AMRL-TR-67-230

UNSUPERVISED SEQUENTIAL CLASSIFICATION OF NONSTATIONARY TIME SERIES

THOMAS J. HARLEY JR.

FOREWORD

This report was prepared at the Philco-Ford Corporation, Communications and Electronics Division, Blue Bell, Pennsylvania 19422, under Contract AF 33(615)-2966 for the Aerospace Medical Research Laboratories, Aerospace Medical Division, Air Force Systems Command, Wright-Patterson Air Force Base, Ohio. The research was performed in support of Project 7233, "Biological Information Handling Systems and Their Functional Analogs," Task 723305, "Theory of Information Handling." The technical contract monitor was Hans L. Oestreicher, Ph.D., Chief, Mathematics and Analysis Branch, Biodynamics and Bionics Division, Biomedical Laboratory.

The author is indebted to his colleagues Dr. Laveen N. Kanal and Dr. Kenneth Abend, for their guidance, patience, and helpful criticism. Much of the introduction to this paper is condensed from Dr. Abend's Ph.D. dissertation.

This technical report has been reviewed and is approved.

WAYNE H. McCANDLESS
Technical Director
Biomedical Laboratory
Aerospace Medical Research Laboratories

ABSTRACT

The problem of unsupervised sequential classification of nonstationary time series is formulated as a compound decision problem. The a priori class probabilities are assumed to be stochastically independent, time varying, and unknown. The class-conditional cumulative distribution functions of the random variable, X, are assumed to be of known parametric form, but with the parameter values unknown and time varying. A Bayesian approach is taken, employing an a priori distribution on the unknown parameters and class probabilities, which leads to a solution in terms of minimizing the sample conditional risk. If the unknown parameters and class probabilities are assumed to have Markov time dependence, then the nonstationary problem can be reformulated in terms of the problem of classifying stationary time series with known parameters and with known Markov dependence on the states of nature. Specific results are presented for two special cases - unknown, time varying a priori class probabilities, and unknown time varying mean.

TABLE OF CONTENTS

SECTION I

INTRODUCTION

COMPOUND DECISION THEORY

Statistical Decision Theory is the classical sense of Wald (1), and Blackwell and Girshick (2), deals with the problem of decision making in a tidy formal way. We have a set Ω of states of nature, a set A of possible actions, a loss function L defined on $\Omega \times A$, and a random variable, X, which has a known cumulative distribution function (c.d.f.), $P_\omega(x)$, conditional on each possible state of nature $\omega \in \Omega$. There also exists an a priori probability distribution, $G(\omega)$, on the states of nature, ω. We may know $G(\omega)$, in which case we can use a Bayes decision rule against $G(\omega)$ which minimizes our average loss. Alternatively, we may not know $G(\omega)$, in which case we will usually use a minimax rule, which minimizes our maximum loss. Thus given a decision rule, everything is clear-cut and completely determined.

Unfortunately, pattern recognition problems don't behave the way classical statistical decision theory asks them to - the c.d.f.'s are not known, at least not completely and the a priori probability functions on Ω are also unknown. In their place, we have available to us a collection, or a source, of patterns from the universe of interest. We may know the true state of nature of each of these patterns, or we may

have unreliable information on their true states of nature, or the patterns may be completely unclassified. Based on an analysis of the available samples, together with whatever knowledge is available about the c.d.f.'s, we must develop a decision rule to classify subsequent patterns. The formal basis for developing such rules is provided by compound decision theory.

In 1951, Herbert Robbins published a paper entitled "Asymptotically Subminimax Solutions of Compound Statistical Decision Problems" (3). This was the first of many papers in the statistical journals treating the compound decision problem (4 to 17).

However, none of this work appeared in the engineering literature until September 1965 when John Van Ryzin presented a paper at the Fourth Prague Conference on Information Theory (16). Since then, the application of these techniques to pattern recognition has been studied extensively by Abend (18), and by Professor Cover (19) and his students at Stanford University, among others.

THE GENERAL COMPOUND DECISION PROBLEM

In a compound decision problem, we deal with a sequence, $\underline{\theta}_n = \{\theta_1, \theta_2, \ldots, \theta_n\}$ of elements from a set $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$ of states of nature (pattern classes). Each possible sequence, $\underline{\theta}_n$, has an a priori

probability of occurrence, $G(\underline{\theta}_n)$. Also, associated with each possible sequence is a cumulative distribution function (c.d.f.), $P(\underline{x}_n|\underline{\theta}_n)$ on a sequence of (vector valued) random variables, $\underline{X}_n = \{X_1, X_2, \ldots, X_n\}$ *. A single observed value, x_k, of an element, X_k, of the sequence $\underline{X}_n$ is called a pattern. There is a space of possible actions, $A = \{\alpha_1, \alpha_2, \ldots, \alpha_s\}$ and a loss function defined on $\Omega \times A$. For the problem of pattern classification, we assume that $A \equiv \Omega$. We must choose a sequence of actions $\underline{a}_n = \{a_1, a_2, \ldots, a_n\}$. The loss function is a matrix $\{L_{ij}(k)\} = \{L_k(\omega_i, \alpha_j)\}$ which specifies the loss for the k^{th} decision if we choose action $a_k = \alpha_j$ when the true state of nature is $\theta_k = \omega_i$.

A randomized decision function $t(\alpha_j|x)$ is, for each x, a distribution over A, i.e., $t(\alpha_j|x)$ is the probability with which action α_j is selected when x is observed. If, for every x, $t(\alpha_j|x) = 1$ for one particular action $\alpha_j = \alpha_j(x)$, t is said to be non-randomized.

If all observations, $\underline{x}_n$, are at hand before the individual decisions must be made, one can use a <u>compound decision rule</u> $\underline{t}_n = (t_1, \ldots, t_n)$ where $t_k = t_k(\alpha_j|\underline{x}_n)$. If only the observations $\underline{x}_{k+r}$, $r \geq 0$, are at hand when the k^{th} decision must be made, one can use a <u>sequential compound decision rule</u>, where $t_k = t_k(\alpha_j|\underline{x}_{k+r})$. If $r \neq 0$, then the decision is said to be deferred r units of time. A <u>simple rule</u> is one where $t_k = t_k(\alpha_j|x_k)$, that is, one where the decision about θ_k depends on x_k alone. For a <u>simple symmetric</u> rule, $t_k = t(\alpha_j|x_k)$ for all k. Classical decision theory is restricted to simple symmetric rules.

