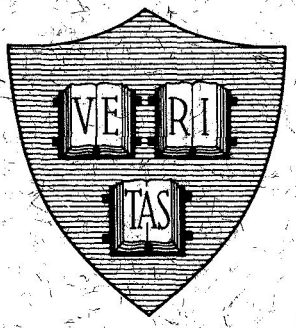
Office of Naval Research
Contract N00014-67-A-0298-0006 NR-372-012
NATIONAL AERONAUTICS AND SPACE ADMINISTRATION
Grant NGL 22-007-143

ON FEATURE REDUCTION WITH APPLICATION TO ELECTROENCEPHALOGRAMS



By

Karkal Pulkeri Sheshagiri Prabhu

September 1970

Technical Report No. 615

Division of Engineering and Applied Physics
Harvard Univerity - Cambridge, Massachusetts

Office of Naval Research

Contract N00014-67-A-0298-0006

NR-372-012

National Aeronautics and Space Administration

Grant NGL 22-007-143

ON FEATURE REDUCTION WITH APPLICATION TO ELECTROENCEPHALOGRAMS

By

Karkal Pulkeri Sheshagiri Prabhu

Technical Report No. 615

September 1970

The research reported in this document was made possible through support extended the Division of Engineering and Applied Physics, Harvard University by the U. S. Army Research Office, the U. S. Air Force Office of Scientific Research and the U. S. Office of Naval Research under the Joint Services Electronics Program by Contracts N00014-67-A-0298-0006, 0005, and 0008 and by the National Aeronautics and Space Administration under Grant NGL 22-007-143.

Division of Engineering and Applied Physics

Harvard University · Cambridge, Massachusetts

ON FEATURE REDUCTION WITH APPLICATION TO ELECTROENCEPHALOGRAMS

By

Karkal Pulkeri Sheshagiri Prabhu

Division of Engineering and Applied Physics

Harvard University · Cambridge, Massachusetts

ABSTRACT

The problem considered in this report is that of discriminating between two kinds of electroencephalogram (EEG) signals recorded from a human subject -- spontaneous EEG and EEG driven by photic stimuli at the alpha frequency of the subject. Since an EEG record represents a large amount of data, efficient feature reduction methods are required to pick out a few features which are significant for discrimination purposes.

The feature reduction methods available in the literature are first examined critically. A nonparametric feature reduction method based on a distance measure is developed, using the sampled values of the EEG as features. The computations involved in feature reduction also yield the best separating hyperplane at each stage. The error rate is less than five percent when the decisions are based on twenty periods of the alpha frequency.

A random process model is developed for the two kinds of EEG signals based on the fact that the EEG driven at the alpha frequency has more phase coherence than the spontaneous EEG. The model is then employed for feature reduction and pattern classification. The model

provides a four dimensional vector of sufficient statistics, which contains all the information necessary for discrimination purposes. The sufficient statistics are functions of the phase values of the EEG. They are in the form of cumulative sums which can be updated as more data becomes available. Moreover, the Bayes optimal separating surface is linear in terms of these sufficient statistics.

The error rates obtained by the two methods are compared. It is seen that in the 5% range of error rate, which is of practical interest, the two methods perform equally well. The computational simplicity of the model-based method gives it a decisive advantage.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER I

A SURVEY OF FEATURE REDUCTION

1.1 Role of Feature Reduction in Pattern Classification

The two class pattern classification problem can be formulated as follows. We are permitted to make N measurements on a given pattern and thus extract an N-dimensional pattern vector $\underline{x} = (x_1, x_2, \ldots, x_N)$ from the pattern. The components x_i of the pattern vector are called features. Depending on the numbers x_i the pattern has to be classified as coming from one of two pattern classes H^0 or H^1.

From a geometrical point of view, each pattern can be considered as a point in N-dimensional space. The classification problem is to find a separating surface which divides the N-dimensional space into two parts corresponding to pattern classes H^0 and H^1 respectively. Any given pattern is then classified according to the position of the point representing it in N-dimensional space.

From the analytical point of view, the classification problem reduces to finding a scalar function $f: R^N \rightarrow R^1$ such that the pattern $\underline{x}$ is classified as coming from H^0 or H^1 according as $f(\underline{x}) \gtrless 0$. $f(\underline{x}) = 0$ will then describe the separating surface.

In most cases of interest, it is found that the two pattern classes do overlap to some extent, and therefore

are not separable in N-dimensional space. In such cases the objective is to construct the separating surface (or the function f) in such a way that the number of misclassifications (or probability of error, in a statistical sense) is minimized.

There are several algorithms available in the literature to find a suitable function f which is optimal in some sense. Ho and Agrawala [1] have classified these algorithms depending on the nature of the information available about the two pattern classes. This information may be knowledge about the statistical distributions of the two pattern classes or learning patterns (also called training samples) of known or unknown classification.

This thesis does not seek to develop any new algorithms for pattern classification. Rather, the purpose is to deal with the problem of feature reduction (also called dimensionality reduction in the literature) which arises from the following considerations. The computational complexity involved in finding a suitable separating function f increases rather rapidly as the dimension N of the pattern vector goes up. This is true regardless of what information is available about the pattern classes or which particular algorithm is used to arrive at the function f. Therefore, it is desirable to transform the pattern vector $\underline{x} = (x_1, x_2, \ldots, x_N)$ into another vector $\underline{y} = (y_1, y_2, \ldots, y_m)$ of considerably lower dimension, and then apply the classification

algorithms to the transformed pattern vector $\underline{y}$.

If the transformation is denoted by $\underline{y} = g(\underline{x})$ and the classification function in the $\underline{y}$-space is $h(\underline{y})$, we can say that the original classification function $f(\underline{x})$ is expressed as a composition of two functions

$$f = h \circ g.$$

This is schematically illustrated in Fig. 1.1. By proper choice of g it is possible that the computations involved in finding g and h are considerably less than those involved in finding f directly. Considering that the feature extraction process might have picked up some features which are not very relevant for discrimination purposes, the transformation of features can also simplify the ultimate physical realization.

It is important that the new features $\underline{y}$ should contain all or most of the discriminatory information contained in the original features $\underline{x}$. In other words, data reduction should be achieved with minimum loss of discriminatory information. The task of finding a suitable function g which maps the original pattern vector $\underline{x}$ into a transformed pattern vector $\underline{y}$ of considerably lower dimension may be called the feature reduction problem.

In looking for a classification function f the search is necessarily restricted to a class of functions. The function f finally arrived at is supposed to be optimal only among the functions within the particular class. This is

SCHEMATIC DIAGRAM SHOWING ROLE OF FEATURE REDUCTION

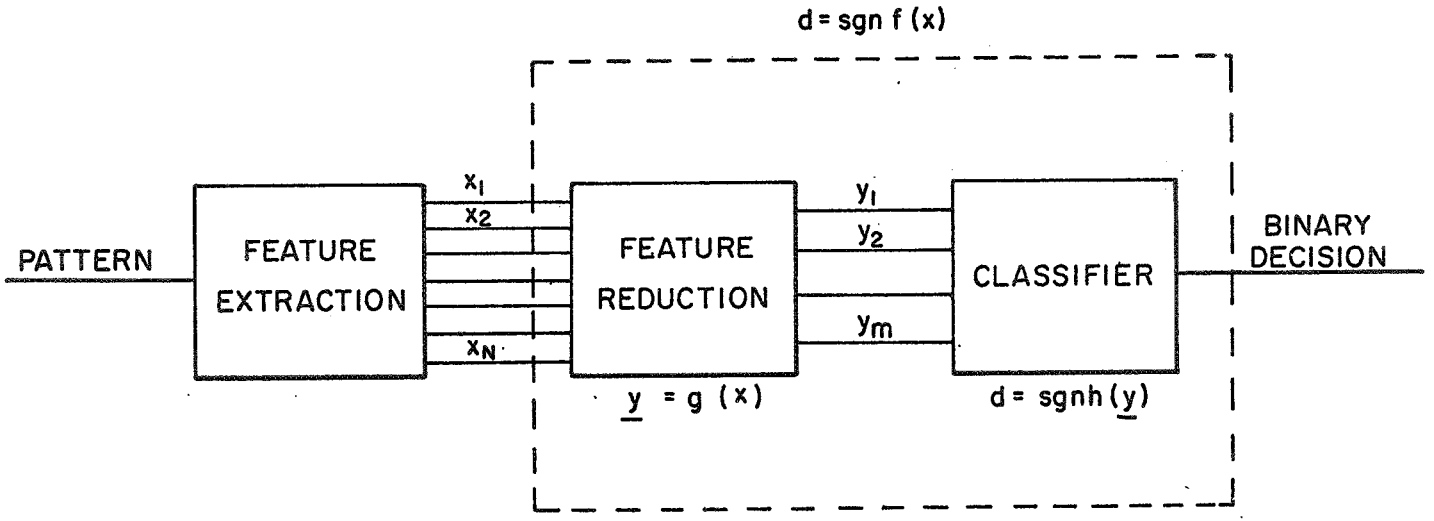


FIGURE 1.1

true of the feature reduction problem too. For fixed dimensions N and m, we look only at a class of functions among all possible mappings $R^N \rightarrow R^m$ and arrive at the optimal map within this class. In Section 1.2 the simplest case, where the mapping g just chooses a subset of the original features $\underline{x}$, will be considered. Section 1.3 deals with the next simplest kind of mappings, namely, those which produce linear combinations of the original features $\underline{x}$. Figures 1.2 and 1.3 show simple two-dimensional examples where such mappings are advantageous. In Fig. 1.1 the feature x_1 alone is sufficient for discrimination. In Fig. 1.2 the classification function can be reduced to a cubic in a single variable $y_1 = \alpha_1 x_1 + \alpha_2 x_2$ instead of a cubic in two variables x_1 and x_2.

1.2 Procedures for Choosing a Subset of Features

Procedures for choosing a subset of features will be considered now; linear combinations will be considered later.

The mapping g: $R^N \rightarrow R^m$ which chooses a subset of the features $\underline{x} \in R_N$ is

$$
\begin{aligned}
y_1 &= x_{i_1} \\
y_2 &= x_{i_2} \qquad \text{where } (i_1, i_2, \ldots, i_m) \subset (1, 2, \\
&\quad \vdots \qquad\qquad \ldots, N) \text{ with } i_1 \neq i_2 \neq \cdots \neq i_m. \\
y_m &= x_{i_m}
\end{aligned}
$$

EXAMPLE FOR SUBJECT
OF FEATURES

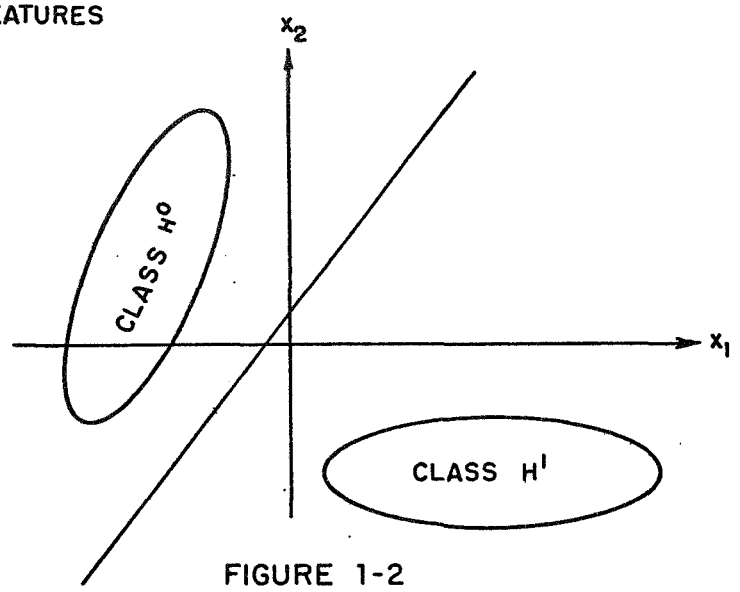


FIGURE 1-2

EXAMPLE FOR LINEAR COMBINATION
OF FEATURES

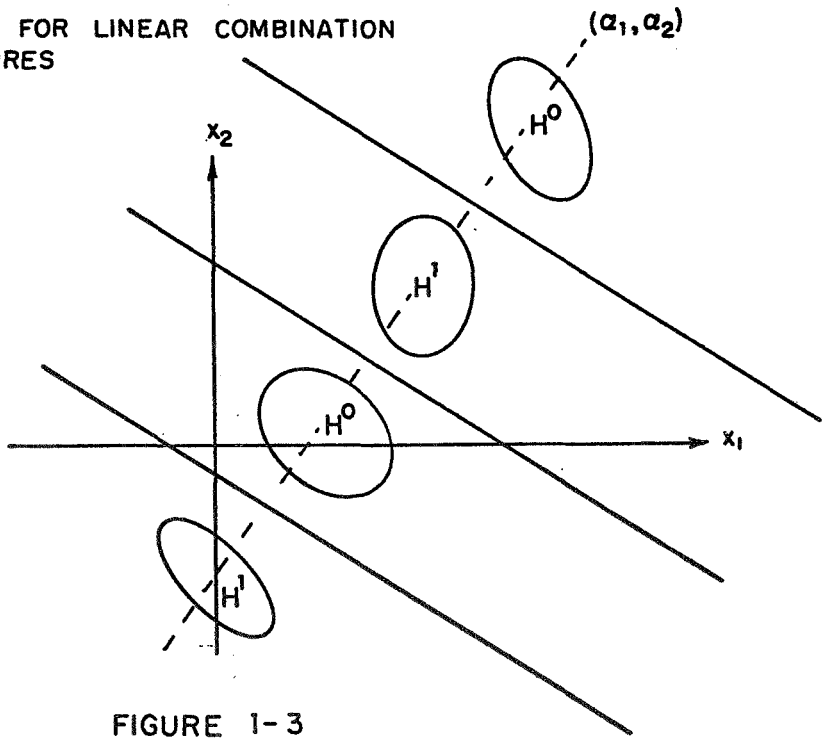


FIGURE 1-3

It is easy to see that there are precisely $\binom{N}{m}$ such mappings. The feature reduction problem amounts to finding one among these $\binom{N}{m}$ mappings which is optimal in some sense. The following subsections are devoted to a discussion of some approaches proposed in the literature. All of them assume that the joint probability densities of the features under each class are known and denoted by $p(x_1, x_2, \ldots, x_N \mid H^i)$, $i = 0, 1$.

1.2.1 Direct minimization of probability of error

Taking the Bayesian approach, let $P(H^0)$ and $P(H^1)$ be the subjective prior probabilities of the two pattern classes, which reflect one's judgment about the frequency of occurrence of each class. Choosing any subset of features $\chi_m = (x_{i_1}, x_{i_2}, \ldots, x_{i_m})$ we can compute the marginal densities $p(\chi_m \mid H^i)$, $i = 0, 1$ simply by integrating out the remaining features. If we are restricted to using only χ_m, the best that can be done is to use the Bayes decision rule, namely,

$$\underline{x} \in \begin{matrix} H^0 \\ H^1 \end{matrix} \quad \text{if} \quad \frac{p(\chi_m \mid H^0)}{p(\chi_m \mid H^1)} \gtrless 1.$$

The Bayes error P_e is given by

$$P_e = \int_{\Omega_m} \text{Min}[P(H^0)p(\chi_m \mid H^0),\ P(H^1)p(\chi_m \mid H^1)]d\chi_m$$

where Ω_m is the space spanned by χ_m. The Bayes error can be computed for all subsets of cardinality m and the subset which minimizes P_e is chosen.

This method suffers from several disadvantages, some of which are listed below.

(i) In a practical application, if N is large, it is difficult to get good estimates of the joint probability densities (this criticism can be levelled against all procedures described in section 1.2).

(ii) The total number of feature subsets to be examined is $\binom{N}{m}$, which can be prohibitively large and make the scheme impractical.

(iii) The integration needed to obtain P_e involves a minimization at each point in R^m and, therefore, may be difficult to carry out in practice.

1.2.2 Sequential decision methods [2]

These methods make use of the apparatus of sequential decision theory. The feature subset is constructed by sequential selection; its size is not specified in advance. The question asked at each stage of the selection process is "Is it worth picking another feature and, if so, which one should be chosen?" The criterion used is the Bayes risk, which takes into account the costs of misclassification and also the cost of observing the features. The selection of features can be done 'off-line' or 'on-line' as described below.

a) Off-line method

The method works 'off-line' in the sense that the whole strategy for sequential selection of features is worked out beforehand from a knowledge of the probability densities of $\underline{x}$ under H^0 and H^1. The actual values attained by the features do not influence the strategy in any way.

Let C_0 be the cost of labelling a pattern as belonging to H^1 when in reality it comes from H^0; let C_1 be the cost of labelling a pattern as belonging to H^0 when in reality it comes from H^1. Let γ be the cost of observing any one feature. If R(m) denotes the Bayes risk in making the decision based on m features, the sequence R(m), m = 0, 1, 2, . . . N can be computed as below.

A decision can be made without observing any features in the following manner. If we arbitrarily assign the pattern to class H^1 we incur an average cost of $C_0 P(H^0)$; whereas if we arbitrarily assign the pattern to class H^0 we incur an average risk of $C_1 P(H^1)$. Therefore,

$$R(0) = \text{Min } [C_0 P(H^0), C_1 P(H^1)].$$

If we decide to observe any one feature, say x_{i_1}, the minimum cost of misclassification will be

$$\int_{x_{i_1}} \text{Min } [C_0 P(H^0) p(x_{i_1} | H^0), C_1 P(H^1) p(x_{i_1} | H^1)] dx_{i_1}.$$

The Bayes risk R(1) will be obtained by minimizing this over all available features and adding the cost of observing one feature. Therefore,

$$R(1) = \min_{i_1 \in (1,2,\ldots,N)} \int \text{Min } [C_0 P(H^0) p(x_{i_1} | H^0), C_1 P(H^1) p(x_{i_1} | H^1)] dx_{i_1} + \gamma.$$

Having carried out the minimization, we know which feature to observe, if at all. At the second stage the choice is limited to the remaining features. It is inductively clear that the expression for R(m) is

$$R(m) = \min_{\substack{i_m : 1 \le i_m \le N \\ i_m \neq i_1, i_2, \ldots, i_{m-1}}} \int \text{Min } [C_0 P(H^0) p(\chi_m | H^0), C_1 P(H^1) p(\chi_m | H^1)] + m\gamma$$

According to sequential decision theory the m^{th} feature should be observed if and only if $R(m) < R(m-1)$ and $m \le N$. The selection procedure is stopped when this condition is violated. It should be noted that if $\gamma = 0$, $R(m) \le R(m-1)$ and the selection stops only when $m = N$.

b) On-line methods

In this method, the probability densities are updated at each stage in the light of the actual values assumed by the features so far chosen. In fact feature reduction

and classification proceed side by side. For example, if $z_{i_1}, z_{i_2}, \ldots, z_{i_{m-1}}$ are the observed values of the first (m-1) features

$$p(x_{i_m} | z_{i_1} z_{i_2}, \ldots, z_{i_{m-1}}, H^j) = \frac{p(z_{i_1} z_{i_2}, \ldots, z_{i_{m-1}}, x_{i_m} | H^j)}{p(z_{i_1} z_{i_2}, \ldots, z_{i_{m-1}} | H^j)}$$

$$j = 0,1$$

are known as functions of x_{i_m} alone since the z's are mere numbers. Therefore the revised Bayes risk R(m) is

$$R(m) = \underset{\substack{i_m:\ 1 \le i_m \le n \\ i_m \ne i_1, i_2, \ldots, i_{m-1}}}{\text{Min}} \int_{x_{i_m}} \text{Min}\,[C_0 P(H^0) p(x_{i_m} | z_{i_1}, \ldots, z_{i_{m-1}}, H^0),\ C_1 P(H^1) p(x_{i_m} | z_{i_1}, \ldots, z_{i_{m-1}}, H^1)]\, dx_{i_m} + m\gamma$$

The decision rule is unchanged. The number of features ultimately chosen is truly a random variable since it depends on $z_{i_1}, z_{i_2}, \ldots$ and may differ from one experiment to another.

The sequential methods suffer from a drawback inherent to all sequential feature selection procedures. This is elaborated in 1.2.5.

