t is a. s. sample-function continuous and has finite average power; (E. 8) follows from Baxter's theorem, see (5.15); (E. 9) is obvious. After some manipulation, and using (E. 5), one can rewrite (4.48) in terms of these sums:

$$f(\underline{Y}_k | y_0, q_1) = \left[\frac{q_1}{\pi a^2 (1-\beta_1^2)} \right]^{\frac{k}{2}} \cdot \exp \left\{ - \frac{q_1}{a^2} \left[\frac{1+\beta_1}{1-\beta_1} \cdot \frac{1}{\delta} \sum_{i=0}^{k-1} y_i^2 \delta - \frac{\beta_1}{1-\beta_1^2} \sum_{1}^{k} (y_i - y_{i-1})^2 + \frac{1}{1-\beta_1} (y_k^2 - y_o^2) \right] \right\} \tag{E.10}$$

where $\beta_1 = - e^{-q_1 \delta}$. Again, the denominator of (E. 6) is similar.

The likelihood ratio function has the form

$$\Lambda_k(\underline{Y}_k|q_1, y_0) = c_k(q_1) \exp\left\{ -\frac{1}{a^2}\left[d_k(q_1) \sum_0^{k-1} y_i^2 \delta + e_k(q_1) \sum_1^k (y_i - y_{i-1})^2 + f_k(q_1)\,(y_k^2 - y_0^2)\right]\right\} \tag{E. 11}$$

The limits of the various coefficient functions will be found separately:

Limit for $c_k(q_1)$:

$$c_k(q_1) = \left[\frac{q_1}{1-e^{-2q_1\delta}} \cdot \frac{1-e^{-2q^*\delta}}{q^*} \right]^{\frac{k}{2}} \tag{E. 12}$$

Now for any $\alpha > 0$,

$$\lim_{k\to\infty} \left[\frac{k}{\alpha}\left(1 - e^{-\alpha/k}\right)\right]^{-k} = \lim_{k\to\infty} \left\{ \frac{k}{\alpha}\left[\frac{\alpha}{k} - \frac{1}{2}\left(\frac{\alpha}{k}\right)^2 + \mathcal{O}\left(\frac{\alpha}{k}\right)^3\right]\right\}^{-k}$$

$$= \exp\left\{ \lim_{k\to\infty} \left[-k\,\ell n \left\{ 1 - \frac{\alpha}{2k} + \frac{k}{\alpha} \cdot \mathcal{O}\left(\frac{\alpha}{k}\right)^3 \right\}\right]\right\}$$

$$= \exp\left(\frac{\alpha}{2}\right)$$

since $\ell n(1-\epsilon) \to -\epsilon$ as $\epsilon \to 0^+$. Using this in (E. 12),

$$c_k(q_1) \to \exp\left[\frac{1}{2}(q_1 - q^*)\,T\right] \quad \text{as} \quad k \to \infty \qquad \text{(E. 13)}$$

■

Limit for $d_k(q_1)$:

$$d_k(q_1) = \frac{1}{\delta}\left[q_1 \frac{1-\exp(-q_1\delta)}{1+\exp(-q_1\delta)} - q^* \frac{1-\exp(-q^*\delta)}{1+\exp(-q^*\delta)}\right] \qquad \text{(E. 14)}$$

$$\to \frac{1}{2}(q_1^2 - q^{*2}) \quad \text{as} \quad \delta \to 0 \qquad \text{(E. 15)}$$

as can be verified using a first-order Taylor approximation. ■

Limit for $e_k(q_1)$:

$$e_k(q_1) = \frac{q_1 \exp(-q_1\delta)}{1-\exp(-2q_1\delta)} - \frac{q^* \exp(-q^*\delta)}{1-\exp(-2q^*\delta)} \qquad \text{(E. 16)}$$

$$= \frac{1}{\delta}\left[q_1\delta\ \operatorname{csch}(q_1\delta) - q^*\delta\ \operatorname{csch}(q^*\delta)\right]$$

$$\to 0 \quad \text{as} \quad \delta \to 0 \qquad \text{(E. 17)}$$

as can be verified using a series approximation for $\operatorname{csch}(\cdot)$.[1] ■

---

[1] See [1], p. 85, #4.5.65.