---

* The elements X_k of the sequence X_n will be treated as column vectors.

In the discussion to follow, we will treat the classification problem, i.e., the case where $A \equiv \Omega$, and decision rules will be functions $t_k(\omega_j | \underline{x}_n)$.

The risk for the compound rule $\underline{t}_n$ is

$$R(\underline{\theta}_n, \underline{t}_n) = \frac{1}{n} \sum_{k=1}^{n} R(\underline{\theta}_n, t_k)$$

where

$$R(\underline{\theta}_n, t_k) = \int_{S^n} \sum_{j=1}^{n} L(\theta_k, \omega_j)\, t_k(\omega_j | \underline{x}_n)\, p(\underline{x}_n | \underline{\theta}_n)\, dx^n$$

is the risk for the k^{th} problem (the k^{th} component risk). The integration is over the n-fold Cartesian product of the measurement space.

One can also talk about a compound Bayes' risk $\overline{R}(G, \underline{t}_n)$ with respect to an a priori distribution $G(\underline{\theta}_n)$ over Ω^n, the n-fold Cartesian product of Ω. The compound Bayes risk is

$$\overline{R}(G, \underline{t}_n) = \sum_{\underline{\theta}_n \in \Omega^n} R(\underline{\theta}_n, \underline{t}_n)\, G(\underline{\theta}_n) = \frac{1}{n} \sum_{k=1}^{n} \overline{R}(G, t_k)$$

where

$$\overline{R}(G, t_k) = \sum_{\underline{\theta}_n \in \Omega^n} R(\underline{\theta}_n, t_k)\, G(\underline{\theta}_n)$$

is the k^{th} component risk. A procedure is compound Bayes against G if it minimizes $\overline{R}(G, \underline{t}_n)$. Thus, the compound Bayes' procedure $\underline{t}_n^G$ is

one that for every k minimizes

$$\bar{R}(G, t_k) = \int \sum_{j=1}^{M} \sum_{\underline{\theta}_n} L(\theta_k, \omega_j)\, t_k(\omega_j|\underline{x}_n)\, p(\underline{x}_n|\underline{\theta}_n)\, G(\underline{\theta}_n)\, dx^n .$$

Hence, $t_k{}^G(\omega_j|x_n) = 1$ for that ω_j which minimizes the quantity

$$\sum_{\underline{\theta}_n} L(\theta_k, \omega_j)\, p(\underline{x}_n, \underline{\theta}_n) = \sum_{\theta_k} L(\theta_k, \omega_j)\, p(\underline{x}_n, \theta_k)$$

where $p(\underline{x}_n, \underline{\theta}_n) = p(\underline{x}_n|\underline{\theta}_n)\, G(\underline{\theta}_n)$. For the special case of 0, 1 loss matrix $t_k{}^G$ chooses the value of θ_k that maximizes $p(\underline{x}_n, \theta_k)$. This is equivalent to maximizing the a posteriori probability

$$G(\theta_k|\underline{x}_n) = \frac{p(\underline{x}_n, \theta_k)}{p(\underline{x}_n)}$$

since the denominator

$$p(\underline{x}_n) = \sum_{\underline{\theta}_n} p(\underline{x}_n|\underline{\theta}_n)\, G(\underline{\theta}_n)$$

is independent of θ_k. In most of the early work in compound decision theory, and in much of the contemporary work, the assumption is made that the patterns are conditionally independent and that the classes also are independent,

i.e.,

$$p(\underline{x}_n | \underline{\theta}_n) = \prod_{k=1}^{n} p_k(x_k | \theta_k)$$

and

$$G(\underline{\theta}_n) = \prod_{k=1}^{n} G_k(\theta_k) .$$

Under these assumptions, a simple rule will be compound Bayes against G. A simple symmetric rule will be compound Bayes against G when $G_k(\theta_k) = G(\theta_k)$ and $p_k(x_k | \theta_k) = p(x_k | \theta_k)$, i.e., the probabilities are independent of k.

REVIEW OF COMPOUND DECISION PROCEDURES

The problem most treated in the early literature on compound decision theory is the symmetric case where $p(x_k | \theta_k)$ is known for each $\theta_k \in \Omega$ but where $G(\theta_k)$ is now known. There are three different approaches taken to this problem.

In the empirical Bayes approach we assume the existence of a $G(\theta)$ for each θ in Ω, but we do not know it. Robbins (7, 8, 9) shows that, if we use a procedure which is Bayes against a consistent estimate of $G(\theta)$ based on the first (n-1) observations, then the n^{th} component Bayes risk $\bar{R}(G, t_n)$ converges to the risk for a simple Bayes decision

against known $G(\theta)$. The non-Bayesian compound approach is to consider the compound risk $R(\underline{\theta}_n, \underline{t}_n)$ without assuming even the existence of an a priori probability function $G(\theta)$. Suppose that $q^n(\omega_i)$ is the true fraction of the θ's in $\underline{\theta}_n$ which are equal to ω_i. This is called the empirical a priori probability distribution of Ω. Hannan and Robbins (4, 9) have shown that if we use a procedure that is Bayes against a consistent estimate of q^n, then the compound risk $R(\underline{\theta}_n, \underline{t}_n)$ converges uniformly for any sequence $\underline{\theta}_n$, to the risk for a simple Bayes decision against known q^n. These two approaches - empirical Bayes and non-Bayesian compound - differ in fundamental philosophy. The non-Bayesians are unwilling to grant the existence of an a priori $G(\theta)$, while the empirical Bayesians assume the existence of a $G(\theta)$ of unknown value. Note, however, that if $G(\theta)$ exists, then the empirical a priori $q^n(\theta)$ is itself a consistent estimate of $G(\theta)$. Therefore, a consistent estimate of $q^n(\theta)$ is also a consistent estimate of $G(\theta)$ and vice versa. Therefore, one finds that although the philosophies differ, the estimation techniques and decision rules employed are the same.