1.2.3 Methods based on entropy [3]

The different approaches outlined so far involve a detailed calculation of the probability of error, which can

be a difficult task under the best of circumstances (for example, both pattern classes Gaussian with different means and covariances). In an effort to overcome this difficulty, methods which rely only on a gross statistical description of the pattern classes have been developed. Entropy and distance measures are examples of such gross descriptions.

The concept of entropy arises in thermodynanics and information theory. To conform with the standard notation, in this subsection only, the symbol H will stand for entropy and the pattern classes will be denoted by C^i (instead of H^i), i = 0, 1. If $p(C^i)$ is the prior probability that an unknown pattern belongs to class i, the quantity

$$H(C) = - \sum_i p(C^i) \log p(C^i)$$

is called the entropy. It is a measure of the uncertainty regarding the correct classification of a pattern. It can be proved that

(i) H(C) is a minimum when all but one of the prior probabilities are zero--in this case one can be certain about the classification of any given pattern.

(ii) H(C) is a maximum when all the prior probabilities are equal to each other--in this case one is most uncertain about the correct classification of any given pattern.

In general, the entropy attains a lower value when the

probabilities are concentrated in a few pattern classes; it attains a higher value when the probabilities are distributed among many classes.

When a set of features $\chi_m = \{x_{i_1}, x_{i_2}, \ldots, x_{i_m}\}$ has been observed, the probabilities associated with the pattern classes are revised according to Bayes' Law, namely,

$$p(C^j|\chi_m) = \frac{p(\chi_m|C^j)p(C^j)}{\sum_j p(\chi_m|C^j)p(C^j)}.$$

Therefore, the <u>conditional entropy</u> after observing χ_m is

$$H(C|\chi_m) = -\sum_j p(C^j|\chi_m) \log p(C^j|\chi_m).$$

Averaging over all possible values which can be attained by the features, we get the average conditional entropy

$$\bar{H}(C|\chi_m) = -\sum_j \int_{\chi_m} p(C^j, \chi_m) \log p(C^j|\chi_m) d\chi_m.$$

The average reduction in entropy achieved is defined as the mutual information, that is, the information which the particular set of features carries about the correct classification of the pattern.

$$I(C|\chi_m) = H(C) - \bar{H}(C|\chi_m).$$

It is proved in textbooks on information theory [3] that the mutual information is a non-negative quantity. The mutual information can be computed for all $\binom{N}{m}$ feature subsets of dimension m and the one which has maximum mutual information with C^i is chosen.

The integrations involved in the computation of the mutual information I are simpler because the minimizing operation at each point in m-dimensional space has been avoided. However, entropy being a gross description of the feature statistics, no unique relation exists between it and the probability of misclassification.

1.2.4 Methods based on distance measures

These methods stem from the intuitive notion that the farther away the two pattern classes are situated in feature space, the less will be the probability of committing an error in classification. Since each pattern class should be properly looked upon as a statistical distribution in feature space, the problem is to define an appropriate distance between any two statistical distributions. It would be desirable for this 'distance' to satisfy the three metric properties of non-negativity, symmetry and triangle inequality. (This last one does not apply to the two-class case with which this thesis is mainly concerned.) In a practical application, the distance is evaluated in the subspace spanned by each of the $\binom{N}{m}$ feature subsets of dimension m. The subset which maximizes the distance is chosen.

A few of the commonly used distance measures and their properties are given below. (Kailath [6] has given a good account of these.) As in the case of entropy, there is no explicit relation between the probability of error and any of the distance measures. However, in some cases bounds can be set on the probability of error.

a) $$d = \int_{\chi_m} |p(\chi_m|H^0) - p(\chi_m|H^1)|^r d\chi_m, \quad r \geq 1$$

It is proved in textbooks on functional analysis [4] that the above distance satisfies the metric properties. No relation with the probability of error is known.

b) Kullback's J-divergence or divergent information [5, 6]

The Bayes decision rule can be interpreted as setting a threshold on the log-likelihood ratio defined as

$$\log L(\chi_m) = \log \frac{p(\chi_m|H^0)}{p(\chi_m|H^1)}.$$

It is reasonable to define a distance as the difference in the average values of the log-likelihood ratio under the two classes. The distance so defined is called the J-divergence and is given by

$$\begin{aligned} J &= E[\log L(\chi_m)|H^0] - E[\log L(\chi_m)|H^1] \\ &= \int_{\chi_m} \log \frac{p(\chi_m|H^0)}{p(\chi_m|H^1)} p(\chi_m|H^0) d\chi_m - \\ &\quad \int_{\chi_m} \log \frac{p(\chi_m|H^0)}{p(\chi_m|H^1)} p(\chi_m|H^1) d\chi_m. \end{aligned}$$

Non-negativity follows by

$$J = \int_{\chi_m} [p(\chi_m|H^0) - p(\chi_m|H^1)] \log L(\chi_m) d\chi_m$$

$$= \int_{\chi_m} [L(\chi_m) - 1] \log L(\chi_m) p(\chi_m|H^1) d\chi_m$$

$$\geq 0,$$

since $(L-1) \log L \geq 0$ with equality if and only if $L = 1$. Symmetry is obvious; J does not obey the triangle inequality. The following theorem due to Karlin [7] is of some interest. "Let χ and χ' be two subsets of features (not necessarily of the same cardinality). If J and J' denote the divergent information contained in these subsets, then $J > J'$ implies the existence of a set of prior probabilities $\{\pi, 1-\pi\}$ associated with the two pattern classes such that $P_e(\chi, \pi) < P_e(\chi', \pi)$."

Karlin's theorem only assures us that the subset containing less divergent information cannot be uniformly better than the subset with more divergent information for all prior probabilities. It does not even tell us for what range of π the latter subset performs better. Kovalewsky [8] has pointed out that the measure J cannot distinguish between features which are distributed as shown in Fig. 1.4, even though the first feature is obviously better (no overlap of classes). This is because J becomes infinite for both the features.

EXAMPLE WHERE J-DIVERGENCE FAILS

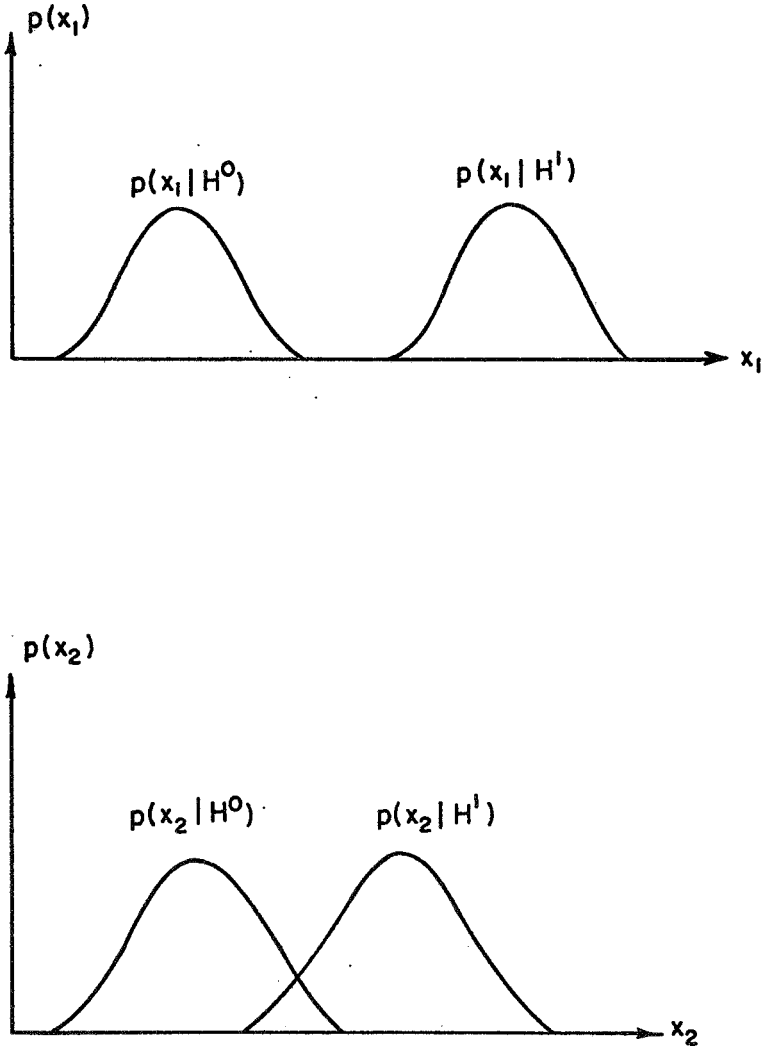


FIGURE 1-4

c) The Bhattacharyya distance (B-distance) [6, 9]

The B-distance between the two pattern classes is defined by

$$B = - \ln \int_{\chi_m} \sqrt{p(\chi_m|H^0)p(\chi_m|H^1)} \; d\chi_m$$

B satisfies non-negativity and symmetry properties; it does not satisfy the triangle inequality. Kailath has given bounds on the probability of error P_e in terms of B.

It was pointed out that distance measures, being only gross statistical descriptions, cannot be related explicitly to the probability of error. The following theorem due to Blackwell [10] brings out this fact forcefully.

"$P_e(\chi,\pi) \leq P_e(\chi',\pi)$ for all π if and only if

$$E_\chi[\psi[L(\chi)]|H'] \leq E_{\chi'}[\psi[L(\chi')]|H']$$

for all continuous concave functions $\psi(L)$." J-divergence and B-distance are only particular cases with $\psi(L) = (1-L) \ln L$ and $\psi(L) = \sqrt{L}$ respectively. Other distance measures can be constructed by taking different concave functions $\psi(L)$. Blackwell's theorem is not useful in practice since it is impossible to test the inequality for all continuous concave functions.

1.2.5 Suboptimality of sequential feature selection procedures

The real drawback of all feature reduction procedures which look for the best subset of given cardinality is that the number of such subsets $\binom{N}{m}$ can be prohibitively large for any reasonably large size problem. It might be thought that the features that make up the best subset of cardinality m can be picked up in a sequential fashion in the order of decreasing discriminatory information. In fact, the sequential decision methods described in 1.2.2 work precisely in this manner. However, such sequential procedures can never be truly optimal for the following reason--the best subset of features of cardinality k is not necessarily a subset of the best subset of features of cardinality (k+1).

A simple example in Fig. 1.5 illustrates why this is so. Each pattern is described by three features (x_1, x_2, x_3); the two pattern classes are distributed on planes $\alpha_1 x_1 + \alpha_2 x_2 = k_0$ and $\alpha_1 x_1 + \alpha_2 x_2 = k_1$ respectively. If projected onto one of these planes, the two classes overlap slightly along the x_3 direction as shown. The marginal densities of x_1, x_2 and x_3 are roughly as indicated. It is clear that x_3 is the best single feature for discrimination since the densities overlap least. However, if we want to consider two features, x_1 and x_2 afford perfect discrimination depending on whether $\alpha_1 x_1 + \alpha_2 x_2$ evaluates to k_0 or k_1.

EXAMPLE SHOWING SUBOPTIMALITY OF SEQUENTIAL FEATURE SELECTION

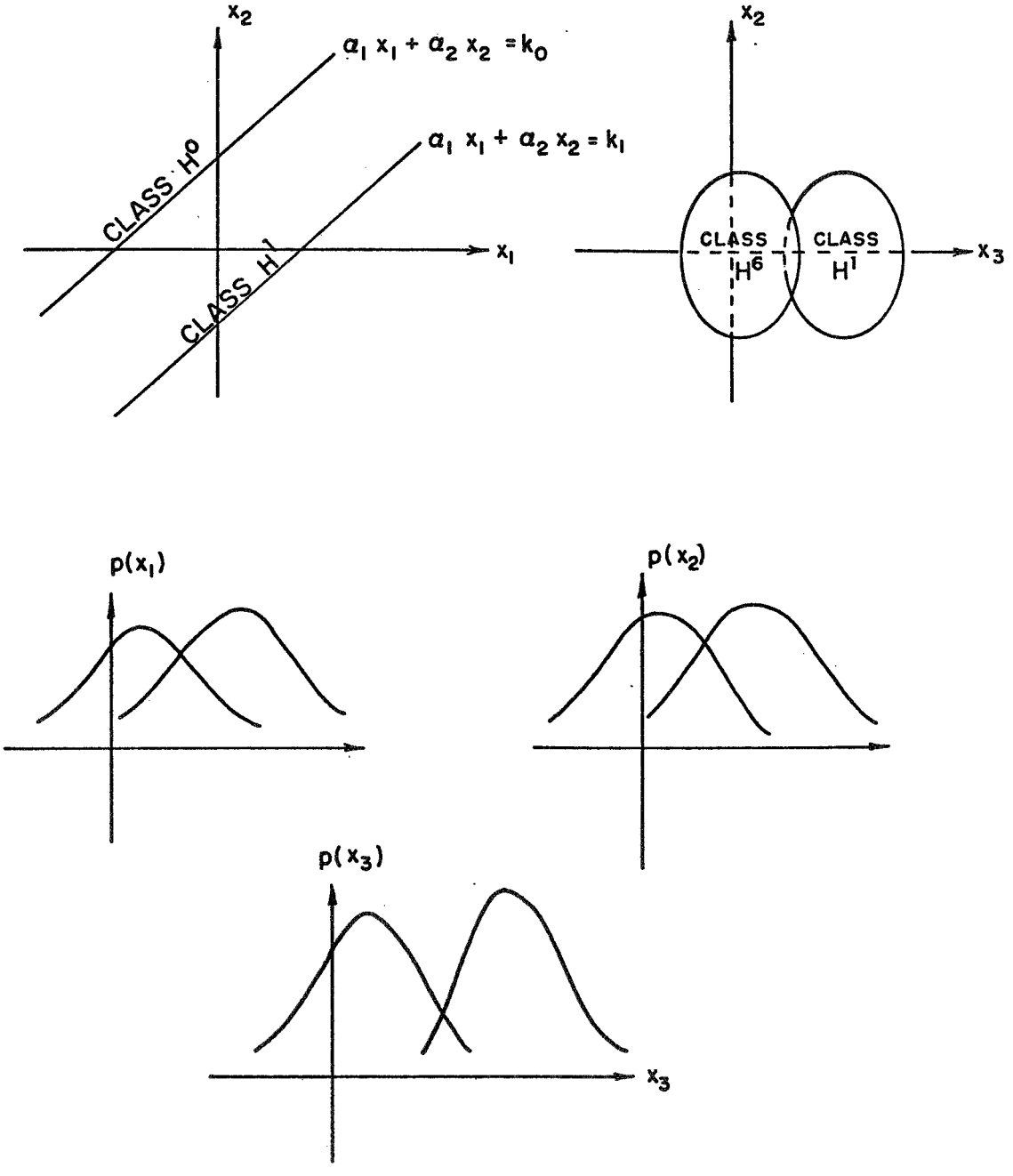


FIGURE 1-5

Neither of the other pairs (x_1, x_3) and (x_2, x_3) can perform as well. The point is proved since $x_3 \not\subset (x_1, x_2)$. A sequential procedure, which picks up x_3 first and retains it, can never discover the perfect pair (x_1, x_2).

1.3 Procedures for Choosing Linear Combinations of Features

Having surveyed methods for choosing a significant subset of features, we shall now consider the next simplest feature mappings, namely, those which effect a linear transformation of the feature space. The problem is to choose a m X N ($m < N$) matrix A such that the new features $\underline{y}$ given by $\underline{y} = A\underline{x}$ contain all or most of the discriminatory information contained in the original features $\underline{x}$. Two procedures given in the literature are outlined below.

1.3.1 The Karhunen-Loéve expansion [11, 12, 13, 14]

Watanabe [11] first considered the problem of information compression applied to continuous signals. Given a continuous signal $\{x(t): t \in (0,T)\}$ with known statistics, the problem is to map it into a finite-dimensional vector $\underline{y} = (y_1, y_2, \ldots, y_m)$ which contains the maximum possible information about the continuous signal, in some sense.

Let $R(t, \tau) = E[x(t)x(\tau)]$ be the correlation function of the random process $x(t)$ and let $\{\psi_i(t)\}$ be the complete, orthonormal set of eigenfunctions of the following integral equation:

$$\int_0^T R(t, \tau)\psi_i(\tau)d\tau = \lambda_i \psi_i(t).$$

Further let the eigenfunctions be ordered in such a way that $\lambda_1 > \lambda_2 > \lambda_3 \ldots$. (The eigenvalues can be proved to be real and positive.) Now if we expand any given signal x(t) chosen from the ensemble representing the random process, in terms of $\{\psi_i(t)\}$

$$x(t) = y_1 \psi_1(t) + y_2 \psi_2(t) + \ldots$$

it can be proved that the coefficients $\{y_i\}$ are uncorrelated, and that $E|y_i|^2 = \lambda_i$. Since correlation implies redundancy we can expect the representation $(y_1, y_2, \ldots)$ to be optimal in some sense. If we consider only the truncated expansion $\sum_{i=1}^{m} y_i \psi_i(t)$, Watanabe [11] proved that the mapping

$$x(t) \rightarrow \begin{pmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ \cdot \\ y_m \end{pmatrix}$$

has the following optimal properties:

(i) Let $\{\theta_i(t)\}$ be any other complete set of orthonormal functions and let

$$x(t) = z_1 \theta_1(t) + z_2 \theta_2(t) + \cdots$$

Then

$$E\left|x(t) - \sum_{i=1}^{m} y_i \psi_i(t)\right|^2 \leq E\left|x(t) - \sum_{i=1}^{m} z_i \theta_i(t)\right|^2;$$

that is, the co-ordinate system $\{\psi_i(t)\}$ minimizes the mean square error of approximation.

(ii) Remembering $\lambda_i = E|y_i|^2$, let $\rho_i = E|z_i|^2$. Then,

$$-\sum_{i=1}^{m} \lambda_i \log \lambda_i \leq -\sum_{i=1}^{m} \rho_i \log \rho_i .$$

From what was said about entropy in section 1.2.3 the following interpretation is obvious. Physically speaking, the energy of the signal will tend to be concentrated on the average in fewer 'modes' if the signal is represented in the Karhunen-Loéve co-ordinate system, rather than any other co-ordinate system.

The signal x(t), instead of being infinite dimensional, can be of large but finite dimension, say $(x_1, x_2, \ldots, x_N)$. The foregoing results still apply if we replace 'correlation function' by 'correlation matrix' and 'eigenfunctions' by 'eigenvectors.' The application to feature reduction is obvious--except for the fact that in a pattern classification problem the signal does not come from a single statistical distribution; rather, it is

generated by several distributions (one corresponding to each pattern class) each having a prior probability associated with it. Chien and Fu [13] have shown that the foregoing results still hold if we replace $E(\cdot)$ by $\sum_i E(\cdot|H^i)p(H^i)$. In other words, the averaging has to be done with respect to a weighted distribution. The mechanics of the so-called 'generalised Karhunen-Loéve expansion' are as follows.

The correlation matrices $E(\underline{x}\underline{x}^T|H^i)$ of each class are either known or estimated from samples of known classification. The eigenvectors $\{\psi_i: \ i = 1, 2, \ldots, N\}$ of the average correlation matrix $\sum_i E(\underline{x}\underline{x}^T|H^i)p(H^i)$ are found and arranged in the descending order of eigenvalues. Any new sample $\underline{x}$ can be expressed as a unique linear combination of the eigenvectors

$$\underline{x} = y_1\underline{\psi}_1 + y_2\underline{\psi}_2 + \cdots + y_N\underline{\psi}_N.$$

Then the mapping

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix} = \begin{pmatrix} \underline{\psi}_1^T \\ \underline{\psi}_2^T \\ \vdots \\ \underline{\psi}_m^T \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{pmatrix}$$

achieves feature reduction optimally in the sense described above.

Unlike any method discussed before, this procedure does not require complete knowledge of probability density functions. Knowledge about second order statistics of the patterns is sufficient.

1.3.2 Optimization of some criterion involving A [15]

The problem of choosing the optimal transformation A can be viewed as that of obtaining the minimising (or maximising) A with respect to a suitable criterion. An example is the following. Suppose $p(\underline{x}|H^0)$ and $p(\underline{x}|H^1)$ are known or can be estimated from training samples. With $\underline{y} = A\underline{x}$, the densities $p(\underline{y}|A, H^0)$ and $p(\underline{y}|A, H^1)$ can be computed as functions of A. We might then ask for the A which maximizes the distance between the transformed pattern classes

$$J(A) = \int |p(y|A, H^0) - p(y|A, H^1)|^2 dy.$$

The maximization can be carried out by gradient procedures. It can also be subjected to certain constraints. For example, one can constrain A to be a projection map, i.e. the rows of A are orthonormal.