Limit for $f_k(q_1)$:

$$f_k(q_1) = \frac{q_1}{1+\exp(-q_1\delta)} - \frac{q^*}{1+\exp(-q^*\delta)} \qquad \text{(E. 18)}$$

$$\to \frac{1}{2}(q_1 - q^*) \quad \text{as} \quad \delta \to 0 \qquad \text{(E. 19)}$$

Put these four results into (E. 11); take the limit, using (E. 7) - (E. 9) :

$$\lambda(y_t|y_0;q_1) = \exp\left\{ -\frac{1}{2a^2}\left[(q_1^2 - q^{*2}) \int_0^T y_t^2\, dt - (q_1-q^*)\left[a^2\, T + y_0^2 - y_T^2\right]\right]\right\} \qquad \text{(E. 20)}$$

The "nuisance parameter" q^* will cancel if this is used in Bayes' rule, leaving (5. 56) ; if the remaining terms are multiplied by $f(y_0|q_1)$ as given in (5.55), one obtains Hajek's result (5.57).

E. 2 The Detection Statistic

Suppose one wishes to solve the detection problem of (5. 1) using (3. 37). To account for the fact that (E. 20) is conditioned on y_0 , one chooses the a priori p. d. f. from the class of (5. 68); upon observing y_0 , the a posteriori p. d. f. is natural conjugate as in (5. 57), with parameter

$$\Psi^{(0)} = \Psi^{-} + \begin{bmatrix} 0 \\ -2y_0^2 \end{bmatrix} \qquad \text{(E. 21)}$$

After observing y_t , $0 \leq t \leq T$, the parameter is further updated using (5.58) and becomes

$$\Psi^{(T)} = \Psi^{-} + \begin{bmatrix} \int_0^T y_t^2 \, dt \\ a^2 T - (y_0^2 + y_T^2) \end{bmatrix} \qquad \text{(E. 22)}$$

The results are <u>not</u> conditioned on y_0 ; analogous results hold for H_1 with γ and $(y_t - s(t))$ replacing Ψ and y_t .

Equation (3. 37) can be written

$$\ell(y_t) = \frac{p_0(q_1;\gamma^{-})}{p_0(q_1;\Psi)} \frac{p(q_1;\Psi^{(T)})}{p(q_1;\gamma^{(T)})} \ell(y_t|q_1) \qquad \text{(E. 23)}$$

Call the ratio of densities above $L(y_t, q_1)$. By direct calculation using (E. 22), (5.68), and (5. 57), it is found that

$$L(y_t, q_1) = \frac{K_0(\gamma^{-})}{K_0(\Psi^{-})} \frac{K(\Psi^{(T)})}{K(\gamma^{(T)})} \exp\left\{ \frac{1}{2a^2} \left[q_1^2 \int_0^T s^2(t)\, dt - 2q_1^2 \int_0^T s(t) y_t dt - q_1 \left[2y_0 s(0) - s^2(0) + 2y_T s(T) - s^2(T) \right] \right] \right\} \qquad \text{(E. 24)}$$

The term $\ell(y_t|q_1)$ has already been found using the Metzger Model, and is given in (1.25) and (1.26); recall that $p_1 = q_1$. Comparing those equations with (E.23) shows that all terms which involve q_1 do indeed cancel, and one is left with

$$\ell(y_t) = \frac{K_0(\chi^-)}{K_0(\psi^-)} \frac{K(\psi^{(T)})}{K(\chi^{(T)})} \exp\left\{ a^{-2}\left[s'(0)\, y_0 + s'(T)\, y_T - \int_0^T s''(t)\, y_t\, dt - \frac{1}{2}\int_0^T [s'(t)]^2\, dt \right]\right\} \qquad \text{(E.25)}$$