The modern Bayesians (20) go further. They assume not only the existence of an a priori $G(\theta)$, which is of unknown value, but theyalso assume an a priori distribution function $F(G)$ on the values of $G(\theta)$. As

shown by Abend (18) this leads to a decision rule which is Bayes against an a posteriori estimate of $G(\theta)$ based on the observations. The same limit theorems hold, of course, because the Bayes estimate of $G(\theta)$ is consistent. Thus, for the case where $G(\theta)$ is unknown, the three techniques give the same basic results, except that any consistent estimator may be used in the first two approaches. However, in the problems of interest to us, we are concerned with more than just unknown $G(\theta)$'s. We are interested in cases where the states of nature are not independent. We are interested in cases where the conditional probabilities of the patterns are unknown. We are interested in cases where the patterns are not conditionally independent. And, most importantly, we are interested in cases where the statistics of the patterns and of the classes vary with time—the non-stationary case. The framework of compound decision theory is still appropriate, but we must extend the techniques described above, and we must be willing to adopt whatever philosophy promises to yield results for the problem at hand.

BAYESIAN TECHNIQUES

Bayesian techniques for the solution of problems in pattern recognition appeared in the engineering literature in 1962 (21). These efforts dealt with the problem of sequential learning with a teacher. Suppose that the conditional c.d.f. of X for a class ω_i is of known

parametric form, but the specific parameter values are unknown. Suppose, also, that we have observed a sequence of N independent observations $\underline{x}_n = \{x_1, \ldots, x_n\}$ from class ω_i. Let the vector B_{i_0} be the true values of the parameters. Let B_i be a random variable corresponding to B_{i_0}. The Bayesian approach is to assume an a priori density function $f(B_i)$ on B_i, and then, based on the sample $\underline{x}_n$, compute from it an a posteriori density function $f(B_i|\underline{x}_n)$. This can be computed in closed form as

$$f(B_i|\underline{x}_n) = \frac{p_i(\underline{x}_n|B_i)\ f(B_i)}{p_i(\underline{x}_n)} = \frac{\prod_{k=1}^{n} p_i(x_k|B_i)\ f(B_i)}{p_i(\underline{x}_n)}$$

where

$$p_i(\underline{x}_n) = \int p_i(\underline{x}_n|B_i)\ f(B_i)\ dB_i \ .$$

Alternatively, it can be computed in an iterative form

$$f(B_i|\underline{x}_n) = \frac{p_i(x_n|B_i)\ f(B_i|\underline{x}_{n-1})}{p_i(x_n|\underline{x}_{n-1})}$$

where

$$p_i(x_n|\underline{x}_{n-1}) = \int p_i(x_n|B_i)\ f(B_i|\underline{x}_{n-1})\ dB_i \ .$$

For many cases of interest, there exists a functional form for the a priori density, $f(B_i)$ such that the a posteriori density $f(B_i|x_n)$ is of the same functional form. Such density functions are called reproducing. Spragins (22) shows that a reproducing density for a parameter B_i exists if and only if the observations $\underline{x}_n$ admit a sufficient statistic expressible as a vector of r dimensions, i.e., the information about B_i contained in a sample $\underline{x}_n$ can be expressed in r statistics for $n > r$, for some fixed value of r. Spragins includes a table which indicates the appropriate reproducing density to use with various common parametric families of c.d.f.'s For these families, an optimum classification device for learning with a teacher can be constructed using a memory of fixed finite size.

More recently, Bayesian techniques have been applied to the problem of learning without a teacher. This approach, and the circumstances where it is applicable, are presented by Patrick and Hancock (23). Assume that the patterns are conditionally independent. Rather than observing a sample from a single class, one observes a sample $\underline{x}_n$ from a mixture of M classes. The c.d.f. for a single independent observation x is a mixture c.d.f. given by

$$p(x \mid B_0) = \sum_{i=1}^{M} Q_{i_0} P_i(x \mid B_{i_0})$$

where

$$Q_{i\,0} = G(\omega_i)$$

and

$$B_0 = \{B_{1_0}, \ldots, B_{M_0}, Q_{1_0}, \ldots, Q_{M_0}\} .$$

The approach is identical to that used in Bayesian learning with a teacher. Let B be a random variable corresponding to B_0. Assume an a priori density f(B) and from the sample $\underline{x}_n$ compute an a posteriori density $f(B|\underline{x}_n)$. The equations follow in the same way as before. Unfortunately, it has been shown that, for the common mixture distributions, no sufficient statistic of finite size exists. This has two unhappy consequences, — no reproducing density exists, and the required memory size for an optimum classifier increases linearly with n. Spragins (24) reviews some of the ways of dealing with this situation under fixed memory size constraints.

Although nothing has appeared in the engineering literature to that effect, Bayesian techniques can also be used in the case of learning with a teacher followed by learning without a teacher. First compute the a posteriori densities for the various B_i's and Q_i's based on learning with a teacher. Then, use the product of these densities as the a priori density of B for subsequent learning without a teacher.

In this discussion of Bayesian techniques for learning, we have discussed only the computation of a posteriori densities for the unknown parameter values. There are two different ways these can be used in a classification problem. First, one can use them to compute Bayes' estimates of the unknown parameter values, i.e.,

$$\hat{B}_i(\underline{x}_n) = \int B_i \, f(B_i|\underline{x}_n) \, dB_i$$

and use these estimates in place of the true parameter values in computing the likelihoods for a Bayes decision rule. This is the approach used by Abramson and Braverman (21). Alternatively, one can compute a sample-conditional, class-conditional density of X and θ for each $\theta \epsilon \Omega$, i.e.,

$$p(x\omega_i|\underline{x}_n) = \int p(x|B_i) \, \Omega_i f(B|\underline{x}_n) \, dB$$

and use these in a Bayes decision rule. This is the approach of Patrick and Hancock (23). In the special case where the class-conditional c.d.f.'s are known, but the class probabilities $G(\omega_i)$ are unknown, these two techniques turn out to be equivalent because, then

$$p(x\omega_i|\underline{x}_n) = p(x|\omega_i) \int \Omega_i \, f(\Omega_i|\underline{x}_n) \, d\Omega_i = p(x|\omega_i) \, \hat{\Omega}_i(\underline{x}_n) \quad .$$

For Gaussian distribution with unknown covariance matrices and mean

vectors, it has been shown by Keehn (25) for the learning with a teacher case, that the sample-conditional, class-conditional density of X is a (non-Gaussian) function of the Bayes estimates, and because a sufficient statistic exists for computing the estimates, there is no need to perform an integration.