The main drawback of such methods is that the number of elements in the matrix A can be quite large for any reasonably large size problem. The gradient search will have to be undertaken in a space of large dimension, making the approach computationally difficult.

1.4 The Contribution of this Thesis

In this thesis, feature reduction and classification techniques are applied to electroencephalographic (EEG) signals. The properties of these signals are given in detail in Chapter II, which also describes how two classes of EEG patterns arise. An EEG record represents a large amount of data and efficient feature reduction techniques are no doubt required.

With the exception of the Karhunen-Loève expansion described in 1.3.1, every other feature reduction method surveyed requires a complete knowledge of the probability density functions of the patterns under each class. This requirement is hardly satisfied by the EEG signals. Chapter III develops a sequential feature selection method based on a distance measure, which relies only on the first and second order statistics of the EEG records. In this sense, the method may be called 'non-parametric.' Even this method becomes unwieldy if we have to base our decisions on long lengths of EEG record.

Chapter IV develops a random process model for the EEG signals in an effort to put some structure into the apparently disorganised data. The model makes rather strong assumptions; however, it is shown that certain statistical and spectral properties of the signals predicted by the model are in accord with the observed facts.

Since the model makes strong assumptions, we can expect a pay-off in the form of simplified feature reduction and classification procedures. In Chapter V it is shown that the model yields a vector of sufficient statistics, whose dimension is independent of the length of the EEG record. In other words, the information contained in the whole record can be summarized by a set of numbers, which can be continuously updated as more data comes in. It is also shown that the optimal separating surface is linear in terms of these sufficient statistics, which leads to simple classification procedures.

CHAPTER II

A QUALITATIVE DESCRIPTION OF EEG

The feature reduction techniques developed in this thesis are generally applicable in any pattern classification problem. However, they are illustrated with special reference to the case of electroencephalograms (EEG). This chapter is devoted to a brief, qualitative description of relevant properties of EEG signals. It is hoped that this discussion will help the reader in better appreciating the results obtained in later chapters.

2.1 Spontaneous EEG

The term electroencephalography refers to the study of the electrical activity of the human brain. This electrical activity is usually studied by recording potential changes of the order of microvolts on the surface of the scalp. The precise origin of these potentials is not yet fully understood. However, there is general agreement that the potentials observed on the scalp are due to the synchronous activity of a large number of cells in the brain.

2.1.1 Methods of recording

The EEG potentials are picked up by electrodes arranged in transverse and longitudinal positions on the scalp as described by Rémond et al. [16]. The recording may

be unipolar or bipolar. In the former case the potentials are measured with reference to an arbitrary level (usually called the indifferent electrode); in the latter case the potentials are measured between pairs of electrodes on the scalp. The bipolar method really measures potential gradients rather than potentials themselves.

In order to achieve some degree of reproducibility in the results, it is necessary that the human subject be kept in the same psychophysiological state during different recordings. The term 'spontaneous EEG' refers to the case where the recording is made with the subject in an alert state, but cut off from external visual or auditory stimuli in a darkened, soundless room. For general experimental conditions see Anliker [43].

2.1.2 The spatio-temporal nature of EEG

At any instant during the recording of the EEG, the observed potential differs from point to point on the scalp. Therefore, a complete description of the phenomenon can only be given by means of potential maps at every instant of time [17]. The spatial distribution of the potential is of some importance in medical applications of the EEG. For example, in most normal subjects the potentials are approximately symmetric about the center line of the skull and any asymmetries are of diagnostic value. However, of greater importance is the temporal behavior of the EEG potentials. In order to study the temporal behavior, the spatial parameter has to be eliminated. This can be done

by observing the potential of only a single electrode (in the case of unipolar records) or only between a single pair of electrodes (in the case of bipolar records), or by averaging the potentials of several electrodes as described by Rémond et al. In succeeding sections we shall be concerned only with the temporal behavior of the EEG, which is typified by Fig. 2.1.

2.1.3 Spectral analysis [18, 19, 20, 21]

As pointed out earlier, the EEG potentials indicate the gross activity of a large number of cells and therefore need a statistical description. That is, a given EEG record can be considered as a sample from a random process. The first question to be asked is whether there are any dominant frequencies (hidden periodicities) in the fluctuating signal. The techniques of spectral analysis (or, equivalently, correlation analysis) of random signals can be used to answer this question. Blackman and Tukey [18] have described how to get good estimates of power spectra from finite samples of a random process. The basic facts are that the estimate gets better if the spectrum is computed from a longer length of the signal, and, in case we have only sampled values of the signal, the resolution of the spectral estimate can be improved by sampling more frequently. Various authors [19, 20, 21] have applied correlation and spectral techniques to EEG signals. A typical spectral estimate of a spontaneous EEG record is shown in Fig. 2.2. It was

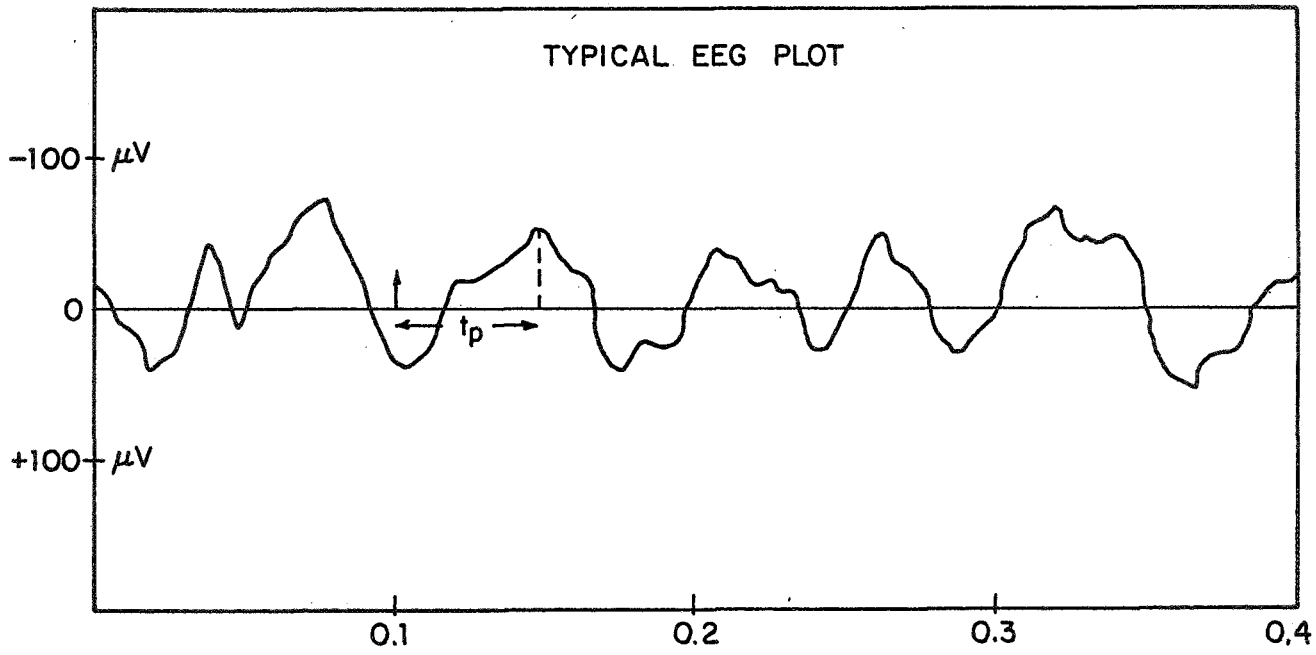


FIGURE 2-1

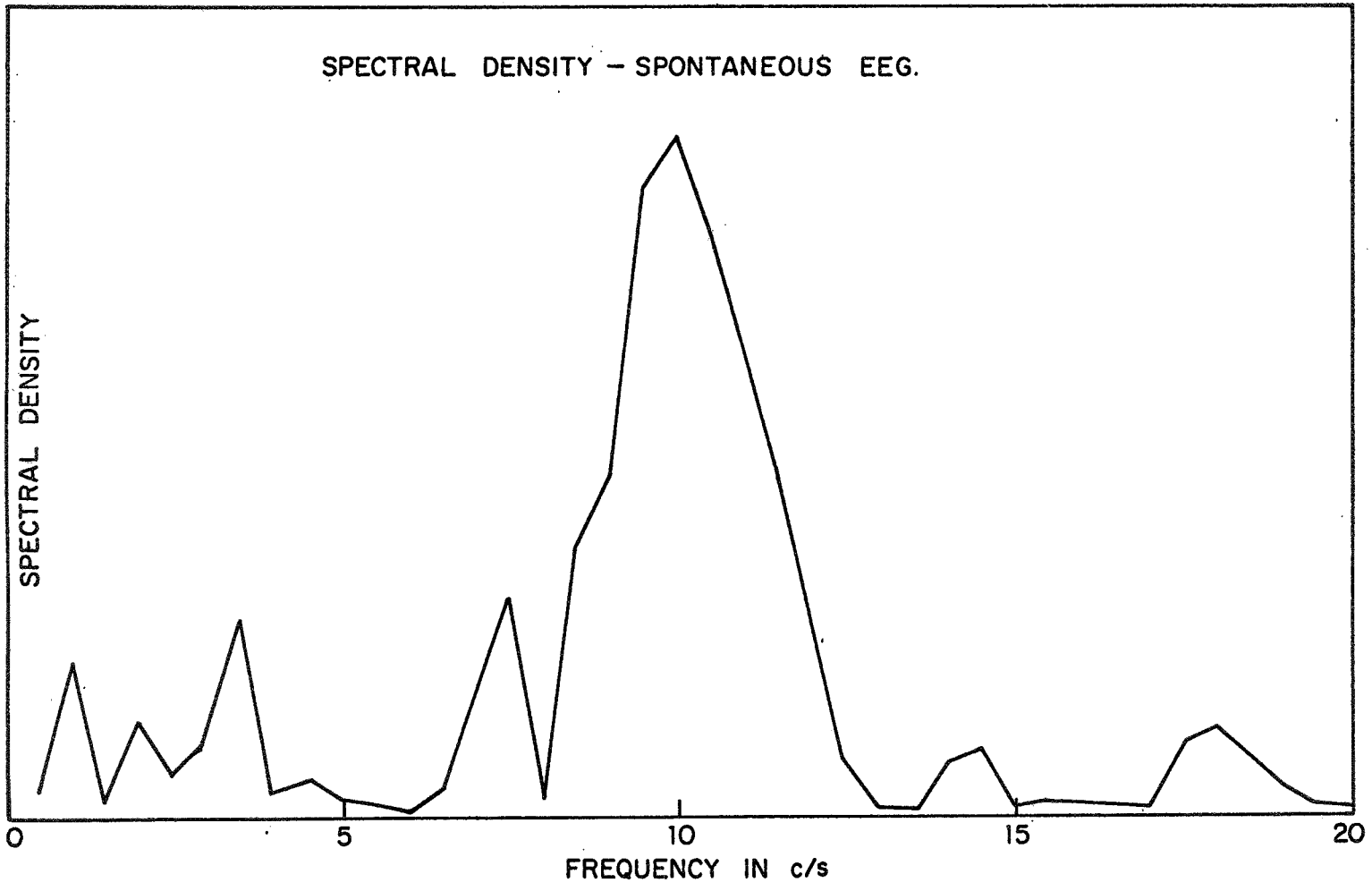

SPECTRAL DENSITY – SPONTANEOUS EEG.
SPECTRAL DENSITY
0
5
10
15
20
FREQUENCY IN c/s

FIG 2-2

computed from $2^{13} = 8192$ sampled values, the sampling interval being one millisecond.

This power spectrum exhibits most of the typical behavior ascribed to such spectra in the literature. There is considerable activity in the band of 8.5-10.5 c/s. This is called the α-activity. It is generally true that the α-frequency predominates in most alert normal adult subjects, even though there are 'nondominant alpha subjects' also. Hereafter, the term 'alpha frequency' shall apply to the mode of the power spectrum. The spectrum has smaller peaks on either side of the alpha frequency--they are the δ-activity (0.5-1.5 c/s), the θ-actitivy (3-8 c/s), the σ-activity (12-15 c/s), and the β-activity (17-20 c/s). In this thesis, we shall be mainly concerned with the behavior of the spectrum near the alpha frequency.

2.1.4 Statistical properties [19]

First and second order statistical properties of the EEG signals are also of interest. If an EEG record is split up into equal parts, with the length of each part equal to the period of the alpha frequency and these parts are averaged, then it is found that the average signal is nearly zero as shown by the dotted line in Fig. 2.3. Mathematically,

$$\mu(t) = \frac{1}{K} \sum_{i=1}^{K} x(t + it_a) \sim 0, \quad 0 \leq t \leq t_a$$

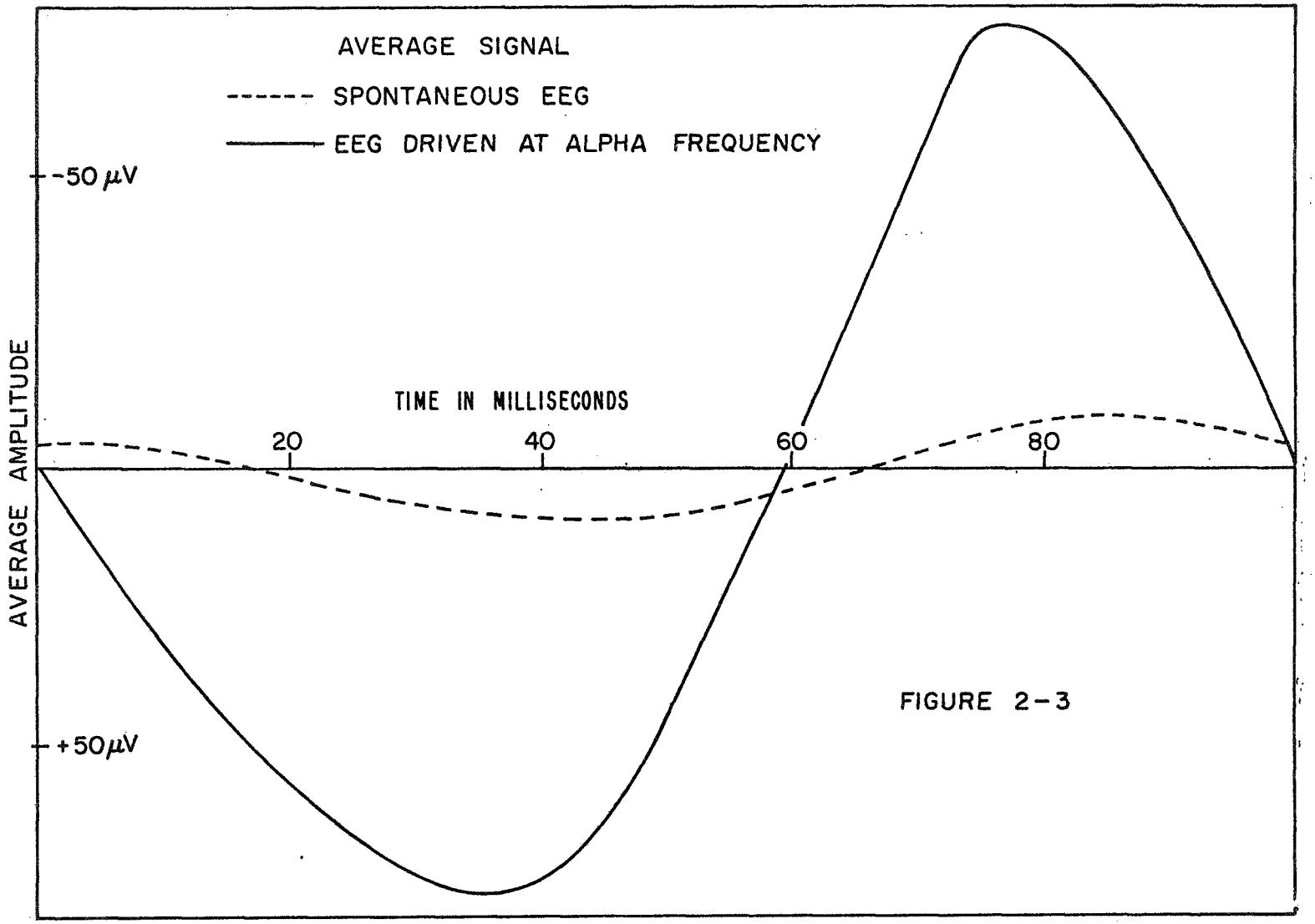

AVERAGE SIGNAL
SPONTANEOUS EEG
EEG DRIVEN AT ALPHA FREQUENCY
-50 μV
AVERAGE AMPLITUDE
TIME IN MILLISECONDS
20
40
60
80
+50μV

FIGURE 2-3

for large K, where $t_\alpha = \frac{1}{f_\alpha}$, f_α being the alpha frequency. This tends to rule out any 'additive noise' models; if indeed the 'noise' were additive in nature, it would have been annulled in the average and the hidden cyclic activity would have shown up. Rather, it suggests that the alpha activity tends to lose its phase coherence with the elapse of time. It should be remembered that if a large number of sinusoids with different phase relationships are added, they might average to zero.

The second order statistical properties are generally displayed by means of the autocorrelogram defined by

$$R_T(\tau) = \frac{1}{T} \int_0^T x(t)x(t + \tau)dt.$$

If x(t) is a stationary random process with autocorrelation function $R_{xx}(\tau)$ then $ER_T(\tau) = R_{xx}(\tau)$, showing that $R_T(\tau)$ is an unbiased estimate of $R_{xx}(\tau)$. It can also be proved [19] that the variance of this estimate is proportional to $\frac{1}{T}$. The Fourier transform of $R_T(\tau)$ is of course the spectral estimate discussed in the previous subsection.

A typical autocorrelogram of the spontaneous EEG is shown by the dotted line in Fig. 2.4. It exhibits a decaying cosine behavior, with its period equal to that of the alpha frequency and the rate of decay reflecting the bandwidth of the power spectrum around the alpha peak.