For simplicity, assume that the a priori densities on q_1 were identical under H_0 and H_1 (i.e., $\psi^- = \chi^-$) so the first ratio above is 1. The second ratio can be found using (E.22) and (5.61), and then

$$\ell(y_t) = \left[\frac{\psi_1^- + \int_0^T y_t^2\, dt}{\psi_1^- + \int_0^T [y_t - s(t)]^2\, dt}\right]^{\frac{1}{2}} \Omega \left[\frac{\psi_2^- + a^2 T - (y_0^2 + y_T^2)}{2a\left(\psi_1^- + \int_0^T y_t^2\, dt\right)}\right]$$

$$\cdot\ \Omega^{-1}\left[\frac{\psi_2^- + a^2 T - [y_0 - s(0)]^2 - [y_T - s(T)]^2}{2a\left[\psi_1^- + \int_0^T \{y_t - s(t)\}^2\, dt\right]}\right]$$

$$\cdot\ \exp\left\{ a^{-2}\left[s'(0) y_0 + s'(T)\, y_T - \int_0^T s''(t)\, y_t\, dt - \frac{1}{2}\int_0^T \{s'(t)\}^2\, dt \right]\right\} \qquad \text{(E.26)}$$

It does not appear that any further significant simplifications are possible.

REFERENCES

1. M. Abramowitz and I. Stegun, ed., Handbook of Mathematical Functions with Formulas, Graphs, and Tables, National Bureau of Standards, U.S. Government Printing Office, Washington, 1970.

2. K. J. Aström, Introduction to Stochastic Control Theory, Academic Press, New York, 1970.

3. R. R. Bahadur, "Sufficiency and Statistical Decision Functions," Annals of Mathematical Statistics, 25, 1954, pp. 423-462.

4. G. Baxter, "A Strong Limit Theorem for Gaussian Processes," Proc. of the American Mathematical Society, 7, 1956, pp. 522-528.

5. T. G. Birdsall, "The Theory of Signal Detectability: ROC Curves and Their Character," Ph. D. Dissertation, The University of Michigan, Ann Arbor, August, 1966.

6. T. G. Birdsall, Adaptive Detection Receivers and Reproducing Densities, Cooley Electronics Laboratory Technical Report No. TR-194, The University of Michigan, Ann Arbor, July 1968.

7. T. Birdsall and J. Gobien, "Sufficient Statistics and Reproducing Densities in Simultaneous Sequential Detection and Estimation," to be published in the IEEE Trans. on Information Theory.

8. H. Cramer, Mathematical Methods of Statistics, Princeton University Press, Princeton, New Jersey, 1946.

9. W. B. Davenport and W. L. Root, An Introduction to the Theory of Random Signals and Noise, McGraw-Hill Book Company, Inc., New York, 1958.

10. M. H. DeGroot, Optimal Statistical Decisions, McGraw-Hill Book Company, New York, 1970.

11. J. L. Doob, "The Elementary Gaussian Processes," Annals of Mathematical Statistics, 15, 1944, pp. 229-282.

REFERENCES (Cont.)

12. J. L. Doob, Stochastic Processes, John Wiley and Sons Inc., New York, 1953.

13. E. B. Dynkin, "Necessary and Sufficient Statistics for a Family of Probability Distributions," Selected Translations in Mathematical Statistics and Probability, 1, 1961, pp. 17-40.

14. V. N. Faddeeva, Computational Methods of Linear Algebra, Dover Publications Inc., New York, 1959.

15. J. Feldman, "Equivalence and Perpendicularity of Gaussian Processes," Pacific Journal of Mathematics, 8, No. 4, 1958, pp. 699-708.

16. T. S. Ferguson, Mathematical Statistics, A Decision Theoretic Approach, Academic Press, New York, 1967.

17. A. Fredriksen, D. Middleton, D. Vandelinde, "Simultaneous Signal Detection and Estimation Under Multiple Hypotheses," IEEE Trans. On Information Theory, IT-18, No. 5, 1972, pp. 607-614.