Just about all the engineering literature on Bayesian techniques for learning with or without a teacher treats the stationary case where the statistics do not vary with time. One exception is the early paper of Abramson and Braverman (21) which treats the Gaussian case with a time varying mean which acts as a Martingale. Their work deals with the case of sequential learning with a teacher, that is they assume that after each decision is made, they are told whether or not it was correct. They compute an iterative Bayesian estimate for the mean at time K.

SECTION II

CLASSIFICATION OF NONSTATIONARY TIME SERIES

CLASSIFICATION OF NONSTATIONARY TIME SERIES- A COMPOUND DECISION PROBLEM

A nonstationary time series is, by definition, a time series which is not wide sense stationary, i.e., a time series is nonstationary if either the mean $\bar{x}_t$ or the autocovariance function $R(t,\tau) = \mathcal{E}\{(x_t - \bar{x}_t)(x_{t-\gamma} - \bar{x}_{t-\gamma})\}$ depends on t. In the cases of interest, we are dealing with sequences of vector random variables $\{X_t;\ t = 1, 2, \ldots, n\}$, and therefore the mean $\bar{x}_t$ is a vector and the autocovariance function $R(t, \gamma)$ is a matrix of covariance functions given by

$$\left[R(t, \gamma)\right] = \mathcal{E}\{(x_t - \bar{x}_t) \cdot (x_{t-\gamma} - \bar{x}_{t-\gamma})^T\}$$

where the diagonal elements of $R(t, \gamma)$ are the autocovariance functions of the elements of the vector x_t, and the off-diagonal elements are the cross-covariance functions between terms of the vector.

The general statement of the compound decision problem does not impose any assumptions dealing with stationarity and neither does the expression for the optimum decision rule for the k^{th} element of the sequence.

Thus, classification of nonstationary time series fits within "compound decision theory", and the general classification problem for nonstationary time series is identical to the general compound decision problem.

Compound decision theory can be used to develop either compound rules where we make the k^{th} decision, for every k, after observing the entire sequence of patterns $\underline{x}_n$, or sequential compound rules when we must make each k^{th} decision on the basis of the subsequence $\underline{x}_{k+r}$ for a fixed r. In our treatment of the nonstationary case, we will restrict our attention to the sequential compound case because it is the most common situation. In this case, the solution for the k^{th} decision is the same as that for the general compound decision problem, with n = k+r. Here, however, n will not be fixed throughout the sequence of problems, and the risk of the individual decisions will tend to decrease with increasing k. For simplicity, this paper will consider only the case r = 0.

For r = 0, i.e., for n = k, the risk for the k^{th} problem is given by

$$\bar{R}(t_k) = \int \sum_{j=1}^{m} \sum_{\theta_k \in \Omega} t_k(\omega_j | \underline{x}_k)\, L(\theta_k, \omega_j)\, p(\underline{x}_k\, \theta_k)\, d\underline{x}_k \quad .$$

Since $p(\underline{x}_k \theta_k) = p(x_k \theta_k | \underline{x}_{k-1})\ p(\underline{x}_{k-1})$, $\bar{R}(t_k)$ can be rewritten as

$$\bar{R}(t_k) = \int \bar{R}(t_k | \underline{x}_{k-1})\ p(\underline{x}_{k-1})\ d\underline{x}_{k-1}$$

where

$$\bar{R}(t_k | \underline{x}_{k-1}) = \int \sum_{j=1}^{M} \sum_{\theta_k \in \Omega} t_k(\omega_j | \underline{x}_k)\ L(\theta_k, \omega_j)\ p(x_k\ \theta_k | \underline{x}_{k-1})\ dx_k$$

is the sample conditional risk. A rule which minimizes $\bar{R}(t_k | \underline{x}_{k-1})$ will also minimize $\bar{R}(t_k)$. Let $t_k(\omega_j | \underline{x}_k) = 1$ for that ω_j such that

$$\sum_{\theta_k \in \Omega} L(\theta_k, \omega_j)\ p(x_k \theta_k | \underline{x}_{k-1}) = \inf_{\omega_i \in \Omega} \sum_{\theta_k \in \Omega} L(\theta_k, \omega_j)\ \pi(x_k \theta_k | \underline{x}_{k-1}),$$

and $t_k(\omega_i | \underline{x}_k) = 0$ for $\omega_i \neq \omega_j$, i.e., choose action $a_k = \omega_j$. To do this, we must be able to compute $p(x_k \theta_k | \underline{x}_{k-1})$ or $p(\underline{x}_k \theta_k)$.

THE GENERAL NONSTATIONARY PROBLEM - A BAYESIAN APPROACH

Assume that, at each time k, each class conditional c.d.f., $p(x_k |\ \omega_i\ B_{i0}(k))$ is dependent on a set of parameters, $B_{i0}(k)$, whose true values are not known. We will assume that the patterns in the sequence are class conditionally stochastically independent, i.e.,

$$p(\underline{x}_n | \underline{\theta}_n\ \underline{B}(n)) = \prod_{k=1}^{n} p(x_k | \theta_k\ B(k)).$$

We will also assume that the states of nature are stochastically independent, i.e.,

$$G(\underline{\theta}_n) = \prod_{k=1}^{n} G(\theta_k) \quad .$$

From these, it can be shown that the patterns are also parameter conditionally independent, i.e.,

$$p(\underline{x}_n | \underline{B}(n)) = \prod_{k=1}^{n} p(x_k | B(k)) \quad .$$

Let the true value of $G_k(\omega_i), i = 1, 2, \ldots, M$ be $Q_{i_0}(k)$, also known.

Define $B_0(k) = \left\{B_{1_0}(k), \ldots, B_{M_0}(k), Q_{1_0}(k), \ldots, Q_{M_0}(k)\right\}$ *

and $\underline{B}_0(n) = \left\{B_0(1), B_0(2), \ldots, B_0(k)\right\}$.