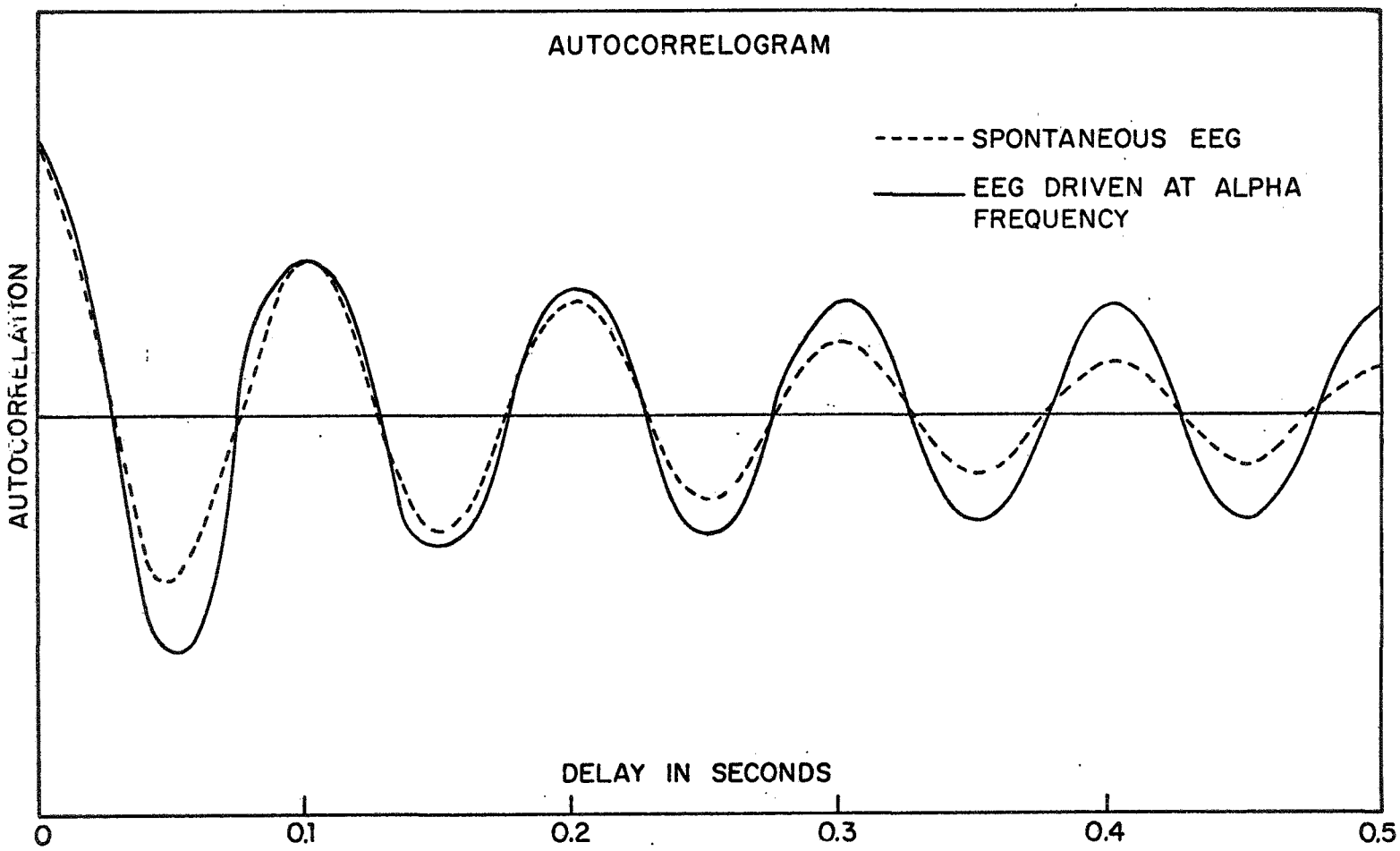

AUTOCORRELOGRAM
SPONTANEOUS EEG
EEG DRIVEN AT ALPHA FREQUENCY
AUTOCORRELATION
DELAY IN SECONDS
0
0.1
0.2
0.3
0.4
0.5

FIGURE 2-4

2.2 EEG with Photic Stimuli

2.2.1 Evoked potentials [19, 22, 23]

The spontaneous EEG, as we have seen, represents the electrical activity of the brain when the subject is cut off from external visual and auditory stimuli. Various authors [19, 22] have studied the effect of a sudden flash of light (photic stimulus) on the observed EEG potentials. If the procedure is repeated several times, the statistical average of the EEG potential immediately following the presentation of a stimulus would look as shown in Fig. 2.5. (The stimuli should be presented with enough time interval between them to reduce the inevitable correlations). The potential shows a rise (conventionally EEG plots have negative scalp potentials drawn upwards) to a peak and then dips to a smaller peak in the negative direction before returning to normal. The period of this oscillation is approximately that of the alpha frequency and it is called the evoked potential for obvious reasons. Kitajima [23] has found that the magnitude of the evoked potential is approximately proportional to the logarithm of the flash intensity.

2.2.2 Response to repetitive stimuli [24]

The evoked potential can be likened to the impulse response of a dynamical system comprising of the brain and the associated neural paths. It is of interest to study the dynamic steady state response (frequency response) of

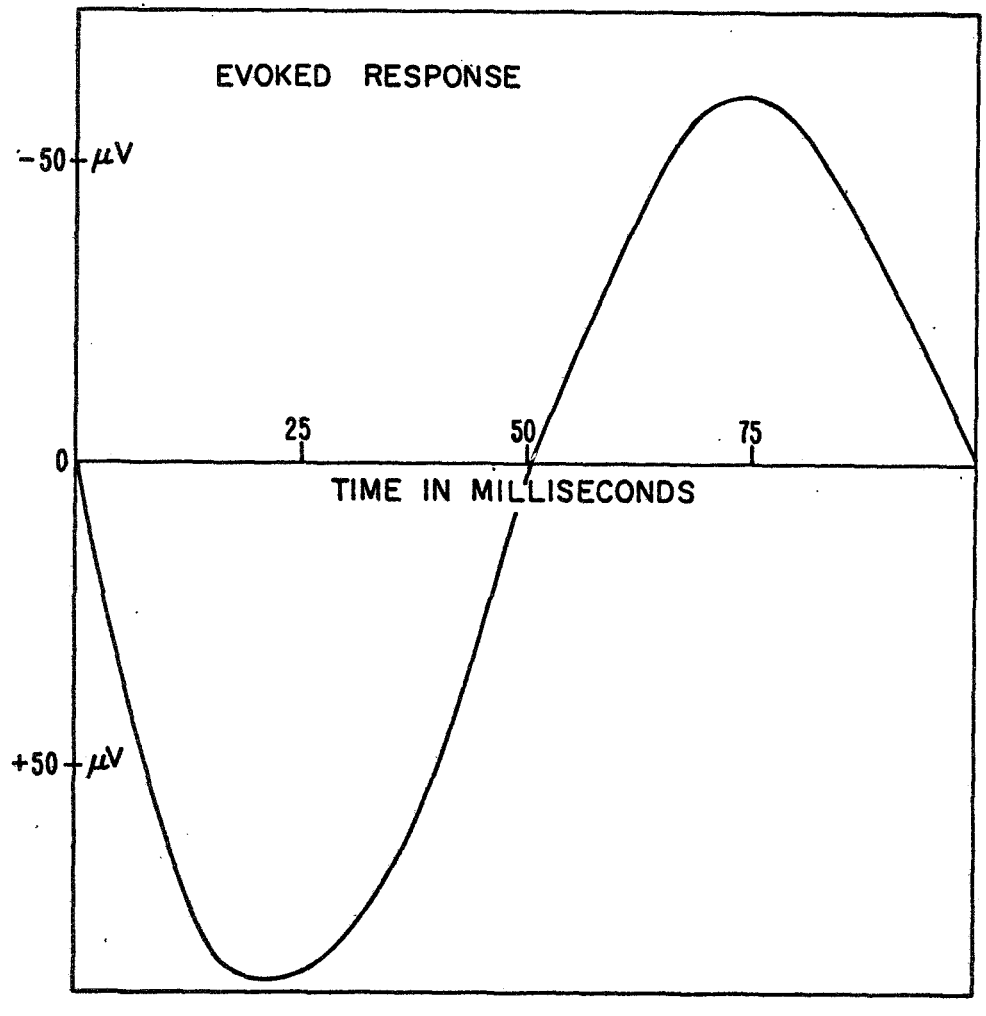


FIGURE 2-5

this system, which can be achieved by presenting repetitive stimuli at a fixed rate. The response could then be studied as a function of the stimulus frequency. (The frequency response can properly characterize only a linear time invariant system. The system representing the brain and the neural paths is certainly more complicated.)

The result of these studies [24] has been as follows. When the stimulus frequency is very near the alpha frequency, the stimuli have the effect of driving the EEG into resonance. The power spectrum of the EEG would then show a sharper peak at the alpha frequency as in Fig. 2.6, which should be compared with Fig. 2.2. The statistical properties of the signal also undergo a change. The average signal, shown by the solid line in Fig. 2.3, is no longer zero, but looks very much like an evoked response in Fig. 2.5. The autocorrelogram (solid line in Fig. 2.4) decays at a slower rate reflecting the decreased bandwidth around the alpha frequency. It is as though the stimuli have the effect of introducing phase coherence into the alpha activity.

If the stimulus frequency is slightly different (say 1 c/s away) from the alpha frequency sometimes it is found that the alpha frequency locks onto the new frequency. However, if the stimulus frequency differs considerably from the alpha frequency, the phase coherence is lost and once more we have spontaneous EEG behavior. Sometimes, the alpha activity may even be blocked completely.

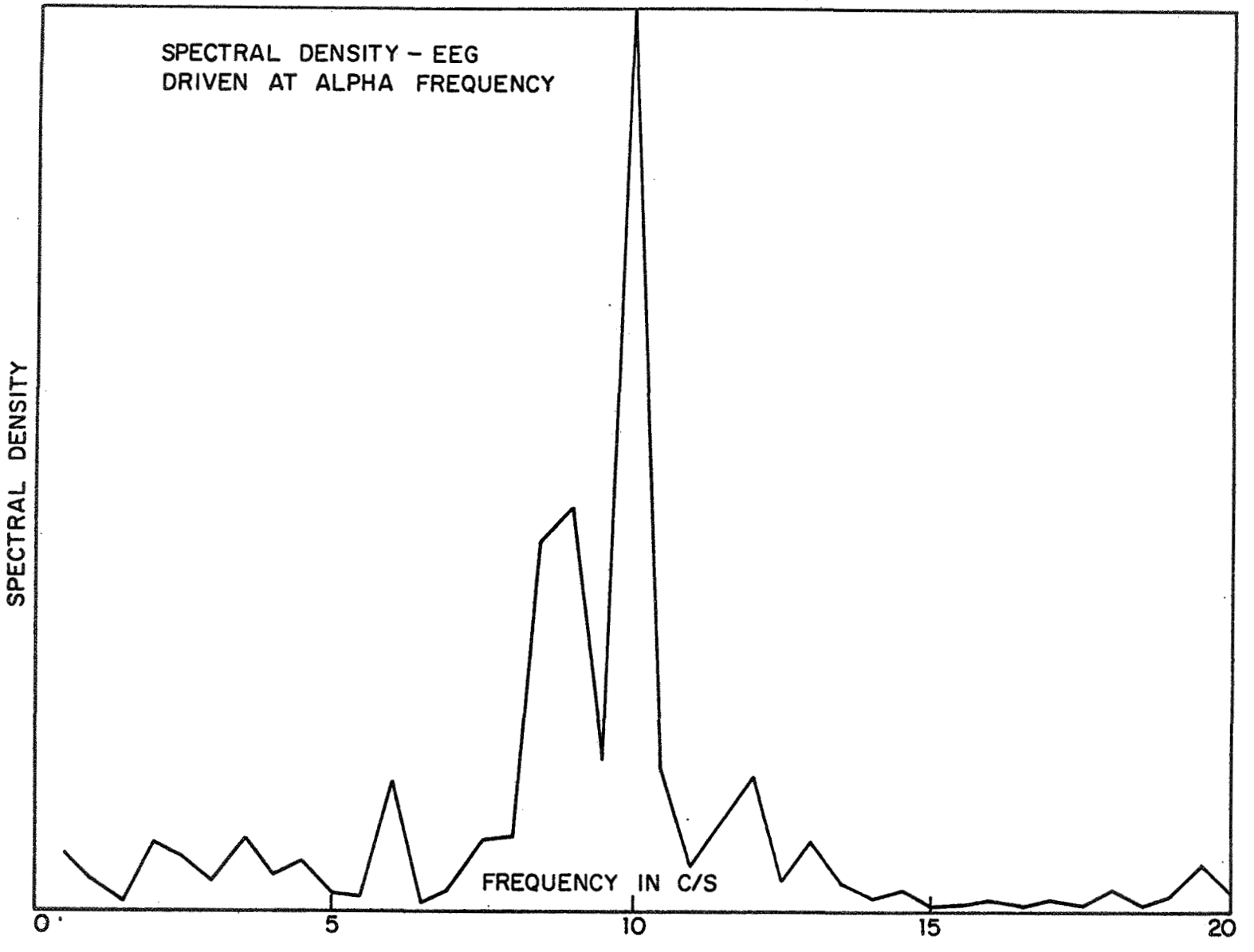

SPECTRAL DENSITY – EEG
DRIVEN AT ALPHA FREQUENCY
SPECTRAL DENSITY
FREQUENCY IN C/S
0
5
10
15
20

FIGURE 2-6

2.3 Statement of the Classification Problem

The problem considered in this thesis is one of distinguishing between a spontaneous EEG record and an EEG record with photic stimuli applied at the alpha frequency. In the latter case, the instants at which the stimuli are applied are assumed to be known. From what has been said before, it is clear that if we have a sufficiently long record, we can get a good spectral estimate. We could then compute the energy within a certain narrow band centered at the alpha frequency and, if this energy is less than a certain threshold, the record can be classified as coming from a spontaneous EEG. However, if we have to make the decision from a short length of the record the frequency domain analysis would not be very accurate and hence the decisions based on it would not be very reliable.

The following time domain approach has been taken by physiologists to solve the classification problem using the phase coherence (time-locking) feature of the EEG when driven by repetitive flashes at the alpha frequency.

Let t_p be the time taken by the EEG record in Fig. 2.1 to attain its maximum positive value after the reference mark (which coincides with the stimulus if one is present). t_p is evaluated for a large number of responses known to belong to the same class. t_p is treated as a random variable; its mean and variance are then computed. If the stimulus were present at the beginning of each waveform,

then the variance of t_p tends to be smaller because of the time-locking feature. On the contrary, if the stimulus were absent, there is no time-locking and so the variance of t_p tends to be large. The disadvantage of this method is that a fairly large number of responses is needed to get a good estimate of the variance and hence the decision cannot be reached quickly. Typically, the number of responses needed to make a reliable decision is of the order of 600.

The object of this thesis is to develop feature reduction and classification methods based on short lengths of EEG record (typically 20 - 50 periods of the alpha frequency). Such methods can be useful in the case where the application of the stimuli may be in short bursts.

2.4 Description of Data

The data were obtained from EEG records of two different subjects. The potentials were recorded by means of a pair of electrodes located in the left occipital parietal area. The records were of 10 minutes duration in each case. Stroboscopic light was flashed into the eye of the subject at the frequency of the alpha rhythm. The light was periodically blocked so it did not reach the eye of the subject. Thus the entire EEG record was split up into several lengths of spontaneous and driven nature. Each of these lengths was about 25 seconds and so contained approximately 250

periods of the alpha frequency. The entire record can be split up arbitrarily into 'training' and 'test' samples.

The continuous time signals were first recorded on FM tape and then digitised. The sampling rate was chosen carefully to avoid any distortion due to 'aliasing' [25]. The sampling interval was taken as one millisecond. The 'folding frequency' [25] would then be 500 c/s and the frequency band of interest (say 0 - 20 c/s) is far below it. There would be 100 sampled values in every period of the alpha frequency.

CHAPTER III

FEATURE REDUCTION BASED ON A DISTANCE MEASURE

In this chapter the sampled values of the EEG will be treated as features in a linear classification scheme. Section 3.1 deals with the case where a classification is sought on the basis of a length of EEG record equal to one period of the alpha frequency, that is, on the basis of 100 features (remembering that one period of the alpha frequency is appriximately 100 milliseconds and that the sampling interval is one millisecond). An algorithm, based on a statistical distance measure between the two pattern classes, is developed to pick out significant ones among the 100 features. A feature of the algorithm is that the computations involved in the feature reduction also yield the optimal linear separating surface. However, it is found that error rates of about 20% occur because of the inherent overlap of the two pattern classes in 100 dimensional space. Section 3.2 extends the algorithm to the case where the classification is based on a length of EEG record which is a multiple of the period of the alpha frequency. It is shown that by basing the decisions on a sufficient length of EEG (20 - 50 periods of the alpha frequency) the error rate can be brought down to 1% or less. In the following, the word <u>response</u> connotes the length of the EEG record between two successive stroboscopic flashes, whether or not they are seen by the

subject.

3.1 Decision Based on a Single Response

3.1.1 The distance measure and its properties

One possible figure of merit which measures the effectiveness of any particular feature for discrimination purposes is the normalised square of the distance between the means (which is also a distance measure between the distributions of the two pattern classes)

$$d_i = \frac{[\mu_i^0 - \mu_i^1]^2}{\sigma_{ii}^0 + \sigma_{ii}^1} \tag{3.1}$$

where μ_i^j and σ_{ii}^j denote the mean and the variance of feature x_i under class j.

When we examine the combined effectiveness of a group of features, we have to take into account the correlations between features. The distance measure generalises to

$$d = (\underline{\mu}^0 - \underline{\mu}^1)^T (\Sigma^0 + \Sigma^1)^{-1} (\underline{\mu}^0 - \underline{\mu}_1) \tag{3.2}$$

where $\underline{\mu}^j$ and Σ^j are the mean vector and covariance matrix of the features under consideration for the distribution of pattern class j.

This distance measure, which is sometimes called the Mahalanobis D^2 statistic [26], can be explicitly related

[27, 28] to the Bayes error if the pattern classes are Gaussian with equal covariance matrices, $N(\underline{\mu}^0, \Sigma)$ and $N(\underline{\mu}^1, \Sigma)$ respectively. The Bayes separating surface, assuming equal prior probabilities and equal costs of misclassification, is

$$\log L(\underline{x}) = -\frac{1}{2}(\underline{x}-\underline{\mu}^0)^T\Sigma^{-1}(\underline{x}-\underline{\mu}_0) + \frac{1}{2}(\underline{x}-\underline{\mu}^1)^T\Sigma^{-1}(\underline{x}-\underline{\mu}^1) = 0$$

or

$$-(\underline{\mu}^0 - \underline{\mu}^1)^T\Sigma^{-1}\underline{x} + \frac{1}{2}\underline{\mu}^{0\,T}\Sigma^{-1}\underline{\mu}^0 - \frac{1}{2}\underline{\mu}^{1\,T}\Sigma^{-1}\underline{\mu}^1 = 0.$$

The decision rule is to classify $\underline{x}$ under H^0 if the left hand expression is positive and under H^1 if the left hand expression is negative. The expression, being a linear transformation of a Gaussian vector, is itself a Gaussian scalar. It is easily seen that if $\underline{x} \in H^i$ the expression is $N(m_i, \sigma)$, where

$$m_0 = \frac{1}{2}(\underline{\mu}^0 - \underline{\mu}^1)^T\Sigma^{-1}(\underline{\mu}^0 - \underline{\mu}^1) = d,$$

$$m_1 = -\frac{1}{2}(\underline{\mu}^0 - \underline{\mu}^1)^T\Sigma^{-1}(\underline{\mu}^0 - \underline{\mu}^1) = -d,$$

$$\sigma = (\underline{\mu}^0 - \underline{\mu}^1)^T\Sigma^{-1}(\underline{\mu}^0 - \underline{\mu}^1) = 2d.$$

Hence the Bayes error is given by

$$P_e = \frac{p(H^0)}{\sqrt{2\pi\sigma}}\int_{-\infty}^{0}\exp -\frac{1}{2\sigma}(y-m_0)^2\,dy + \frac{p(H^1)}{\sqrt{2\pi\sigma}}\int_{0}^{\infty}\exp -\frac{1}{2\sigma}(y-m_1)^2\,dy$$

$$= \frac{p(H^0)}{\sqrt{2\pi}} \int_{-\infty}^{-\frac{m_0}{\sqrt{\sigma}}} \exp - \frac{1}{2} y^2 dy + \frac{p(H^1)}{\sqrt{2\pi}} \int_{-\frac{m_1}{\sqrt{\sigma}}}^{\infty} \exp - \frac{1}{2} y^2 dy$$

$$= \frac{1}{\sqrt{2\pi}} \int_{\sqrt{\frac{1}{2}d}}^{\infty} \exp \left(- \frac{1}{2} y^2\right) dy$$

from which it is seen that P_e is a monotonically decreasing function of d. Hence maximisation of d ensures minimisation of P_e.