18. U. Grenander, "Stochastic Processes and Statistical Inference," Arkiv for Matematik, 1, 1950, pp. 195-277.

19. U. Grenander and G. Szego, Toeplitz Forms and Their Application, University of California Press, Berkeley, 1958.

20. T. L. Grettenberg, A New Class of Structurally Invariant Learning Machines, Communication Theory Laboratory Technical Report No. T5-685/3111, California Inst. of Technology, Pasadena, Calif., July 1965.

21. J. Hajek, "On A Property of Normal Distribution of Any Stochastic Process," Selected Translations in Mathematical Statistics and Probability, 1, 1958, pp. 245-253.

22. J. Hajek, "On Linear Statistical Problems in Stochastic Processes," Czech Mathematics Journal, 12, 1962, pp. 404-443.

REFERENCES (Cont.)

23. P. R. Halmos, Measure Theory, D. Van Nostrand Company Inc., New York, 1950.

24. P. Halmos and L. Savage, "Application of the Radon-Nikodym Theorem to the Theory of Sufficient Statistics," Annals of Mathematical Statistics, 20, 1949, pp. 225-241.

25. C. W. Helstrom, "Solution of the Detection Integral Equation for Stationary Filtered White Noise," IEEE Trans. on Information Theory, IT-11, 1965, pp. 335-339.

26. C. W. Helstrom, Statistical Theory of Signal Detection, Pergamon Press, Oxford, 1968.

27. R. A. Howard, "Decision Analysis: Perspectives on Inference, Decision, and Experimentation," Proceedings of the IEEE, 58, No. 5, 1972, pp. 632-643.

28. R. Hogg and A. Craig, Introduction to Mathematical Statistics, The Macmillan Company, New York, 1965.

29. A. Jaffer and S. Gupta, "Recursive Bayesian Estimation with Uncertain Observation," IEEE Trans. on Information Theory, IT-17, No. 5, 1971, pp. 614-616.

30. A. Jaffer and S. Gupta, "Coupled Detection-Estimation of Gaussian Processes in Gaussian Noise," IEEE Trans. on Information Theory, IT-18, No. 1, 1972, pp. 106-110.

31. R. L. Kashyap, "Prior Probability and Uncertainty," IEEE Trans. on Information Theory, IT-17, No. 6, 1971, pp. 641-650.

32. E. J. Kelly, I. S. Reed, and W. L. Root, "The Detection of Radar Echoes in Noise Parts I and II," Journal of the Society on Industrial and Applied Mathematics, 8, 1960, pp. 309-341 and 481-505.

33. B. O. Koopman, "On Distributions Admitting a Sufficient Statistic," Transactions of the American Mathematical Society, 39, 1936, pp. 399-409.

REFERENCES (Cont.)

34. L. LeCam, "On Some Asymptotic Properties of the Maximum Likelihood Estimate and the Related Bayes Estimate," University of California Publication in Statistics, 1, 1953, pp. 277-330.

35. E. L. Lehmann, Testing Statistical Hypotheses, John Wiley and Sons, Inc., New York, 1959.

36. B. Levin and Y. Shinakov, "Asymptotic Properties of Bayes Estimates of Parameters of a Signal Masked by Interference," IEEE Trans. on Information Theory, IT-18, No. 1, 1972, pp. 102-106.

37. L. A. Liporace, "Variance of Bayes Estimates," IEEE Trans. on Information Theory, IT-17, No. 6, 1971, pp. 665-669.

38. K. Metzger, Personal correspondence consisting of EE 834 notes presented at The University of Michigan on February 13, 1968.

39. D. Middleton and D. Van Meter, "Modern Statistical Approaches to Reception in Communication Theory," IRE Trans. on Information Theory, IT-4, 1954.

40. D. Middleton and R. Esposito, "Simultaneous Optimum Detection and Estimation of Signals in Noise," IEEE Trans. on Information Theory, IT-14, No. 3, 1968, pp. 434-444.