Finally, since the various parameters may be unknown, let $\underline{B}(n)$, $B(k)$, $B_i(k)$, and $Q_i(k)$ be random variables corresponding to $\underline{B}_0(n)$, $B_0(k)$, $B_{i_0}(k)$, and $Q_{i_0}(k)$ respectively. In using a Bayesian approach we assume an a priori c.d.f., $F(\underline{B}(n))$ on the unknown parameters, with $f(\underline{B}(n))$ being the corresponding density function. The minimum risk rule for the k^{th} decision given x_k requires computation of

---

* This notation is an extension of that used by Patrick and Hancock (23).

$p(\underline{x}_k\ \theta_k)$ for each $\theta_k = \omega_1, \omega_2, \ldots, \omega_M$. We have then

$$p(\underline{x}_k\ \omega_i) = \int P(\underline{x}_k\ \omega_i\ \underline{B}(n)\ )\ d\underline{B}(n)$$

$$= \int p(x_k|\underline{x}_{k-1}\ \omega_i\ \ \underline{B}(n)\ )\ p(\omega_i|\underline{x}_{k-1}\ \underline{B}(n)\ )\ p(\underline{x}_{k-1}|\underline{B}(n)\ )\ f(\underline{B}(n)\ )\ d\underline{B}(n),$$

$$\therefore \quad p(\underline{x}_k\ \omega_i) = \int p(x_k|\omega_i\ B_i(k)\ )\ Q_i(k) \prod_{j=1}^{k-1} p(x_j|B(j)\ )\ f(\underline{B}(n)\ )\ d\underline{B}(n)\ .$$

This is a general (Bayesian) expression for the case of conditionally independent patterns. Alternatively, we can divide both sides of the equation by $p(\underline{x}_{k-1})$ and get

$$p(x_k\ \omega_i|\underline{x}_{k-1}) = \frac{\int p(x_k|\omega_i B_i(k)\ )\ Q_i(k) \prod_{j=1}^{k-1} p(x_j|B(j))\ f(\underline{B}(n)\ )\ d\underline{B}(n)}{\int \prod_{i=1}^{k-1} p(x_j|B(j)\ )\ f(\underline{B}(n)\ )\ d\underline{B}(n)}\ .$$

This could also have been derived in the form

$$p(x_k\ \omega_i|\underline{x}_{k-1}) = \int p(x_k|\omega_i\ B_i(k)\ )\ Q_i(k)\ f(\underline{B}(n)|\underline{x}_{k-1})\ d\underline{B}(n)$$

where

$$f(\underline{B}(n)|x_{k-1}) = \frac{\prod_{j=1}^{k-1} p(x_j|B(j)\ )\ f(\underline{B}(n)\ )}{\int \prod_{j=1}^{k-1} p(x_j|\ B(j))\ f(\underline{B}(n)\ )\ d\underline{B}(n)}$$

This expression for $p(x_k\ \omega_i | \underline{x}_{k-1})$ is a direct extension to the nonstationary case of Equation (28) of Patrick and Hancock (23).

In the above equations then, we have the expressions necessary to state the minimum risk decision rule for nonstationary time series when some of the parameters are unknown. The general Bayesian solution for nonstationary time series also subsumes stationary timeseries as a special case.

MARKOV DEPENDENCY OF THE TIME-VARYING PARAMETERS

Suppose that the parameter B(k) can take on values in the finite space $\Gamma = \{\gamma_1, \gamma_2, \ldots, \gamma_g\}$, independent of k. Then, for $\theta_k = \omega_i$

$$p(\underline{x}_k, \omega_i) = \sum_{B \in \Gamma} p(\underline{x}_k\ \omega_i\ B(k)) = \sum_{j=1}^{g} p(\underline{x}_k\ \omega_i\ \gamma_j) \ .$$

Define the space $\Psi = \{\psi_{11}, \psi_{21}, \ldots, \psi_{Mg}\} = \Omega \times F$ where each element $\psi_{ij} \in \Psi$ consists of a pair $\{\omega_i\ \gamma_j\}$ where $\omega_i \in \Omega$ and $\gamma_j \in \Gamma$. We have defined $\underline{\theta}_n$ as a sequence of states of nature, and $\underline{B}(n)$ as a sequence of parameter values. Now, let $\underline{\Phi}(n) = \{\Phi(1), \Phi(2), \ldots, \Phi(n)\}$, where each $\Phi(k) \in \Psi$, be a sequence of generalized (parameter) states of nature. Then, for $\theta_k = \omega_i$, and $\Phi(k) = \psi_{ij}$, we get

$$p(\underline{x}_k\ \omega_i) = \sum_{j=1}^{g} p(\underline{x}_k \psi_{ij}) \ .$$

Assume that the parameter values, $\underline{B}(n)$, form a first order Markov chain, i.e.,

$$P(B(k)|\underline{B}(k-1)) = P(B(k)|B(k-1)) \quad \text{for all } k.$$

(Because Γ is a discrete space, we use the notation P designating a probability, rather than f designating a probability density function.)

Then

$$P(\Theta(k) \mid \underline{\Theta}(k-1) = P(\theta_k\ B(k)|\underline{\theta}_{k-1}\ \underline{B}_{(k-1)})$$

$$= P(\theta_k|B(k))\ P(B(k)|B(k-1))$$

and therefore

$$P(\Theta(k)|\underline{\Theta}(k-1)) = P(\Theta(k)|\Theta(k-1)),$$

i.e., the generalized states of nature $\underline{\Theta}(n)$ also form a Markov chain.