If the pattern classes are Gaussian with unequal covariance matrices, $N(\mu^0, \Sigma^0)$ and $N(\mu^1, \Sigma^1)$ respectively, it can be shown that the distance measure can be related to the B-distance of 1.2.4 in the following way.

$$B = \frac{1}{4} d + \frac{1}{2} \ln \frac{\det \left(\frac{\Sigma^0 + \Sigma^1}{2}\right)}{\sqrt{\det \Sigma^0 \cdot \det \Sigma^1}}.$$

3.1.2 Principle on which algorithm is based

If we let $\underline{\mu} = \underline{\mu}^0 - \underline{\mu}^1$ and $\Sigma = \Sigma^0 + \Sigma^1$ equation (3.2) can be written as

$$d = \underline{\mu}^T \Sigma^{-1} \underline{\mu}. \qquad (3.3)$$

Equation (3.3) suggests a sequential algorithm for feature selection. At each step, the proposed algorithm would choose the feature which leads to a maximum increase in the distance measure. If n features have already been chosen,

$$d_n = \underline{\mu}^T \Sigma^{-1} \underline{\mu}$$

where $\underline{\mu}$ is a n X 1 vector of the difference in means and Σ is the n X n matrix of the sum of the covariances. If a new feature x_{n+1} is added,

$$d_{n+1} = (\underline{\mu}^T \mu_{n+1}) \left(\begin{array}{c|c} \Sigma & C \\ \hline C^T & \sigma_{n+1} \end{array}\right)^{-1} \left(\begin{array}{c} \underline{\mu} \\ \hline \mu_{n+1} \end{array}\right)$$

where μ_{n+1} is the scalar difference in means for x_{n+1}, σ_{n+1} is the scalar variance of x_{n+1} and C is a n X 1 vector whose components are the covariances between the new feature x_{n+1} and the old features x_1 through x_n. Applying the Frobenius inversion formula* [29] it is easy to see that

$$d_{n+1} - d_n = (\underline{\mu}^T \mu_{n+1}) \left(\begin{array}{c|c} \Sigma^{-1} + \dfrac{\Sigma^{-1} C C^T \Sigma^{-1}}{\sigma_{n+1} - C^T \Sigma^{-1} C} & -\dfrac{\Sigma^{-1} C}{\sigma_{n+1} - C^T \Sigma^{-1} C} \\ \hline -\dfrac{C^T \Sigma^{-1}}{\sigma_{n+1} - C^T \Sigma^{-1} C} & \dfrac{1}{\sigma_{n+1} - C^T \Sigma^{-1} C} \end{array}\right) \left(\begin{array}{c} \underline{\mu} \\ \hline \mu_{n+1} \end{array}\right) - \underline{\mu}^T \Sigma^{-1} \underline{\mu}$$

$$= \frac{(\mu_{n+1} - C^T \Sigma^{-1} \underline{\mu})^2}{\sigma_{n+1} - C^T \Sigma^{-1} C}.$$

---

*Let $P = \left[\begin{array}{c|c} A & B \\ m \text{ X } m & m \text{ X } n \\ \hline C & D \\ n \text{ X } m & n \text{ X } n \end{array}\right]$.

Then $P^{-1} = \begin{bmatrix} R + RB\Delta^{-1}CR & -RB\Delta^{-1} \\ -\Delta^{-1}CR & \Delta^{-1} \end{bmatrix}$

where $\Delta = D - CRB$ and $R = A^{-1}$.

The feature which maximises $d_{n+1} - d_n$ is chosen as the next feature to be included.

The following facts emerge from this expression for the increase in the distance measure. They are proved in the Appendix.

i) Since any covariance matrix is at least positive semi-definite, it can be proved that

$$\sigma_{n+1} - C^T \Sigma^{-1} C \geq 0$$

which implies

$$d_{n+1} - d_n \geq 0.$$

Therefore, bringing in an additional feature can never worsen the discriminating capacity of the features already chosen.

ii) The increase in distance is zero if the new feature x_{n+1} is a linear combination of the old features (x_1, x_2, . . . , x_n).

iii) The increase in distance is infinite if the two classes are singularly distributed on separate hyperplanes in (n + 1) dimensional space.

iv) Since

$$\mu_{n+1} - C^T \Sigma^{-1} \underline{\mu} = E[x_{n+1} \mid x_1 = x_2 = \cdots = x_n = 0]$$

and

$$\sigma_{n+1} - C^T\Sigma^{-1}C = \text{Var}\,[x_{n+1} | x_1, x_2, \ldots, x_n]$$

it is seen that the algorithm chooses the (n + 1)th feature in the same manner as it chooses the first one, except that the mean and the variance now refer to the conditional distribution of x_{n+1}, given that the previous features x_1 through x_n are all zero.

3.1.3 Details of the algorithm

i) Using a sufficiently large training set, estimates of the N X 1 vector of the difference in means, and the N X N matrix of the sum of the covariances are obtained.

ii) The first feature x_1 is chosen such that

$$\frac{\mu_1^2}{\sigma_{11}} = \max_j \frac{\mu_j^2}{\sigma_{jj}}$$

iii) At each subsequent step, the increase in the distance measure $(d_{n+1} - d_n)$ is computed for each of the remaining features. The feature which gives rise to the maximum increase is chosen.

iv) Having chosen the best feature, the inverse covariance matrix Σ^{-1} is updated according to the Frobenius inversion formula.

v) The best separating hyperplane in the subspace spanned by the features chosen so far is

$$\underline{\alpha}_{opt.}^{T} \underline{x} + \alpha_o = 0$$

where

$$\underline{\alpha}_{opt.} = (\Sigma^o + \Sigma^1)^{-1} (\underline{\mu}^o - \underline{\mu}^1)$$

$$\alpha_o = -\frac{1}{2} (\underline{\mu}^o + \underline{\mu}^1)^T (\Sigma^o + \Sigma^1)^{-1} (\underline{\mu}^o - \underline{\mu}^1).$$

The weighting vector $\underline{\alpha}_{opt.}$ and the threshold α_o are computed. (The optimality of this hyperplane is discussed in 3.1.4.)

vi) The process is stopped either after all features have been exhausted or when the increase in distance is smaller than a preset value.

3.1.4 Optimality of the algorithm

As pointed out in 1.2.5, it is an inherent characteristic of the feature reduction problem that no sequential algorithm can be truly optimal. This is so because the best subset of n features is not necessarily a subset of the best subset of (n + 1) features. Only by an exhaustive search of all $\binom{N}{n}$ possible feature combinations, at each step, one can construct a truly optimal scheme; however, such an exhaustive search rapidly becomes infeasible as the total number of features increases.

Subject to this qualification, the algorithm is optimal in the sense that, at each step, it picks the particular feature which adds most to the effectiveness of the feature

set already chosen. Moreover, the particular weighting vector $\alpha_{opt.}$ given by equation (4) maximises the Fischer criterion [30], which is expressed by

$$\frac{[\underline{\alpha}^T(\underline{\mu}^o - \underline{\mu}^1)]^2}{\underline{\alpha}^T(\Sigma^o + \Sigma^1)\underline{\alpha}}$$

and interpreted as the ratio of the interclass distance to the intraclass dispersion along the direction $\underline{\alpha}$.

3.1.5 Results

A training set of 2000 responses of known classification, roughly equally divided between the two pattern classes, was used to compute the mean vectors and covariance matrices. The separating surfaces given by the algorithm were tested against a test set of 1000 responses. Fig. 3.1 shows how the actual error rate comes close to the predicted error rate based on a Gaussian model. This fact also seems to justify the stationarity assumption implicit in this approach, since the training and test sets were separated by about two minutes of EEG record.

It is also clear that decisions based on a small number of important features do almost as well as those based on a much larger number of features. This leads to the conclusion that discriminatory information is contained in the low frequencies (say 0 - 20 c/s). The best two features correspond to the positive and negative peaks of the average

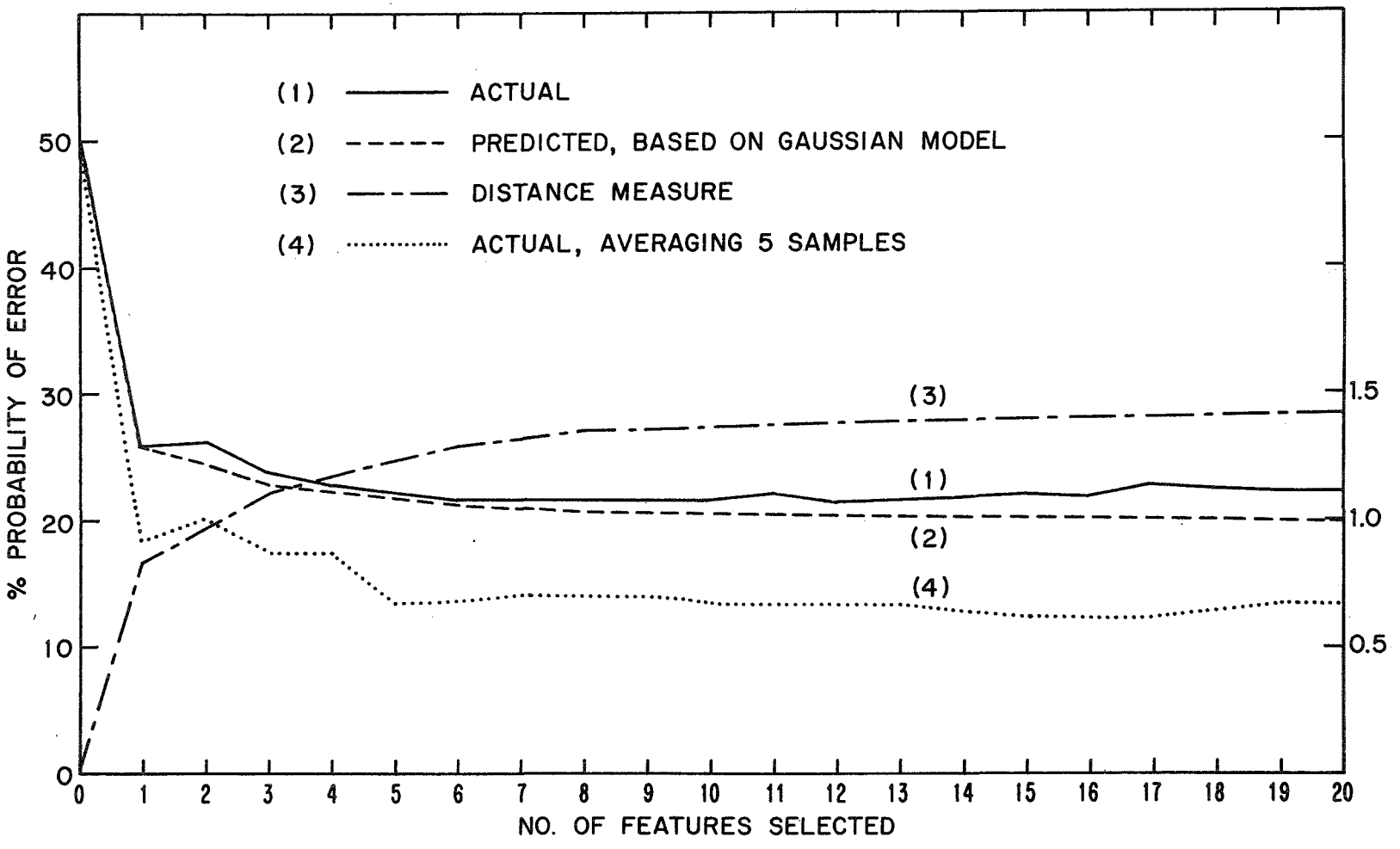

(1) ACTUAL
(2) PREDICTED, BASED ON GAUSSIAN MODEL
(3) DISTANCE MEASURE
(4) ACTUAL, AVERAGING 5 SAMPLES
% PROBABILITY OF ERROR
NO. OF FEATURES SELECTED
0
10
20
30
40
50
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
1.5
1.0
0.5
(1)
(2)
(3)
(4)

FIG 3-1

evoked response.

The error rate cannot be brought below about 20% by decisions based on a single response alone. Fig. 3.2, which is a plot of the patterns in the two-dimensional subspace spanned by the best two features, explains why this is so. There is considerable overlap between the two classes as a result of the large variances of the features within each pattern class.

3.2. Decisions Based on More Than One Response

3.2.1 Effect of correlations between responses

We have seen that the poor error rate in recognition results from the large deviations of the individual responses of each type from the average response of the same type. One way of circumventing this difficulty would be to average several statistically independent responses known to belong to the same pattern class; the decision would be based on the average response. It is well known that averaging K statistically independent random variables reduces the variance by a factor of K.

The implementation of the averaging operation would be rather difficult. This is so because the responses to be averaged will have to be picked out at random points in time from the EEG record, if they are to be statistically independent. In other words, decisions cannot be made in real time.

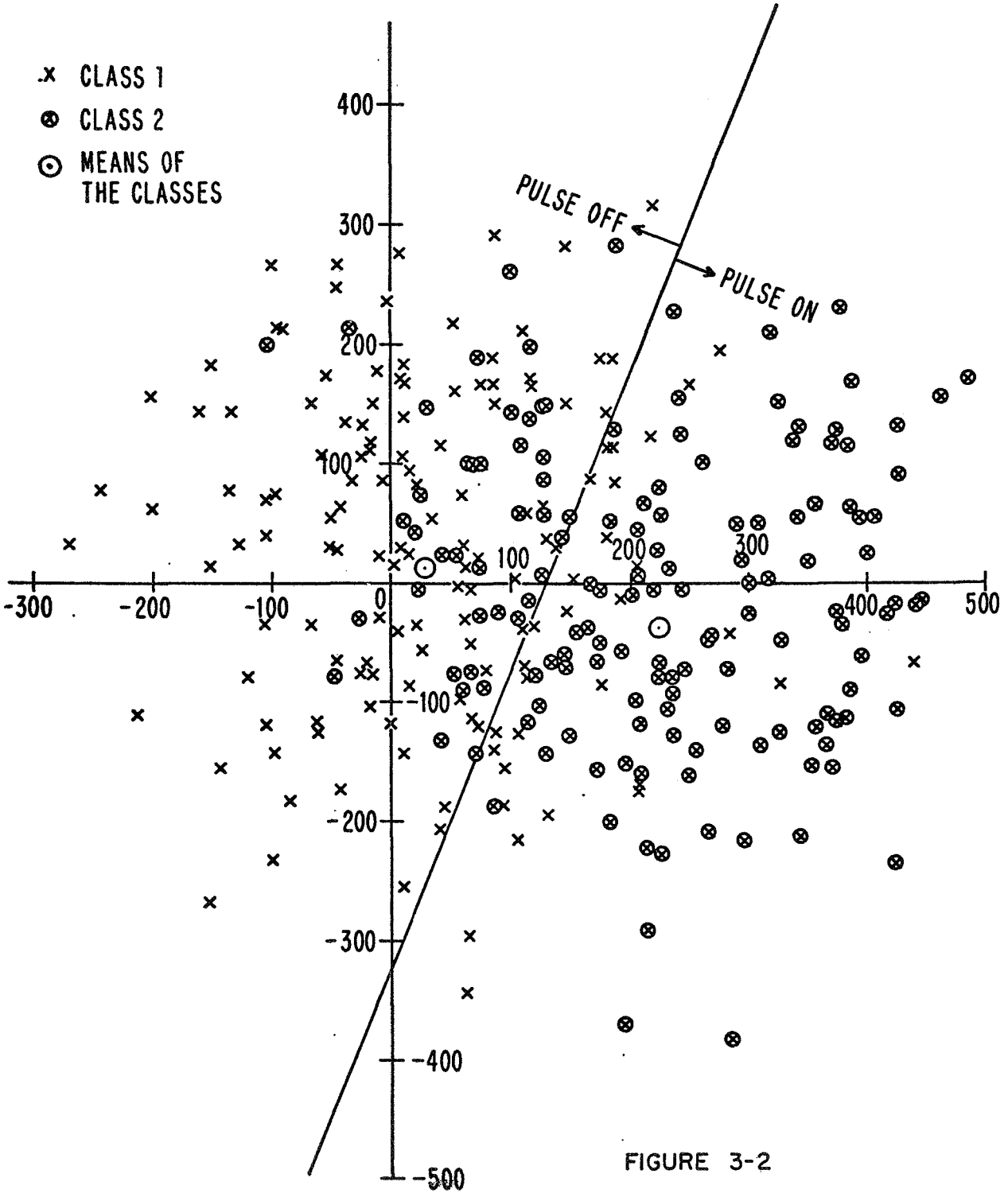

CLASS 1
CLASS 2
MEANS OF THE CLASSES
PULSE OFF
PULSE ON
400
300
200
100
-100
-200
-300
-400
-500
-300
-200
-100
0
100
200
300
400
500

FIGURE 3-2

Real time operation would be a very desirable characteristic of any practical scheme.

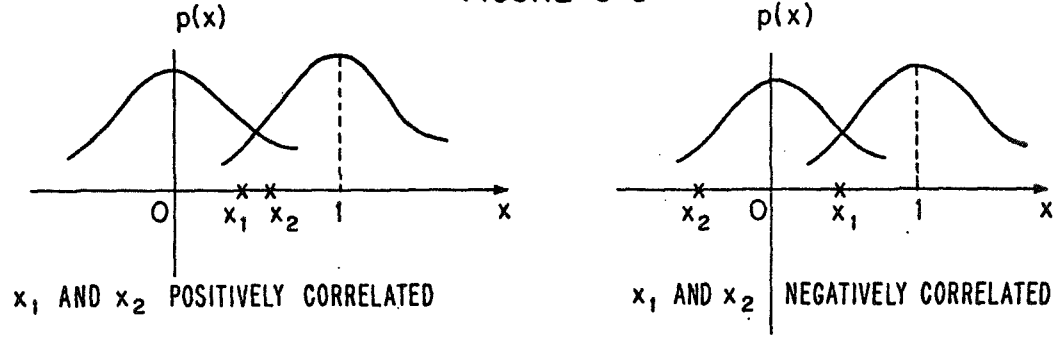
If the decision is to be based on several consecutive responses on the EEG record, we cannot afford to disregard the correlations between them. A simple example in Fig. 3.3 will illustrate the effects of positive and negative correlations upon the decision-making process. Let us consider the case where the two pattern classes are normally distributed N(0, 1) and N(1, 1) respectively. x_1 and x_2 are two successive measurements from the same class. It is obvious that if x_1 and x_2 are positively correlated, two measurements are not much better than one for discrimination; whereas, if x_1 and x_2 are negatively correlated, they tend to lie on either side of the mean and contain much greater information than x_1 alone.

3.2.2. The algorithm

The algorithm is essentially the same as in the single response case, except that the available set of features now extends over K consecutive responses. Let these be denoted by $\underline{x}(1), \underline{x}(2), \ldots, \underline{x}(K)$, each being a N-dimensional vector. The NK X 1 vector of the difference in means is

$$\begin{pmatrix} \underline{\mu}(1) \\ \underline{\mu}(2) \\ \vdots \\ \underline{\mu}(K) \end{pmatrix}$$

FIGURE 3-3


p(x)
0
x_1
x_2
1
x
x_1 AND x_2 POSITIVELY CORRELATED
p(x)
x_2
0
x_1
1
x
x_1 AND x_2 NEGATIVELY CORRELATED

where $\underline{\mu}(i) = E_{H^0}[\underline{x}(i)] - E_{H^1}[\underline{x}(i)]$. The NK X NK matrix of the sum of covariances is

$$\begin{pmatrix} C_{11} & C_{12} & \cdots & C_{1K} \\ C_{21} & C_{22} & \cdots & C_{2K} \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ C_{K1} & C_{K2} & \cdots & C_{KK} \end{pmatrix}$$

where each C is a N X N matrix and

$$C_{ij} = E_{H^0}[\underline{x}(i) - \underline{\mu}(i)][\underline{x}(j) - \underline{\mu}(j)]^T + E_{H^1}[\underline{x}(i) - \underline{\mu}(i)][\underline{x}(j) - \underline{\mu}(j)]^T.$$

It is reasonable to assume stationarity in the sense that

$$\underline{\mu}(1) = \underline{\mu}(2) = \cdots = \underline{\mu}(K) = \underline{\mu}$$

and

$$C_{ij} = C_{|i-j|} \text{ if } i < j$$

$$= C^T_{|i-j|} \text{ if } i > j.$$

Now the distance measure becomes

$$d_{NK} = (\underline{\mu}^T \underline{\mu}^T \cdots \underline{\mu}^T) \begin{pmatrix} C_0 & C_1 & C_2 & \cdots & C_{K-1} \\ C_1^T & C_0 & C_1 & \cdots & C_{K-2} \\ C_2^T & C_1^T & C_0 & \cdots & C_{K-3} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ C_{K-1}^T & C_{K-2}^T & \cdots & & C_0 \end{pmatrix}^{-1} \begin{pmatrix} \underline{\mu} \\ \underline{\mu} \\ \cdot \\ \cdot \\ \underline{\mu} \end{pmatrix}.$$

It is easy to see that if all the C's except C_0 are zero (that is, if the K consecutive responses are all statistically independent),

$$d_{NK} = Kd_N$$

and the discriminatory information contained in K consecutive responses would be K times that contained in a single response. However, in practice, because of the positive correlations that exist between consecutive responses

$$d_N < d_{NK} < Kd_N.$$

The algorithm naturally favors those features which are least positively correlated with the ones already chosen. The significant features are picked out from one response at a time as follows. The single-response algorithm is applied to $\underline{x}(1)$ and the optimum linear combination

$$y_1 = \underline{\alpha}^T_{opt.}(1)\underline{x}(1)$$

is obtained. y_1 is now combined with $\underline{x}(2)$ to yield a $(N + 1)$-dimensional pattern vector $(y_1, \underline{x}(2))$ for which the vector of the difference in means is $(\underline{\alpha}^T_{opt.}(1)\underline{\mu}, \underline{\mu})$ and the matrix of the sum of covariances is

$$\left(\begin{array}{c|c} E[\underline{\alpha}^T(1)\underline{x}(1)]^2 & E[\underline{\alpha}^T(1)\underline{x}(1)\underline{x}^T(2)] \\ \hline E[\underline{x}(2)\underline{x}(1)^T\underline{\alpha}(1) & E[\underline{x}(2)\underline{x}(2)^T] \end{array}\right) = \left(\begin{array}{c|c} \underline{\alpha}^T(1)C_o\underline{\alpha}(1) & \underline{\alpha}^T(1)C_1 \\ \hline C_1^T\underline{\alpha}(1) & C_o \end{array}\right)$$

The single response algorithm is again applied to produce the best linear combination of the $(N+1)$ features. The process continues, the pattern vector always being of dimension $(N + 1)$.