41. N. E. Nahi, "Optimal Recursive Estimation with Uncertain Observation," IEEE Trans. on Information Theory, IT-15, No. 4, 1969, pp. 457-462.

42. L. W. Nolte, Adaptive Realizations of Optimum Detectors for Synchronous and Sporadic Recurrent Signals in Noise, Cooley Electronics Laboratory Technical Report No. TR-163, The University of Michigan, Ann Arbor, March 1965.

43. W. W. Peterson, T. G. Birdsall, and W. C. Fox, "The Theory of Signal Detectability," IRE Trans. on Information Theory, IT-4, 1954.

REFERENCES (Cont.)

44. V. F. Pisarenko, "The Detection of a Random Signal on a Background of Noise," Radiotekhnika i Elektronika, 6, No. 4, 1961, pp. 514-528.

45. R. Price, "Optimum Detection of Random Signals in Noise, with Applications to Scatter-Multipath Communication, Part I," IRE Trans. on Information Theory, IT-2, 1956, pp. 125-135.

46. H. Raiffa and R. Schlaiffer, Applied Statistical Decision Theory, The M.I.T. Press, Cambridge, Massachusetts, 1961.

47. R. A. Roberts, Theory of Signal Detectability: Composite Deferred Decision Theory, Ph.D. Dissertation, The University of Michigan, Ann Arbor, 1965.

48. R. A. Roberts, "On the Detection of a Signal Known Except for Phase," IEEE Trans. on Information Theory, IT-11, 1965, pp. 76-82.

49. W. L. Root, "Singular Gaussian Measures in Detection Theory," Proceedings of the Symposium on Time Series Analysis (Brown University, 1962), M. Rosenblatt ed., J. Wiley and Sons, New York, 1963, pp. 292-315.

50. H. L. Royden, Real Analysis, The Macmillan Company, New York, 1968.

51. A. P. Sage and J. R. Melsa, Estimation Theory with Applications to Communications and Control, McGraw-Hill Book Company, New York, 1971.

52. L. J. Savage, The Foundations of Statistics, John Wiley and Sons, 1954.

53. L. J. Savage, Joint Statistics Seminar, The University of London, 1959, John Wiley and Sons, New York, 1962.

54. L. Scharf and D. Lytle, "Signal Detection in Gaussian Noise of Unknown Level: An Invariance Application," IEEE Trans. on Information Theory, IT-17, No. 4, 1971, pp. 409-411.

REFERENCES (Cont.)

55. S. M. Selby, ed., CRC Standard Mathematical Tables, The Chemical Rubber Company, Cleveland, 1970.

56. D. Slepian, "Some Comments on the Detection of Gaussian Signals in Gaussian Noise," IRE Trans. on Information Theory, 4, No. 2, 1958, pp. 65-68.

57. R. L. Spooner, "On the Detection of a Known Signal in a Non-Gaussian Noise Process," Journal of the Acoustical Society of America, 44, 1968, pp. 141-147.

58. R. L. Spooner, The Theory of Signal Detectability: Extension to the Double-Composite Hypothesis Situation, Cooley Electronics Laboratory Technical Report No. TR-192, The University of Michigan, Ann Arbor, April 1968.

59. J. Spragins, "A Note on the Iterative Application of Bayes' Rule," IEEE Trans. on Information Theory, IT-11, No. 4, 1965, pp. 544-549.

60. C. T. Striebel, "Densities for Stochastic Processes," Annals of Mathematical Statistics, 30, 1959, pp. 559-567.

61. H. L. Van Trees, Detection, Estimation, and Modulation Theory, Part I, John Wiley and Sons, Inc., New York, 1968.

62. E. Wong, Stochastic Processes in Information and Dynamical Systems, McGraw-Hill Book Company, New York, 1971.

63. A. M. Yaglom, "On the Equivalence and Perpendicularity of Two Gaussian Measures in Function Space," Proceedings of the Symposium on Time Series Analysis (Brown University, 1962), M. Rosenblatt ed., J. Wiley and Sons, New York, 1963, pp. 327-346.