This leads to

$$p(\underline{x}_k\Theta(k)) = p(x_k|\Theta(k)\ \underline{x}_{k-1})\ p(\underline{x}_{k-1}\ \Theta(k))$$

$$= p(x_k|\Theta(k)) \sum_{\substack{\Theta(k-1) \\ \epsilon\ \Psi}} p(\underline{x}_{k-1}\Theta(k)\Theta(k-1))$$

and finally

$$p(\underline{x}_k\ \Theta(k)) = p(x_k|\Theta(k) \sum_{\substack{\Theta(k-1) \\ \epsilon\ \Psi}} P(\Theta(k)|\Theta(k-1))\ p(\underline{x}_{k-1}\ \Theta(k-1))$$

This result shows that the problem of unsupervised sequential classification of conditionally independent patterns, assuming independent

states of nature, when the unknown parameters form a Markov chain, is identical in structure to the problem of sequential classification of conditionally independent patterns, assuming Markov dependency on the states of nature, with all the parameters known. The solution to this equivalent problem has been presented by Abend (18). The result can be generalized to the case of an m-th order Markov chain.

SECTION III

SPECIFIC EXAMPLES

UNKNOWN, TIME VARYING A PRIORI PROBABILITIES

Suppose all the parameters were known except the a priori probability of each ω_i at each time k. Let us consider only the two-class case, i.e., M = 2, where $Q_1(k) = 1 - Q_2(k)$. Define a new random variable q(k) such that $Q_1(k) = 1/2 + q(k)$ and $Q_2(k) = 1/2 - q(k)$. Then $f(\underline{B}(n)) = f(\underline{q}(n))$ where $\underline{q}(n) = \{q(1)\ q(2), \ldots, q(n)\}$. We have then

$$p(x_k\ \omega_i|\underline{x}_{k-1}) =$$

$$\frac{\int p(x_k|\omega_1)(1/2+q(k))\prod_{j=1}^{k-1}\left[\left(p(x_j|\omega_1)+p(x_j|\omega_2)\right](1/2)+q(j)\left[p(x_j|\omega_1)-p(x_j|\omega_2)\right]\right] f(\underline{q}(n))\,d\underline{q}(n)}{\int \prod_{j=1}^{k-1}\left[\left[p(x_j|\omega_1)+p(x_j|\omega_2)\right](1/2)+q(j)\left[p(x_j|\omega_1)-p(x_j|\omega_2)\right]\right] f(\underline{q}(n))\,d\underline{q}(n)}$$

If we let

$$b_j = \frac{p(x_j|\omega_1) - p(x_j|\omega_2)}{p(x_j|\omega_1) + p(x_j|\omega_2)}$$

we get

$$p(x_k\ \omega_1|\underline{x}_{k-1}) = p(x_k|\omega_1)\left[(1/2) + \frac{N_k(\underline{x}_{k-1})}{D_k(\underline{x}_{k-1})}\right]$$

and

$$p(x_k\,\omega_2|\underline{x}_{k-1}) = p(x_k|\omega_2)\left[(1/2) - \frac{N_k(\underline{x}_{k-1})}{D_k(\underline{x}_{k-1})}\right]$$

where

$$N_k(\underline{x}_{k-1}) = \int q(k) \prod_{j=1}^{k-1} \left[1 + 2b_j\, q(j)\right] f(\underline{q}(n))\, d\underline{q}(n)$$

and

$$D_k(\underline{x}_{k-1}) = \int \prod_{j=1}^{k-1} \left[1 + 2b_j\, q(j)\right] f(\underline{q}(n))\, d\underline{q}(n)\,.$$

We can expand the product in the denominator to get

$$D_k(\underline{x}_{k-1}) = \int f(\underline{q}(n)) \left[1 + 2\sum_{j_1=1}^{k-1} b_{j_1} q(j_1) + 4\sum_{j_2>j_1=1}^{k-1} b_{j_1} b_{j_2} q(j_1)q(j_2) + \cdots \right.$$

$$\left. + 2^{k-1} b_1 b_2 \ldots b_{k-1}\, q(1)\ldots q(k-1)\right] d\underline{q}(n)$$

$$= 1 + 2\sum_{j_1=1}^{k-1} b_{j_1}\overline{q(j_1)} + \ldots + 2^{k-1} b_1 b_2 \ldots b_{k-1}\,\overline{q(1)\ldots q(k-1)}$$

where

$$\overline{q(j_1)\, q(j_2)\ldots q(j_m)} = \int f(\underline{q}(n))\, q(j_1)\, q(j_2)\ldots q(j_m)\, d\underline{q}(n)\,, \quad j_1<j_2<\ldots<j_m\,.$$

Similarly, we get

$$N_k(\underline{x}_{k-1}) = \overline{q}_k + 2 \sum_{j_1=1}^{k-1} b_{j_1} \overline{q(k)q(j_1)} + \ldots + 2^{k-1} b_1 b_2 \ldots b_{k-1} \overline{q(1)q(2)\ldots q(k)} .$$

For a 0, 1 loss matrix, i.e., if we wish to minimize the total number of errors, the resulting decision rule is, choose ω_1 if

$$\frac{p(x_k|\omega_1)}{p(x_k|\omega_2)} > \frac{D_k(\underline{x}_{k-1}) - 2 N_k(\underline{x}_{k-1})}{D_k(\underline{x}_{k-1}) + 2 N_k(\underline{x}_{k-1})}$$

and ω_2 otherwise. By appropriate manipulation, we can obtain the rule, choose ω_1 if

$$-b(k) < 2 \frac{N_k(\underline{x}_{k-1})}{D_k(\underline{x}_{k-1})} = 2 E(q(k)|\underline{x}_{k-1}) .$$

This is the general solution for the case where the a priori's vary with time. The solution is explicit, and in closed form. It requires computation of the expected values of the multiple products of the q(k)'s. In the general nonstationary case, the value of $\overline{q(j_1)q(j_2)\ldots q(j_m)}$ will be different for every sequence $j_1, j_2, \ldots, j_m$, and these will require considerable computation. For the special case of stationarity, we will show that, for a given m, the values will all be the same.