3.2.3 Results

In the actual application, the responses were discretised into 20 values ($N = 20$) to facilitate storage of several correlation matrices ($K = 40$). Fig. 3.4 shows how the distance measure increases and predicted error rate (based on a Gaussian model) decreases as more responses are used to arrive at a decision. The actual error rate deviates somewhat from the predicted one (unlike in the single response case). Thus the Gaussian model may not be adequate to describe the joint distribution of several pattern vectors.

Fig. 3.5 compares the performance of this scheme to the simple averaging technique using the same number of responses and features. It is seen that the averaging method performs poorly for small numbers of responses,

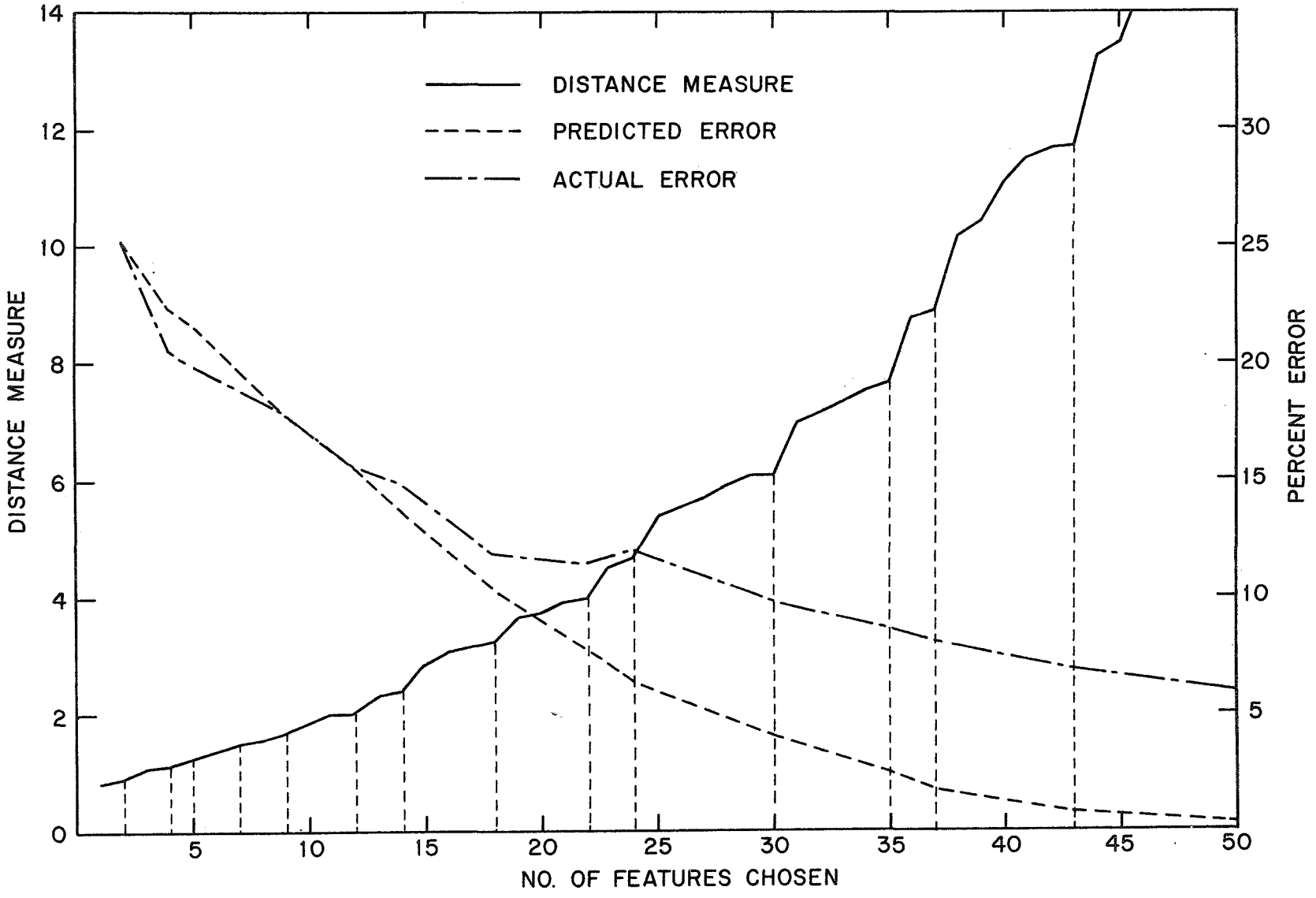

DISTANCE MEASURE
PREDICTED ERROR
ACTUAL ERROR
DISTANCE MEASURE
PERCENT ERROR
NO. OF FEATURES CHOSEN
0
2
4
6
8
10
12
14
5
10
15
20
25
30
35
40
45
50
5
10
15
20
25
30

FIG 3-4

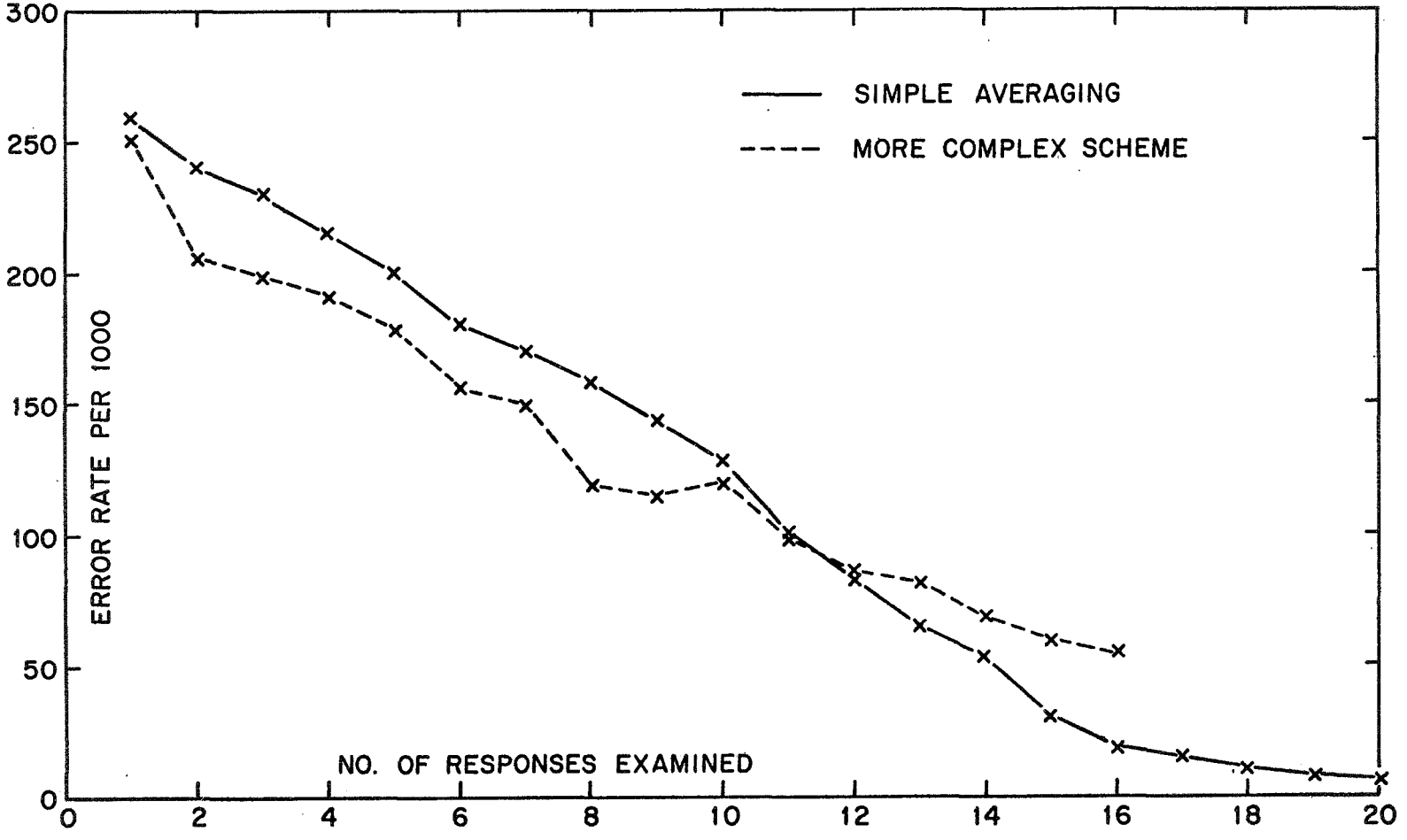

SIMPLE AVERAGING
MORE COMPLEX SCHEME
ERROR RATE PER 1000
NO. OF RESPONSES EXAMINED
0
50
100
150
200
250
300
0
2
4
6
8
10
12
14
16
18
20

FIGURE 3-5

but does better than the complex scheme when the number of samples averaged exceeds 12. The explanation for this phenomenon perhaps lies in the assumptions underlying the two schemes. The averaging method requires that the density function $p[\underline{x}(i)]$ be independent of i. The more complex scheme requires that the joint density function of K responses, $p[\underline{x}(i), \underline{x}(i + 1), \ldots, \underline{x}(i + K - 1)]$, be independent of i, which is certainly a more stringent assumption (stationarity of higher order) which is less likely to be satisfied in practice. What is more, the order of stationarity required increases as more responses are used to make a decision, and some deterioration in performance is only to be expected.

CHAPTER IV

THE RANDOM PROCESS MODEL

4.1 A Preview

This section is written especially for those readers who are interested in the EEG from a physiological point of view. Only an intuitive feel of some concepts associated with a random process, such as mean, variance, correlation and power spectrum, is assumed on the part of the reader. No mathematical details are given. Such readers are advised to proceed directly to Section 4.5 after reading through this section. One who is mathematically inclined can find all the details in Sections 4.2, 4.3 and 4.4.

As pointed out in Section 1.4, the modelling effort is motivated by the hope that by putting some structure into the EEG data, more efficient feature reduction and classification techniques might be developed. The EEG naturally requires a statistical description, since it indicates the combined activity of a large population of cells in the brain. Technically, a given EEG record should be considered as a sample obtained from an underlying random process, and efforts should be directed towards modelling this underlying process. Any model developed should adequately explain observed statistical properties of the spontaneous EEG, such as the behavior of the average signal and the autocorrelogram.

It should also explain the spectral properties of the spontaneous EEG, particularly the concentration of power in a narrow band around the alpha frequency. Besides, the model should be capable of being modified to explain the change in the statistical and spectral properties when the EEG is driven by repetitive stimuli at the alpha frequency. The statistical and spectral attributes are singled out because they are common to a large number of subjects. The statistical model would not be expected to explain peculiarities which might be observed only infrequently.

It is necessary to examine some of the models proposed in the literature for spontaneous EEG. A straightforward approach is to model the spontaneous EEG signal as a narrow band Gaussian process. This model implies that the sampled values of the signal obey a multivariate normal distribution in the statistical sense. Also, the correlation between any two sampled values as a function of the interval between the sampling instants should be such that the random process has a narrow band spectrum around a central frequency. Such a model can be termed 'linear' because it can be obtained by linear operations (passing through a linear network in a physical sense) on the most elementary random process, namely, the white Gaussian noise (that is, a Gaussian process with an ideally flat spectrum). The reverse is also true; it is possible to extract Gaussian white noise from a Gaussian narrow band process by a linear transformation.

A more sophisticated approach is to perform an amplitude and phase analysis of the EEG and try to give statistical descriptions of these parameters. This approach is consistent with the physically appealing assumption that the brain contains a large number of 'oscillators,' all of them at the alpha frequency, but with different phase relationships. The phase process is usually modelled as a linear transformation of white Gaussian noise. The statistics of the amplitude are not very important, especially for classification purposes, as will be seen in Chapter V. The model is termed 'nonlinear' because of the nonlinear relationship between the signal and its phase.

To the author's knowledge, there has not been an attempt to extend any of the above models to the case of the EEG driven by repetitive stimuli at the alpha frequency. The proposed nonlinear model, which bridges this gap, is as follows.

Let the EEG signal be represented by $\sin(\omega_a t + \theta(t))$, where the total phase is made up of two parts--(i) a non-random part $\omega_a t$ due to the alpha frequency $f_a = \frac{\omega_a}{2\pi}$, and (ii) a random part $\theta(t)$. If the random part $\theta(t)$ were zero, the total phase would increase linearly at a rate determined by the alpha frequency and the signal would be a pure sinusoid at the alpha frequency. The effect of the randomness expressed by $\theta(t)$ is to make the signal deviate from its pure sinusoidal shape and make it look like a random signal. The

amplitude is assumed to be constant for simplicity.

Let us now turn to the statistics of the random process $\theta(t)$. In the case of the spontaneous EEG, $\theta(t)$ is modelled as a zero mean Brownian process. Now the Brownian process is one of the most important physical random processes and has been studied in great detail. Imagine a particle which can move along a straight line and which is initially at the origin. At times Δt, $2\Delta t$, $3\Delta t$, . . . let the particle be given displacements x_1, x_2, x_3, . . . which are independent and normally distributed with zero mean and variance σ. After k such displacements the position of the particle is given by $x_1 + x_2 + \cdots + x_k$ which has zero mean and variance $k\sigma$. Thus it is seen that, even though the mean position of the particle is always at the origin, the uncertainty in its position, as expressed by its variance, increases linearly in time. As $\Delta t \to 0$, we can imagine the particle being given continuous displacements and its position $x(t)$ is then defined to be a zero mean Brownian process. In general, the initial position may also have an uncertainty about the origin; its variance then merely adds to the variance of the process at any instant. The Brownian process can be obtained by passing white Gaussian noise through an integrating network.

Reverting to the EEG signal, the effect of a Brownian component in the total phase of the signal is to make the uncertainty in the deviation from pure sinusoidal behavior

increase linearly in time. This is seen clearly in Fig. 4.1, which shows the actual phase histories of several short lengths of spontaneous EEG. (The precise method of determining the phase is described in the next chapter.) The uncertainty in the phase manifests itself as a dispersion of phase values around the mean values represented by the straight line corresponding to the alpha frequency. The standard deviation of $\theta(t)$ increases as $\sqrt{t}$, which explains the curvature in the lines showing the $1-\sigma$ bounds of phase dispersion.

The restoration of phase coherence by repetitive stimuli at the alpha frequency can be explained as follows. The component $\theta(t)$ would still be a Brownian process between any two successive stimuli. The stimuli, however, have the effect of restoring the uncertainty to its initial value. In effect, $\theta(t)$ is split into several Brownian processes, which, however, are assumed to be correlated among themselves. The correlations are assumed to fall off in a geometrical fashion. Fig. 4.2, obtained from an actual EEG record with repetitive stimuli, shows how the phase dispersion is controlled by the stimuli.

It is proved mathematically in Sections 4.3 and 4.4 that the statistical and spectral properties of the processes described above correspond to the observed properties of the two types of EEG considered. The reader who is not interested in mathematical details should proceed in Section 4.5, which compares the predicted and actual properties.

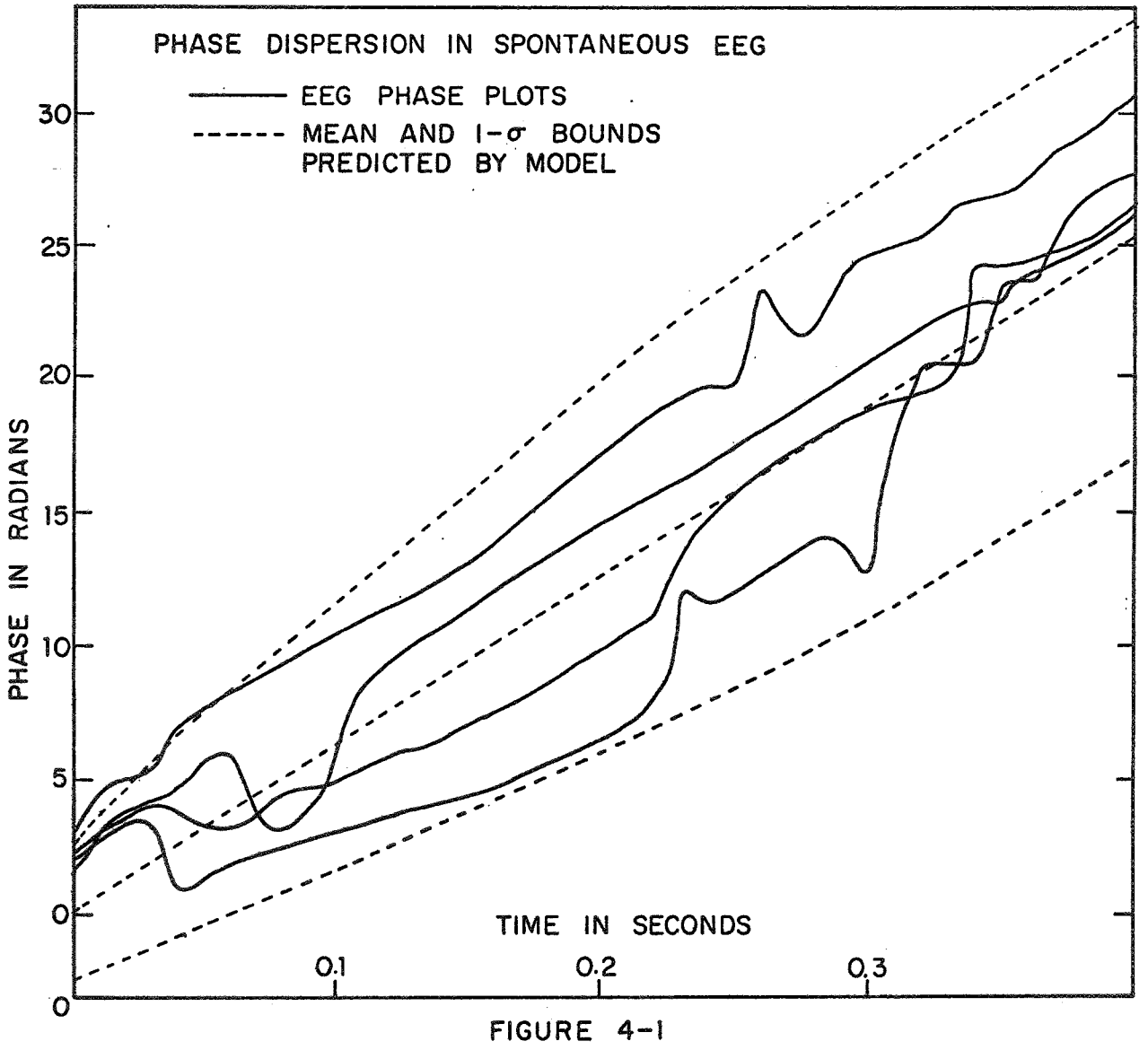

PHASE DISPERSION IN SPONTANEOUS EEG
EEG PHASE PLOTS
MEAN AND 1-σ BOUNDS
PREDICTED BY MODEL
PHASE IN RADIANS
30
25
20
15
10
5
0
0
TIME IN SECONDS
0.1
0.2
0.3