64. L. Zadeh and J. Ragazzini, "Optimum Filters for the Detection of Signals in Noise," Proceedings of the IRE, 40, 1952, p. 1223.

DISTRIBUTION LIST

	No. of Copies
Office of Naval Research (Code 468) (Code 102-OS) (Code 480) Navy Department Washington, D.C. 20360	1 1 1
Director, Naval Research Laboratory Technical Information Division Washington, D.C. 20390	6
Director Office of Naval Research Branch Office 1030 East Green Street Pasadena, California 91101	1
Dr. Christopher V. Kimball Special Studies Group IAR/PGI Suite 4 9719 South Dixie Highway Miami, Florida 33156	1
Director Office of Naval Research Branch Office 495 Summer Street Boston, Massachusetts 02210	1
Office of Naval Research New York Area Office 207 West 24th Street New York, New York 10011	1
Director Office of Naval Research Branch Office 536 S. Clark Street Chicago, Illinois 60605	1
Director Naval Research Laboratory Attn: Library, Code 2029 (ONRL) Washington, D.C. 20390	8

DISTRIBUTION LIST (Cont.)

	No. of Copies
Commander Naval Ordnance Laboratory Acoustics Division White Oak, Silver Spring, Maryland 20907 Attn: Dr. Zaka Slawsky	1
Commanding Officer Naval Ship Research & Development Center Annapolis, Maryland 21401	1
Commander Naval Undersea Research & Development Center San Diego, California 92132 Attn: Dr. Dan Andrews Mr. Henry Aurand	2
Chief Scientist Navy Underwater Sound Reference Division P. O. Box 8337 Orlando, Florida 32800	1
Commanding Officer and Director Navy Underwater Systems Center Fort Trumbull New London, Connecticut 06321	1
Commander Naval Air Development Center Johnsville, Warminster, Pennsylvania 18974	1
Commanding Officer and Director Naval Ship Research and Development Center Washington, D. C. 20007	1
Superintendent Naval Postgraduate School Monterey, California 93940	1
Commanding Officer & Director Naval Ship Research & Development Center* Panama City, Florida 32402	1

*Formerly Mine Defense Lab.

DISTRIBUTION LIST (Cont.)

	No. of Copies
Naval Underwater Weapons Research & Engineering Station Newport, Rhode Island 02840	1
Superintendent Naval Academy Annapolis, Maryland 21401	1
Scientific and Technical Information Center 4301 Suitland Road Washington, D. C. 20390 Attn: Dr. T. Williams Mr. E. Bissett	2
Commander Naval Ordnance Systems Command Code ORD-03C Navy Department Washington, D. C. 20360	1
Commander Naval Ship Systems Command Code SHIPS 037 Navy Department Washington, D. C. 20360	1
Commander Naval Ship Systems Command Code SHIPS 00V1 Washington, D. C. 20360 Attn: CDR Bruce Gilchrist Mr. Carey D. Smith	2
Commander Naval Undersea Research & Development Center 3202 E. Foothill Boulevard Pasadena, California 91107	1
Commanding Officer Fleet Numerical Weather Facility Monterey, California 93940	1

DISTRIBUTION LIST (Cont.)

	No. of Copies
Defense Documentation Center Cameron Station Alexandria, Virginia 22314	5
Dr. James Probus Office of the Assistant Secretary of the Navy (R&D) Room 4E741, The Pentagon Washington, D. C. 20350	1
Mr. Allan D. Simon Office of the Secretary of Defense DDR&E Room 3E1040, The Pentagon Washington, D. C. 20301	1
Capt. J. Kelly Naval Electronics Systems Command Code EPO-3 Washington, D. C. 20360	1
Chief of Naval Operations Room 5B718, The Pentagon Washington, D. C. 20350 Attn: Mr. Benjamin Rosenberg	1
Chief of Naval Operations Rm 4C559, The Pentagon Washington, D. C. 20350 Attn: CDR J. M. Van Metre	1
Chief of Naval Operations 801 No. Randolph St. Arlington, Virginia 22203	1
Dr. Melvin J. Jacobson Rensselaer Polytechnic Institute Troy, New York 12181	1
Dr. Charles Stutt General Electric Co. P. O. Box 1088 Schenectady, New York 12301	1

DISTRIBUTION LIST (Cont.)