- Here we assume that q(k) is independent of k. If we assume a uniform a priori on q, i.e.,

$$f(q) = \begin{cases} 1 & \text{if } -\frac{1}{2} \le q \le \frac{1}{2} \\ 0 & \text{otherwise} \end{cases} .$$

Then

$$\overline{q(j_1)q(j_2)\ldots q(j_m)} = E(q^m) = \begin{cases} 0 & \text{for } m \text{ odd} \\ \left(\frac{1}{2}\right)^m \frac{1}{m-1} & \text{for } m \text{ even} , \end{cases}$$

This leads to

$$D_k(\underline{x}_{k-1}) = 1 + \frac{1}{3} \sum_{j_2 > j_1 = 1}^{k-1} b_{j_1} b_{j_2} + \ldots + \frac{1-(-1)^k}{2k} \; b_1 \ldots b_{k-1}$$

and

$$N_k(\underline{x}_{k-1}) = \frac{1}{2}\left[\frac{1}{3} \sum_{j_1 = 1}^{k-1} b_{j_1} + \ldots + \frac{1+(-1)^k}{2(k+1)} \; b_1 \ldots b_{k-1} \right] .$$

- Suppose that q(k) is independent of k and that we know that class ω_1 is more probable than class ω_2, i.e., $q \ge 0$. Then we can assume

$$f(q) = \begin{cases} 2 & \text{for } 0 \le q \le \frac{1}{2} \\ 0 & \text{otherwise} \end{cases}$$

for which

$$E(q^m) = \frac{1}{2^m(m+1)}$$

Then we get

$$D_k(\underline{x}_{k-1}) = 1 + \frac{1}{2}\sum b_{j_1} + \frac{1}{3}\sum b_{j_1}b_{j_2} + \ldots + \frac{b_1 b_2 \ldots b_{k-1}}{k}$$

and

$$N_k(\underline{x}_{k-1}) = \frac{1}{2}\left[\frac{1}{2} + \frac{1}{3}\sum b_{j_1} + \ldots + \frac{b_1 b_2 \ldots b_{k-1}}{k+1}\right].$$

TIME VARYING MEAN VALUE

Again we treat the two class case. Let each pattern x_k be scalar valued. Assume that q is known, and that the c.d.f. is normal with unit variance, and that the patterns and classes are independent. The mean for class ω_1 is zero. The mean for class ω_2 is m(k). Assume that m(k) varies randomly with time, k, about a mean value m_o, and that it has a normal c.d.f. and is constrained by a power spectrum $f(\omega)$. Then

$$f(\underline{m}(n)) = (2\pi)^{-n/2}\left|\sum\right|^{-1/2} \exp\left\{-\frac{1}{2}(\underline{m}(n)-\underline{m}_o)^T \sum^{-1}(\underline{m}(n)-\underline{m}_o)\right\}$$

where

$$\sum = \left[\sigma^2_{st}\right] \quad , \qquad s,t = 1,2,\ldots,n$$

and

$$\sigma^2_{s,t} = \sigma^2_{s-t} = \int_{-\pi}^{\pi} f(\omega)\, e^{j\omega(s-t)}\, d\omega \ .$$

Let $\underline{\omega}_i(n)$ be a sequence of n states of nature where each element is the state ω_i. Then

$$P(\underline{x}_n \,|\underline{\omega}_1(n)\,) = \prod_{i=1}^{n} (2\pi)^{-1/2} \exp\left\{-\frac{1}{2}\, x_i^2\right\} = \Phi_I(\underline{x}_n)$$

and

$$P(\underline{x}_n|\underline{\omega}_2(n), \underline{m}(n)\,) = \prod_{i=1}^{n} \left\{(2\pi)^{-1/2} \exp\, -\frac{1}{2}\,(x_i - m(i)\,)^2\right\} = \Phi_I(\underline{x}_n - \underline{m}(n)\,)$$

where I is the n x n identity matrix, i.e., the variances are unity, and the covariances are all zero.

We have been given $p(\underline{x}_n|\underline{\omega}_2\ \underline{m}(n)\,) = \Phi_I(\underline{x}_n - \underline{m}(n)\,)$ and we have assumed that $f(\underline{m}(n)\,) = \Phi_{\Sigma}(\underline{m}(n) - \underline{m}_o)$. Therefore, $(\underline{x}_n - \underline{m}_o)$ is the sum of two normally distributed random variables with zero mean, $(\underline{x}_n - \underline{m}(n)\,)$ and $(\underline{m}(n) - \underline{m}_o)$. The sum also is normally distributed, and the means and covariance matrices add. Therefore

$$P(\underline{x}_n|\underline{\omega}_2(n)\,) = \Phi_{I+\Sigma}\ (\underline{x}_n - \underline{m}_o) \ .$$

We already have

$$P(\underline{x}_n \, \underline{\omega}_1(n)) = \Phi_1(\underline{x}_n) \, .$$

The problem then is equivalent, for known m_o, to a two class problem where samples from class ω_1 are independent with zero mean and unit variance, while the samples from class ω_2 are drawn from a stationary time series which has mean m_o and an autocovariance function given by

$$R(\tau) = \begin{cases} 1 + \sigma_o^2 & \tau = 0 \\ \sigma^2 \, . & \tau = s-t \neq 0 \, , \; s,t = 1,2,\ldots,n. \end{cases}$$

The process corresponding to class ω_2 has a spectrum $g(\omega) = f(\omega) + 1/2\pi$.

SECTION IV

SUMMARY

We have treated the classification of nonstationary time series within the framework of compound decision theory. By using Bayesian techniques, i.e., by assuming an a priori distribution on the unknown parameters of the c.d.f.'s, we were able to develop a general solution to the classification problem.

In the case of Markov dependence of the unknown, time-varying parameters, the classification problem is equivalent in structure to a problem involving stationary time series with known parameters, and Markov dependence on the states of nature.

Detailed results were obtained for two examples of nonstationarity: time varying a priori class probabilities, and time varying class means. Results for stationary time series were derived as special cases of the problem of time varying a priori class probabilities. In the case of time varying means, the solution is equivalent to that for a problem involving conditionally dependent stationary patterns.

REFERENCES

1. Wald, A., Statistical Decision Functions, New York, John Wiley and Sons, 1950.

2. Blackwell, D., and Girshick, M.A., Theory of Games and Statistical Decisions, New York, John Wiley and Sons, 1954.

3. Robbins, H., "Asymptotically Subminimax Solutions of Compound Statistical Decision Problems," Proc. Second Berkley Symposium on Math. Statist. and Prob., University of California Press, pp. 157-163, 1956.

4. Hannan, J.F., and Robbins, H., "Asymptotic Solutions of the Compound Decision Problem for Two Completely Specified Distributions," Ann. Math. Statist., Vol. 26, No. 1, pp. 37-51, March 1955.