FIGURE 4-1

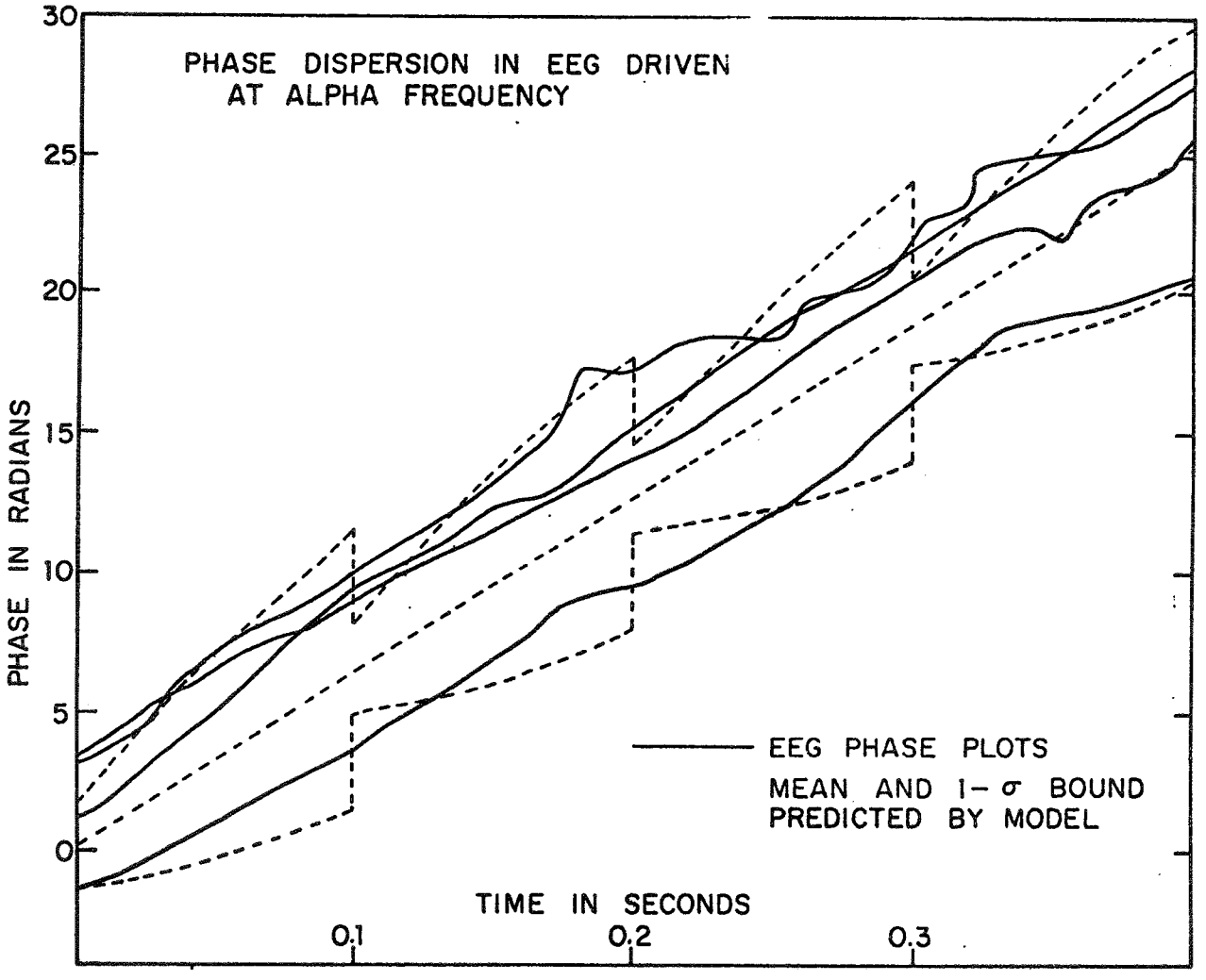

PHASE DISPERSION IN EEG DRIVEN
AT ALPHA FREQUENCY
30
25
20
15
10
5
0
PHASE IN RADIANS
EEG PHASE PLOTS
MEAN AND 1- σ BOUND
PREDICTED BY MODEL
TIME IN SECONDS
0.1
0.2
0.3

FIGURE 4-2

4.2 Details of Earlier Efforts in Spontaneous EEG Modelling

4.2.1 The linear model [19]

The EEG is modelled as a sample from a stationary Gauss-Markov process, which itself is modelled as the output of a linear dynamical system excited by white Gaussian noise. Under certain conditions, it can be proved [31] that the output of such a system is asymptotically a stationary, Gauss-Markov process. Moreover, such a process is uniquely determined by its first and second order statistical properties, namely, the mean, $\mu(t) = E\, x\,(t)$, and the auto-correlation function, $R(\tau) = E\, x\,(t)\, x\,(t + \tau)$. Now it is only a question of finding the proper dynamical system whose output has $\mu(t)$ and $R(\tau)$ similar to the spontaneous EEG average signal and autocorrelogram (shown in dotted lines in Figs. 2.3 and 2.4).

Considering a second order, underdamped dynamical system driven by a zero mean, white Gaussian noise,

$$\ddot{x} + 2\zeta\omega_n \dot{x} + \omega_n^2 x = \upsilon \qquad (\zeta < 1)$$

it can be proved that [31],

$$\lim_{t\to\infty} E\, x\,(t) = 0$$

$$\lim_{t\to\infty} E\, x\,(t)\, x\,(t + \tau) = Ae^{-\zeta\omega_n \tau}\cos\,[\omega_n\sqrt{1-\zeta^2} + \varphi].$$

Asymptotically the random process x(t) will approach a

narrow band Gaussian process. Now ζ and ω_n can be determined to give the exact behavior required.

4.2.2 Non-linear models

Wiener [32] postulated that the brain contains a large number of 'oscillators,' all at the basic frequency of the alpha rhythm, but with different phase relationships. If we denote the activity of a typical oscillator by $a_i \sin(\omega_a t + \theta_i)$, the combined EEG activity will be

$$x(t) = \sum_i a_i \sin(\omega_a T + \theta_i)$$

$$= A \sin(\omega_a t + \theta),$$

where

$$A^2 = \left|\sum_i a_i \cos\theta_i\right|^2 + \left|\sum_i a_i \sin\theta_i\right|^2 \quad \text{and}$$

$$\tan\theta = \frac{\sum_i a_i \sin\theta_i}{\sum_i a_i \cos\theta_i}.$$

It is obvious that only a statistical description can be given for a_i and θ_i, which in turn leads us to model A and θ as random variables. If, in addition, the phase relationships vary with time, A(t) and $\theta(t)$ can be modelled as stochastic processes. Thus,

$$x(t) = A(t) \sin(\omega_a t + \theta(t))$$

will be a process with random amplitude and phase modulation.

It is clear that the spectral behavior of x(t) near the alpha frequency ω_a depends critically on the statistics of the phase process $\theta(t)$. If there is any dominant frequency in A(t), it would tend to shift the peak in the EEG spectrum; but it would not materially affect the behavior of the spectrum near the peak. In the literature A(t) has been usually modelled as a constant.

Wiener [32] modelled the phase as a linear transformation of the Brownian motion process

$$\theta(t) = \epsilon \iint K(t + \tau_1, t + \tau_2)dz(\tau_1)dz(\tau_2)$$

where z(t) is a Brownian motion. Under the assumption that ϵ is small, he has derived the approximate spectrum of the EEG and showed that it is similar to the observed spectrum.

More recently, Joseph et al. [33] have modelled the frequency, rather than the phase, as a Gaussian random variable. They assume that the elementary oscillators are not all of the same frequency; instead, the distribution of frequencies is Gaussian around the alpha frequency ω_a. Therefore, the signal in complex form is

$$x(t) = Ae^{i\omega t} \qquad \text{where } \omega \sim N(\omega_a, \sigma^2).$$

Then,

$$E\,x\,(t) = AEe^{i\omega t}$$

$$= Ae^{i\omega_a t}e^{-\frac{1}{2}\sigma^2 t^2} \qquad \text{(from the characteristic function of a Gaussian variable)}$$

$$R_{xx}(\tau) = Ex(t)x^*(t+\tau)$$

$$= EA^2 e^{i\omega t}e^{-i\omega(t+\tau)}$$

$$= A^2 Ee^{-i\omega\tau}$$

$$= A^2 e^{-i\omega_a \tau}e^{-\frac{1}{2}\sigma^2\tau^2}.$$

The envelope of the average signal has a predicted e^{-kt^2} behavior under this model. The authors note that experimental results support e^{-kt} behavior. The power spectrum is

$$P(\omega) = \int_{-\infty}^{\infty} R_{xx}(\tau)e^{i\omega\tau}d\tau$$

$$\sim e^{-\frac{1}{2\sigma^2}(\omega-\omega_a)^2},$$

which is a Gaussian distribution of power around the alpha frequency.

4.3 Details of Proposed Model for Spontaneous EEG

The nonlinear model was chosen because it admits a natural modification for the case where the EEG is driven by repetitive stimuli at the alpha frequency. The phase is modelled as a simple Brownian motion, since it keeps the mathematical analysis simple and seems to explain experimental facts adequately. The amplitude is modelled as a constant. It will be shown in the next chapter that, under certain reasonable assumptions, the statistics of the amplitude are unimportant for classification purposes.

To keep the mathematics simple, the EEG signal is represented in the complex form in the analysis which follows. Some familiarity with characteristic functions is also assumed. A reader who so desires can read a short exposition of these concepts in the Appendix.

4.3.1 The model

We suppose that the (normalised) spontaneous EEG signal is represented in complex form as

$$x(t) = e^{i[\omega_a t+\theta(t)]}.$$

The random phase $\theta(t)$ is modelled by the Brownian process [34],

$$\dot{\theta} = w$$

where w is a zero mean Gaussian white noise with

$$Ew(t) = 0$$

$$Ew(t)w(\tau) = q\delta(t - \tau).$$

It remains to specify the initial uncertainty in θ. We suppose that $\theta(0)$ is also Gaussian with

$$E\theta(0) = 0$$

$$\text{Var } \theta(0) = p_o .$$

With these assumptions, $\theta(t)$ will be a Gauss-Markov random process with

$$E\theta(t) = 0$$

$$\text{Var } \theta(t) = p_o + qt$$

$$E\theta(t)\theta(\tau) = p_o + qt \quad \text{for } \tau > 0.$$

The variance of the phase increases linearly with time, implying gradual loss of phase coherence.

4.3.2 The predicted average signal

The average signal and the autocorrelogram, as computed from any sample of a random process, are themselves random functions. The model can only predict the expected

values and variabilities of these estimates.

The average signal, as computed from an EEG record whose length is equal to N periods of the alpha frequency, is

$$\mu_N(t) = \frac{1}{N} \sum_{k=0}^{N-1} x(kt_a + t) \qquad 0 \le t \le t_a .$$

Therefore,

$$E\mu_N(t) = \frac{1}{N} \sum_{k=0}^{N-1} Ee^{i\omega_a t + i\theta(kt_a + t)} \quad (\text{since } \omega_a t_a = 2\pi).$$

Now $\theta(t)$ is Gaussian with zero mean and variance $p_o + qt$, and hence has the characteristic function [35]

$$Ee^{i\theta\upsilon} = e^{-\frac{1}{2}(p_o + qt)\upsilon^2} .$$

Therefore, setting $\upsilon = 1$ in the above formula, we obtain for $E\mu_N(t)$

$$E\mu_N(t) = \frac{1}{N} \sum_{k=0}^{N-1} e^{i\omega_a t} e^{-\frac{1}{2}(p_o + qkt_a + qt)}$$

$$= A_N e^{-\frac{1}{2}(p_o + qt)} e^{i\omega_a t}$$

where $A_N = \frac{1}{N} e^{-\frac{1}{2}qt_a} \left(\frac{1 - e^{-\frac{1}{2}Nqt_a}}{1 - e^{-\frac{1}{2}qt_a}} \right)$

showing that the expected average signal is an exponentially decaying oscillation. Since A_N is $0(\frac{1}{N})$, the model predicts

that the average signal falls off to zero as greater lengths of EEG are used in computing it, which is supported by experiments. Also to be noted is the e^{-kt} behavior of the average signal, compared with the e^{-kt^2} behavior predicted by the model proposed by Joseph et al. [33]. The Gaussian assumption is perhaps more valid with the phase rather than the frequency.

The variance of the estimate is derived in the Appendix.

4.3.3 The predicted autocorrelogram

$$R_{xx}(t,\ t+\tau) = Ee^{i(\omega_a t+\theta(t))}e^{-i(\omega_a t+\omega_a \tau+\theta(t+\tau))}$$

$$= e^{-i\omega_a \tau}Ee^{i\theta(t)-i\theta(t+\tau)}.$$

Now $\theta(t)$ and $\theta(t+\tau)$ are jointly Gaussian with zero mean and covariance matrix (for $\tau \geq 0$)

$$\begin{bmatrix} p_o + qt & p_o + qt \\ p_o + qt & p_o + qt + q\tau \end{bmatrix}$$

Therefore, the joint characteristic function [35] of $\theta(t)$ and $\theta(t+\tau)$ is

$$Ee^{i\theta(t)\upsilon_1+i\theta(t+\tau)\upsilon_2} = e^{-\frac{1}{2}[(p_o+qt)\upsilon_1^2+2(p_o+qt)\upsilon_1\upsilon_2 + (p_o+qt+q\tau)\upsilon_2^2]}.$$

Setting $\upsilon_1 = 1$, $\upsilon_2 = -1$,

$$R_{xx}(t,\ t+\tau) = e^{-i\omega_a\tau}e^{-\frac{1}{2}[p_o+qt-2p_o-2qt+p_o+qt+q\tau]}$$

$$= e^{-i\omega_a\tau}e^{-\frac{1}{2}q\tau}$$

$$= R_{xx}(\tau).$$

The process x(t) is not stationary in the wide sense since E x (t) is not a constant, However, since the autocorrelation function is only a function of τ, the autocorrelogram $R_N(\tau)$ continues to be an unbiased estimate of $R_{xx}(\tau)$. Hence,

$$ER_N(\tau) = R_{xx}(\tau) = e^{-i\omega_a\tau}e^{-\frac{1}{2}q\tau}.$$

The variability of this estimate is derived in the Appendix.

4.4 Details of Proposed Model for EEG with Repetitive Stimuli at the Alpha Frequency

4.4.1 The model

As discussed in Chapter II, repetitive photic stimuli at the alpha frequency lead to recovery of phase coherence. The mathematical model can reflect this in the following way. Each stimulus, occurring at instants t_a, $2t_a$, $3t_a$, . . . is supposed to reset the variance of the phase (not the phase itself) to its initial value p_o. That is,

$$\text{var } \theta(0) = \text{var } \theta(t_a^+) = \text{var } \theta(2t_a^+) \cdots = p_o.$$

The dynamics of $\theta(t)$ between the photic stimuli is unaltered; $\theta(t)$ continues to be a zero mean Brownian process. In effect, the stimuli occurring at t_a, $2t_a$, . . . split the random process $\theta(t)$ into a vector of Brownian processes

$$\begin{pmatrix} \theta_1(t) \\ \theta_2(t) \\ \cdot \\ \cdot \\ \cdot \\ \theta_N(t) \end{pmatrix} \qquad \text{where } \theta_i(t) = \theta[(i-1)t_a + t].$$

The statistics of each component $\theta_i(t)$ are determined by the Brownian model as before, namely,

$$\dot{\theta}_i = w_i \qquad E\theta_i(0) = 0, \ \text{var } \theta_i(0) = p_0$$

$$Ew_i(t) = 0, \ Ew_i(t)w_j(\tau) = q\delta_{ij}\delta(t-\tau)$$

To specify the vector process completely, we also need the covariance matrix of the initial values. We assume

$$E\theta_i(0^+)\theta_j(0^+) = p_0\alpha^{|i-j|}.$$

The process thus described is periodically stationary with period t_a in the sense that the statistics of the process are invariant with respect to a translation by a multiple of

t_a.

$$p[x(t_1), x(t_2), \ldots, x(t_n)] = p[x(t_1 + kt_a), x(t_2 + kt_a), \ldots, x(t_n + kt_a)]$$

for any integer k.

4.4.2 The predicted average signal

Now the random variables $x(kt_a + t)$ for $k = 0, 1, 2, \ldots, N - 1$ have all the same distribution. Hence,

$$E\mu_N(t) = Ee^{i[\omega_a t+\theta(t)]}$$

which by similar reasoning as in section 4.3.2

$$= e^{i\omega_a t} e^{-\frac{1}{2}qt}.$$

Therefore, the expected average signal is an exponentially decaying oscillation, whose amplitude is independent of the length of the EEG record used in computing the average. This is borne out by experimental evidence. The variance of this estimate is derived in the Appendix.

4.4.3 The predicted autocorrelogram

$$R_{xx}(t, t + \tau) = Ex(t)x^*(t + \tau)$$

$$= Ee^{i[\omega_a t+\theta(t)]} e^{-i[\omega_a t+\omega_a \tau+\theta(t+\tau)]}$$

$$= e^{-i\omega_a \tau} E e^{i\theta(t)-i\theta(t+\tau)}.$$

Now $Ee^{i\theta(t)-i\theta(t+\tau)}$ depends on the relative positions of the instants t and $t + \tau$ on the EEG record. Without loss of generality, we can assume $0 \le t \le t_a$ because of the periodic stationarity of the process.

(i) If $0 \le t + \tau \le t_a$, then

$$E|\theta(t)|^2 = p_o + qt$$

$$E\theta(t)\theta(t + \tau) = p_o + qt$$

$$E|\theta(t + \tau|^2 = p_o + qt + q\tau.$$

Therefore, from the joint characteristic function of $\theta(t)$ and $\theta(t + \tau)$,

$$Ee^{i\theta(t)-i\theta(t+\tau)} = e^{-\frac{1}{2}(p_o+qt-2p_o-2qt+p_o+qt+q\tau)}$$

$$= e^{-\frac{1}{2}q\tau} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (4.1)$$

(ii) If $mt_a \le t + \tau \le (m + 1)t_a$ (where $m = 1, 2, 3, \ldots$)

$$E|\theta(t)|^2 = p_o + qt$$

$$E\theta(t)\theta(t + \tau) = E\theta(0^+)\theta(mt_a^+) = \alpha^m p_o$$

$$E|\theta(t + \tau)|^2 = p_o + q(t + \tau - mt_a).$$

Thus we have

$$Ee^{i\theta(t)-i\theta(t+\tau)} = e^{-\frac{1}{2}(p_o+qt-2\alpha^m p_o+p_o+qt+q\tau-qmt_a)}$$

$$= e^{-(1-\alpha^m)p_o} e^{-\frac{1}{2}q(\tau-mt_a)} e^{-qt} \quad . \quad . \quad . \quad (4.2)$$

and the autocorrelation function is no longer purely a function of the delay τ. However, the autocorrelogram defined as

$$R_T(\tau) = \frac{1}{T}\int_o^T x(t)x^*(t+\tau)dt$$

performs a smoothing operation with respect to t and gives the result purely as a function of the delay τ. In practice, the autocorrelogram is computed from a length of EEG record equal to N periods of the alpha frequency

$$R_N(\tau) = \frac{1}{Nt_a}\int_o^{Nt_a} x(t)x^*(t+\tau)dt.$$

The expected value of this estimate will now be derived.