	No. of Copies
Dr. Alan Winder EDO Corporation College Point, New York 11356	1
Dr. T. G. Birdsall Cooley Electronics Laboratory The University of Michigan Ann Arbor, Michigan 48105	1
Mr. Morton Kronengold Director, Institute for Acoustical Research 615 S.W. 2nd Avenue Miami, Florida 33130	1
Mr. Robert Cunningham Bendix Corporation 11600 Sherman Way North Hollywood, California 91606	1
Dr. H. S. Hayre University of Houston Cullen Boulevard Houston, Texas 77004	1
Mr. Ray Veenkant Texas Instruments, Inc. North Central Expressway Dalla, Texas 75222 Mail Station 208	1
Dr. Stephen Wolff John Hopkins University Baltimore, Maryland 21218	1
Dr. Bruce P. Bogert Bell Telephone Laboratories Whippany Road Whippany, New Jersey 07981	1
Dr. Albert Nuttall Navy Underwater Systems Center Fort Trumbull New London, Connecticut 06320	1

DISTRIBUTION LIST (Cont.)

	No. of Copies
Dr. Philip Stocklin Raytheon Company P. O. Box 360 Newport, Rhode Island 02841	1
Dr. H. W. Marsh Navy Underwater Systems Center Fort Trumbull New London, Connecticut 06320	1
Dr. David Middleton 35 Concord Ave., Apt. #1 Cambridge, Massachusetts 02138	1
Mr. Richard Vesper Perkin-Elmer Corporation Electro-Optical Division Norwalk, Connecticut 06852	1
Dr. Donald W. Tufts University of Rhode Island Kingston, Rhode Island 02881	1
Dr. Loren W. Nolte Dept. of Electrical Engineering Duke University Durham, North Carolina 27706	1
Dr. Thomas W. Ellis Texas Instruments, Inc. 13500 North Central Expressway Dallas, Texas 75231	1
Mr. Robert Swarts Honeywell, Inc. Marine Systems Center 5303 Shilshole Ave., N.W. Seattle, Washington, 98107	1
Mr. Charles Loda Institute for Defense Analyses 400 Army-Navy Drive Arlington, Virginia 22202	1

DISTRIBUTION LIST (Cont.)

	No. of Copies
Mr. Beaumont Buck General Motors Corporation Defense Research Division 6767 Holister Ave. Goleta, California 93017	1
Dr. M. Weinstein Underwater Systems, Inc. 8121 Georgia Avenue Silver Spring, Maryland 20910	1
Dr. Harold Saxton 1601 Research Blvd. TRACOR, Inc. Rockville, Maryland 20850	1
Dr. Thomas G. Kincaid General Electric Company P. O. Box 1088 Schenectady, New York 12305	1
Applied Research Laboratories The University of Texas at Austin Austin, Texas 78712 Attn: Dr. Loyd Hampton Dr. Charles Wood	3
Dr. Paul McElroy Woods Hole Oceanographic Institution Woods Hole, Massachusetts 02543	1
Dr. John Bouyoucos Hydroacoustics, Inc. P. O. Box 3818 Rochester, New York 14610	1
Dr. Joseph Lapointe Systems Control, Inc. 260 Sheridan Avenue Palo Alto, Calif. 94306	1
Cooley Electronics Laboratory University of Michigan Ann Arbor, Michigan 48105	25

DISTRIBUTION LIST (Cont.)

	No. of Copies
Professor Richard A. Roberts Dept. of Electrical Engineering University of Colorado Boulder, Colorado 80302	1
CAPT. Jurgen O. Gobien AFIT - ENE Air Force Inst. of Technology Wright-Patterson AFB, Ohio 45433	1