5. Hannan, J.F., and Van Ryzin, J.R., "Rate of Convergence in the Compound Decision Problem for Two Completely Specified Distributions," Ann. Math. Statist., Vol. 36, No. 6, pp. 1743-1752, December 1965.

6. Johns, M.V., "An Empirical Bayes Approach to Non-parametric Two-way Classification," Studies in Item Analysis and Prediction (ed., H. Solomon), Stanford University Press, pp. 221-232, 1961.

7. Robbins, H., "An Empirical Bayes Approach to Statistics," Proc. Third Berkley Symposium on Math. Statist. and Prob., University of California Press, pp. 157-163, 1956.

8. Robbins, H., "The Empirical Bayes Approach to Statistical Decision Problems," Ann. Math. Statist., Vol. 35, No. 1, pp. -120, March 1964.

9. Robbins, H., and Samuel, E., "Testing Statistical Hypothesis - the 'Compound' Approach," Recent Developments in Information and Decision Processes (ed., R.E. Machol and P. Gray), Macmillin, New York, pp. 63-70, 1962.

10. Samuel, E., "An Empirical Bayes Approach to the Testing of Certain Parametric Hypothesis," Ann. Math. Statist., Vol. 34, No. 4, pp. 1370-1385, December 1963.

11. Samuel, E., "On the Compound Decision Problem in the Non-sequential and the Sequential Case," Ph. D. Thesis, Columbia University, 1961.

12. Samuel, E., "Asymptotic Solutions of the Sequential Compound Decision Problem," Ann. Math. Statist., Vol. 34, No. 3, pp. 1079-1094, September 1963.

13. Samuel, E., "On Simple Rules for the Compound Decision Problem," Journal of the Royal Statistical Society, Series B, Vol. 27, No. 2, pp. 238-240, 1965.

14. Van Ryzin, J.R., "Asymptotic Solutions to Compound Decision Problems," Ph.D. Thesis, Michigan State University, February 1964.

15. Van Ryzin, J.R., "The Sequential Compound Decision Problem with m x n Finite Loss Matrix," Ann. Math. Statist., Vol. 37, No. 4, pp. 954-975, August 1966.

16. Van Ryzin, J.R., "Non-Parametric Bayesian Decision Procedures for (Pattern) Classification with Stochastic Learning," Fourth Prague Conference on Information Theory, Statistical Decision Functions, and Random Processes, September 1965.

17. Van Ryzin, J.R., "Repetitive Play in Finite Statistical Games with Unknown Distributions," Ann. Math. Statist., Vol. 37, No. 4, pp. 976-994, August 1966.

18. Abend, K., "Compound Decision Procedures for Pattern Recognition," Ph.D. Thesis, University of Pennsylvania, 1966.
A portion of this work appears under the same title in the Proceedings of the National Electronics Conference, Vol. XXIII, 1966, pp. 777-999.

19. Alens, N., and Cover, T.M., "Compound Bayes Learning Without a Teacher," Proceedings of the First Annual Princeton Conference on Information Sciences and Systems, 1967.

20. Savage, L. J., The Foundations of Statistics, New York, John Wiley and Sons, 1954.

21. Abramson, N., and Braverman, D., "Learning to recognize patterns in a random environment," IRE Trans. on Information Theory (Supplement), Vol. IT-8, pp. 58-63, September 1962.

22. Spragins, J. D., "A Note on Iterative Application of Bayes' Rule," IEEE Trans. on Information Theory, Vol. IT-11, pp. 544-549, October 1965.

23. Patrick, E. A. and Hancock, J. C., "Nonsupervised Sequential Classification and Recognition of Patterns," IEEE Trans. on Information Theory, Vol. IT-12, pp. 362-372, July 1966.

24. Spragins, J. D., "Learning Without a Teacher," IEEE Trans. on Information Theory, Vol. IT-12, pp. 223-230, April 1966.

25. Keehn, D. G., "A Note on Learning for Gaussian Properties," IEEE Trans. on Information Theory, Vol. IT-11, pp. 126-132, January 1965.

Security Classification

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author) Philco-Ford Corporation Communications and Electronics Division Blue Bell, Pennsylvania 19422	2a. REPORT SECURITY CLASSIFICATION UNCLASSIFIED 2b. GROUP N/A

3. REPORT TITLE

UNSUPERVISED SEQUENTIAL CLASSIFICATION OF NONSTATIONARY TIME SERIES

4. DESCRIPTIVE NOTES (Type of report and inclusive dates)

Final Report, 1 August 1965 - 30 April 1967

5. AUTHOR(S) (First name, middle initial, last name)

Thomas J. Harley, Jr.

6. REPORT DATE October 1968	7a. TOTAL NO. OF PAGES 33	7b. NO. OF REFS 25
8a. CONTRACT OR GRANT NO. AF 33(615)-2966 b. PROJECT NO. 7233 c. Task No. 723305 d.	9a. ORIGINATOR'S REPORT NUMBER(S) 9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report) AMRL-TR-67-230	

10. DISTRIBUTION STATEMENT

This document has been approved for public release and sale; its distribution is unlimited.

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY Aerospace Medical Research Laboratories Aerospace Medical Division, Air Force Systems Command, Wright-Patterson AFB, OH 45433

13. ABSTRACT

The problem of unsupervised sequential classification of nonstationary time series is formulated as a compound decision problem. The a priori class probabilities are assumed to be stochastically independent, time varying, and unknown. The class-conditional cumulative distribution functions of the random variable, X, are assumed to be of known parametric form, but with the parameter values unknown and time varying. A Bayesian approach is taken, employing an a priori distribution on the unknown parameters and class probabilities, which leads to a solution in terms of minimizing the sample conditional risk. If the unknown parameters and class probabilities are assumed to have Markov time dependence, then the nonstationary problem can be reformulated in terms of the problem of classifying stationary time series with known parameters and with known Markov dependence on the states-of-nature. Specific results are presented for two special cases - unknown, time varying a priori class probabilities, and unknown time varying mean.

DD FORM 1 NOV 65 1473

Security Classification

14. KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Statistics Bionics Pattern recognition						