(i) $0 \leq \tau \leq t_a$.

$$ER_N(\tau) = \frac{1}{Nt_a}\int_o^{Nt_a} Ex(t)x^*(t+\tau)dt.$$

The integral can be split up over ranges $(0, t_a - \tau)$, $(t_a - \tau, t_a)$, $(t_a, 2t_a - \tau)$, $(2t_a - \tau, 2t_a)$, . . . , $((n-1)t_a, nt_a - \tau)$, $(nt_a - \tau, nt_a)$. The alternate integrals are all

equal by periodic stationarity of the process. Therefore,

$$ER_N(\tau) = \frac{1}{t_a}\left[\int_0^{t_a-\tau} Ex(t)x^*(t+\tau)dt + \int_{t_a-\tau}^{t_a} Ex(t)x^*(t+\tau)dt\right].$$

In the first integral t and t + τ are such that equation (4.1) applies; whereas in the second integral t and t + τ are such that equation (4.2) applies with m = 1. Hence,

$$ER_N(\tau) = \frac{1}{t_a}\left[\int_0^{t_a-\tau} e^{-i\omega_a\tau}e^{-\frac{1}{2}q\tau}dt + \int_{t_a-\tau}^{t_a} e^{-(1-\alpha)p_0}e^{-qt}e^{-\frac{1}{2}q(\tau-t_a)}e^{-i\omega_a\tau}dt\right]$$

$$= \left[\left(1-\frac{\tau}{t_a}\right) + A_1 \sinh\left(\frac{1}{2}q\tau\right)\right]e^{-i\omega_a\tau}$$

where $A_1 = e^{-p_0(1-\alpha)}\cdot\left(\frac{e^{-\frac{1}{2}qt_a}}{qt_a}\right)$.

(ii) $kt_a \le \tau \le (k+1)t_a$.

Using periodic stationarity again, we can write

$$ER_N(\tau) = \frac{1}{t_a}\left[\int_0^{(k+1)t_a-\tau} Ex(t)x^*(t+\tau)dt + \int_{(k+1)t_a-\tau}^{t_a} Ex(t)x^*(t+\tau)dt\right].$$

Now in the first integral t and t + τ are such that equation (4.2) applies with m = k; whereas in the second integral the same equation applies with m = k + 1. Therefore,

$$ER_N(\tau) = \frac{1}{t_a}\left[\int_0^{(k+1)t_a-\tau} e^{-(1-\alpha^k)p_0} e^{-qt} e^{-\frac{1}{2}q(\tau-kt_a)} e^{-i\omega_a\tau} dt + \int_{(k+1)t_a-\tau}^{t_a} e^{-(1-\alpha^{k+1})p_0} e^{-qt} e^{-\frac{1}{2}q(\tau-(k+1)t_a)} e^{-i\omega_a\tau} dt\right]$$

$$= \left\{A_k \sinh\left[\frac{q}{2}\left((k+1)t_a - \tau\right)\right] + A_{k+1} \sinh\left[\frac{q}{2}\left(\tau - kt_a\right)\right]\right\} e^{-i\omega_a\tau}$$

where $A_j = e^{-(1-\alpha^j)p_0} \cdot \frac{e^{-\frac{1}{2}qt_a}}{qt_a}$.

As $k \to \infty$, A_k, $A_{k+1} \to A_\infty = e^{-p_0} \cdot \left(\frac{e^{-\frac{1}{2}qt_a}}{qt_a}\right)$. Therefore, the expected autocorrelogram is asymptotically periodic with period t_a. This should be contrasted with the exponentially decaying oscillatory behavior of the autocorrelogram of spontaneous EEG.

4.5 Comparison of Predicted and Actual Results

In this section, the validity of the model will be checked by comparing the behavior of actual EEG records with the behavior predicted by the model. One such comparison was already made in Section 4.1 where the phase dispersion

obtained in actual EEG records of the two types was matched against the variance predicted by the model. It is well known that a Gaussian random variable would lie, 68.3% of the time, within one standard deviation of its mean. This is the significance of the $1 - \sigma$ bounds in Figs. 4.1 and 4.2.

Now the EEG signal itself is a nonlinear function of its phase. Given the stochastic behavior of the phase, the stochastic behavior of the signal can be derived in principle. The mathematical analysis in Sections 4.3 and 4.4 was directed towards investigating the predicted behavior of two statistical attributes of the EEG signal, namely, the average signal and the autocorrelogram. (Or, equivalently, the power spectrum.) For those who did not wish to go into mathematical details, the results are briefly recaptured here. The figures show the actual behavior in solid lines and the predicted behavior in broken lines.

If the model holds exactly the following are the predicted results. The average signal for spontaneous EEG is an exponentially decaying oscillation, with its amplitude falling off as $\frac{1}{N}$ as more responses are used in computing it. (Fig. 4.3). With repetitive stimuli, the behavior is the same except that the amplitude becomes independent of N (Fig. 4.4). In other words, the average signal shows a distinct oscillatory behavior if sufficient responses are used in computing it.

The autocorrelogram of spontaneous EEG is

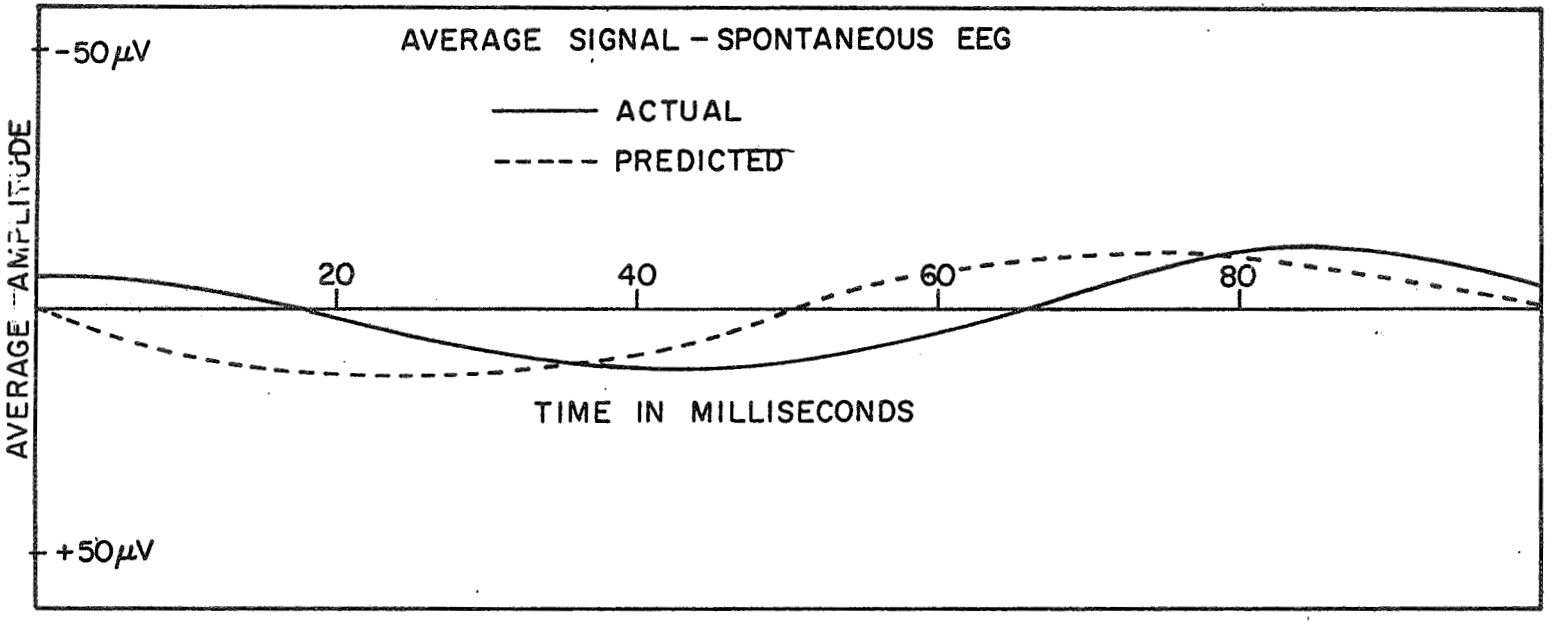

AVERAGE SIGNAL – SPONTANEOUS EEG
ACTUAL
PREDICTED
-50μV
+50μV
AVERAGE AMPLITUDE
20
40
60
80
TIME IN MILLISECONDS

FIGURE 4-3

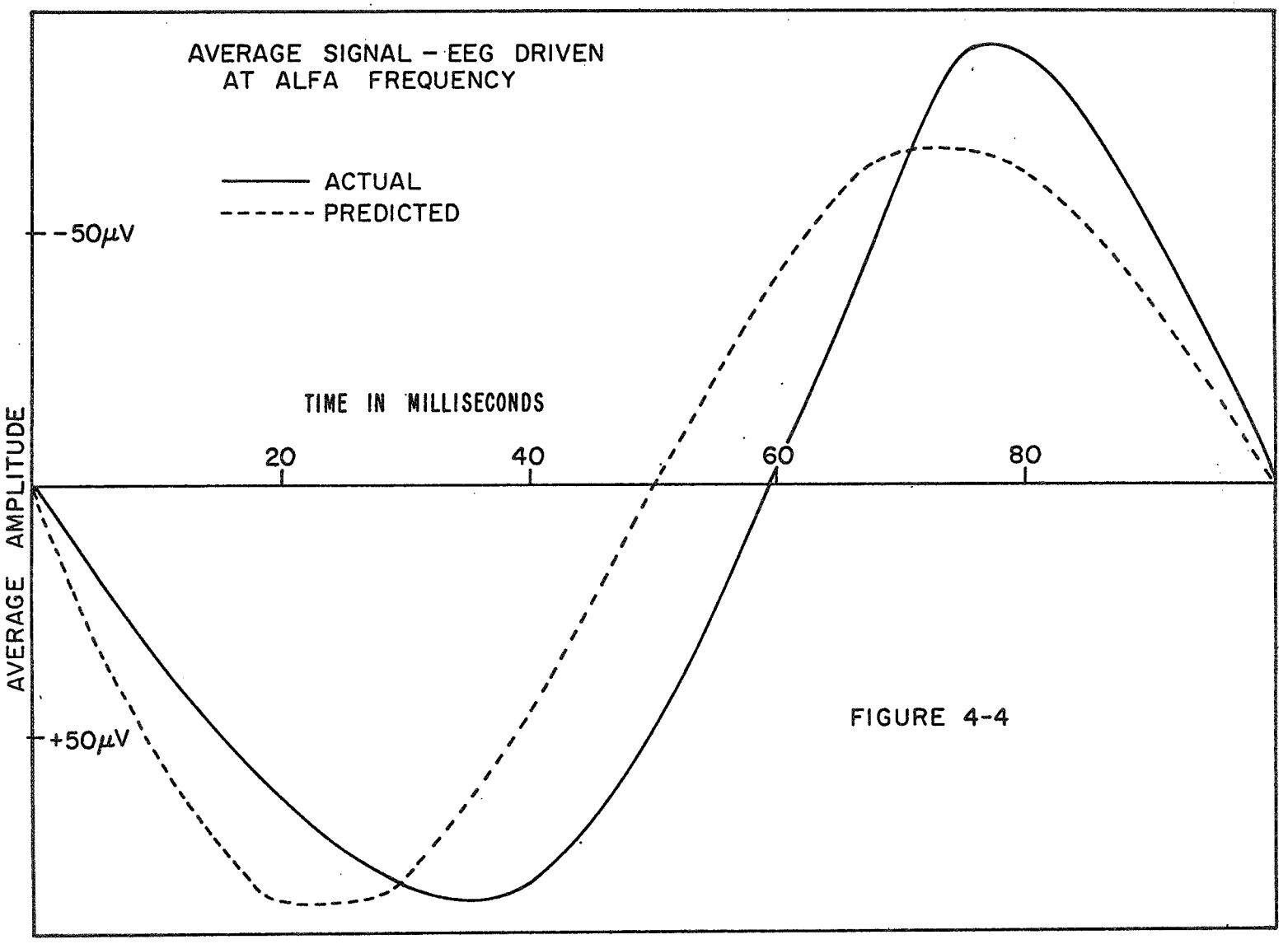

AVERAGE SIGNAL - EEG DRIVEN
AT ALFA FREQUENCY
ACTUAL
PREDICTED
-50μV
TIME IN MILLISECONDS
20
40
60
80
AVERAGE AMPLITUDE
+50μV

FIGURE 4-4

exponentially decaying at a rate proportional to the rate at which the variance of the phase is increasing, and oscillatory at the alpha frequency (Fig. 4.5). With repetitive stimuli, it is still oscillatory at the alpha frequency. However, the amplitude, though it decays for small values of the delay, soon reaches a steady asymptotic value. Thus the autocorrelogram asymptotically becomes periodic (Fig. 4.6).

The power spectrum of spontaneous EEG behaves as $\frac{1}{(\omega-\omega_a)^2+\frac{q^2}{4}}$ near the alpha frequency ω_a, showing that the bandwidth is equal to q, the rate at which the variance of the phase is increasing (Fig. 4.7). With repetitive stimuli, a spike will appear at the alpha frequency, because of asymptotic periodicity of the autocorrelogram (Fig. 4.8).

The figures show that the behavior of actual EEG plots is in good agreement with the predicted results.

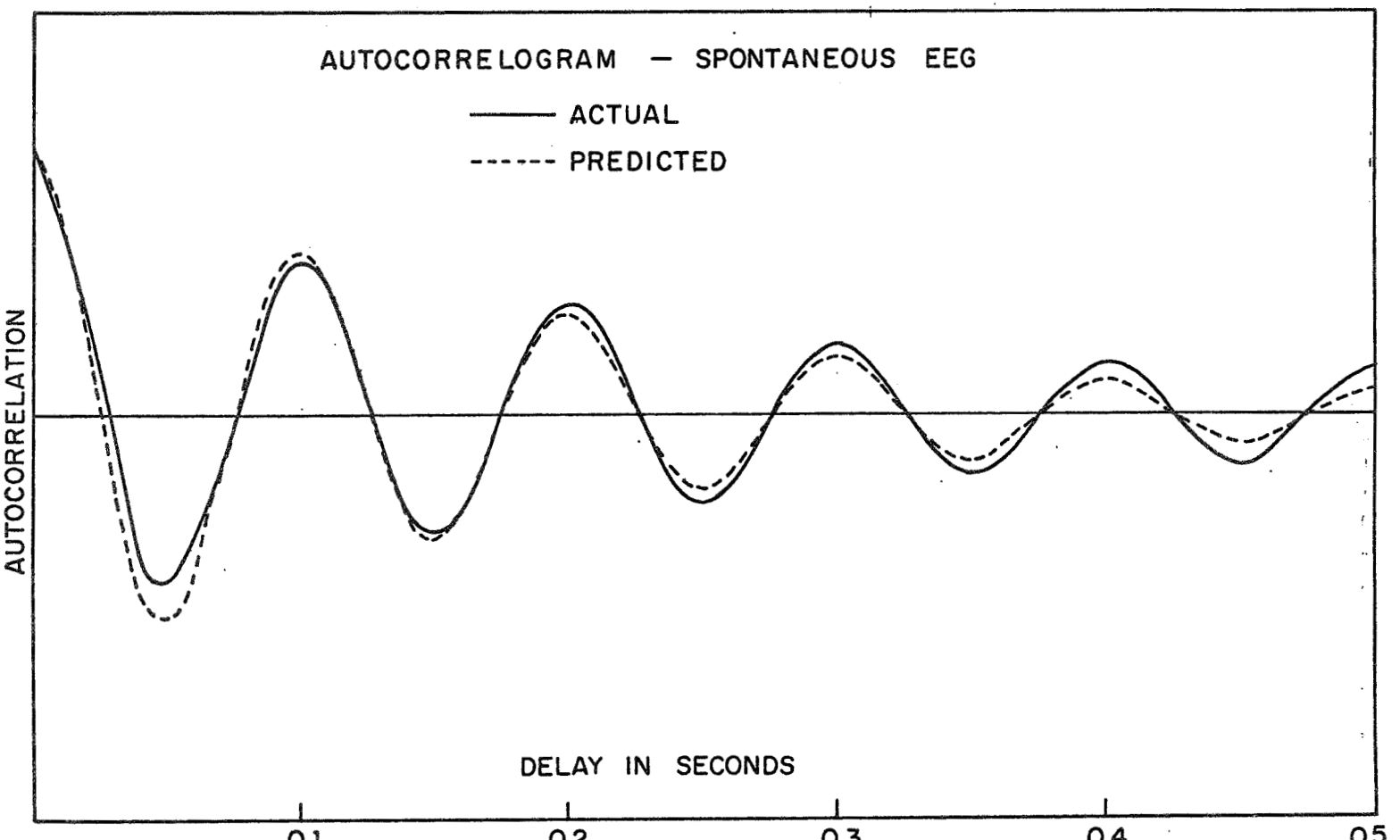

AUTOCORRELOGRAM — SPONTANEOUS EEG
ACTUAL
PREDICTED
AUTOCORRELATION
DELAY IN SECONDS
0.1
0.2
0.3
0.4
0.5

FIG. 4-5

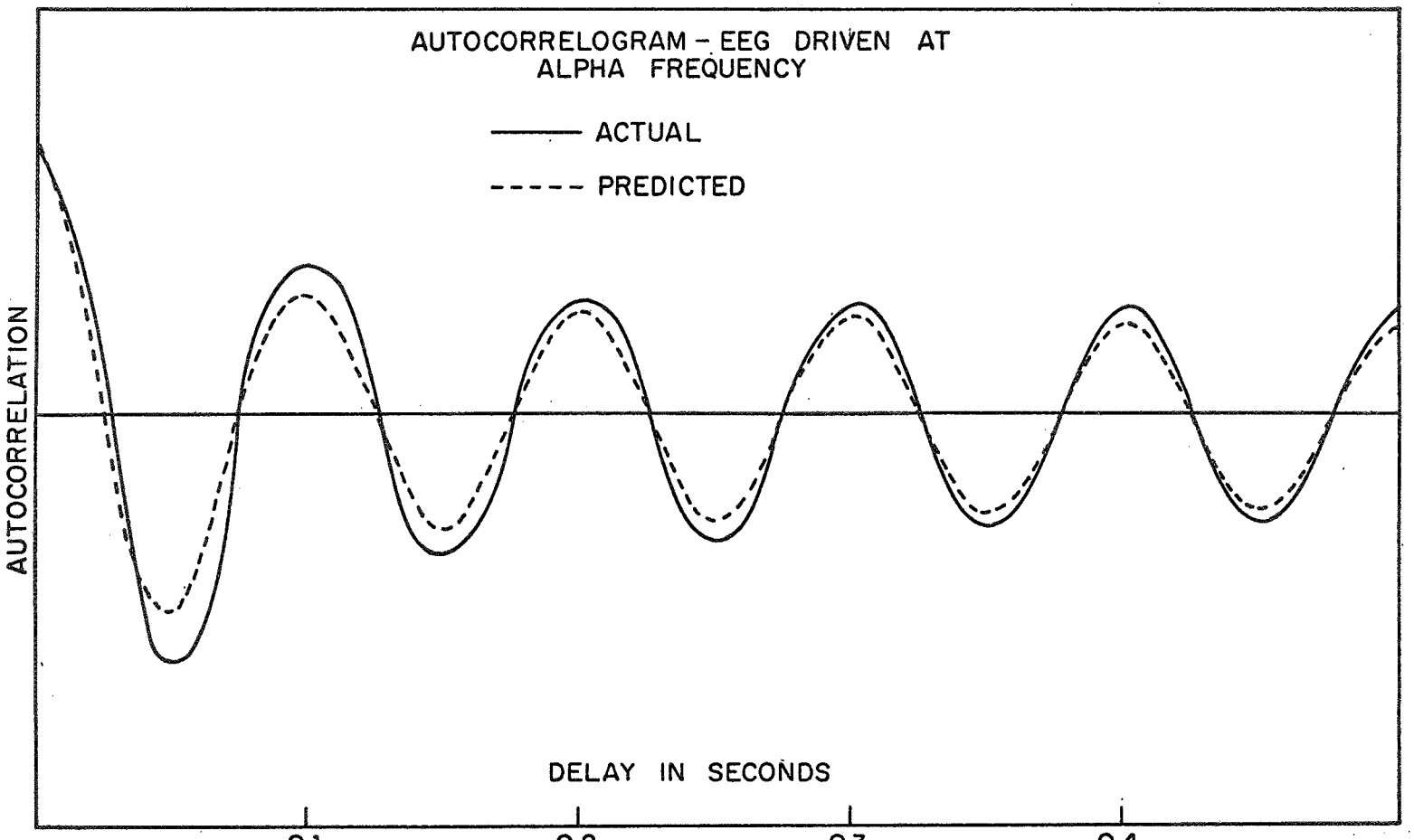

AUTOCORRELOGRAM - EEG DRIVEN AT
ALPHA FREQUENCY
ACTUAL
PREDICTED
AUTOCORRELATION
DELAY IN SECONDS
0.1
0.2
0.3
0.4

FIGURE 4-6

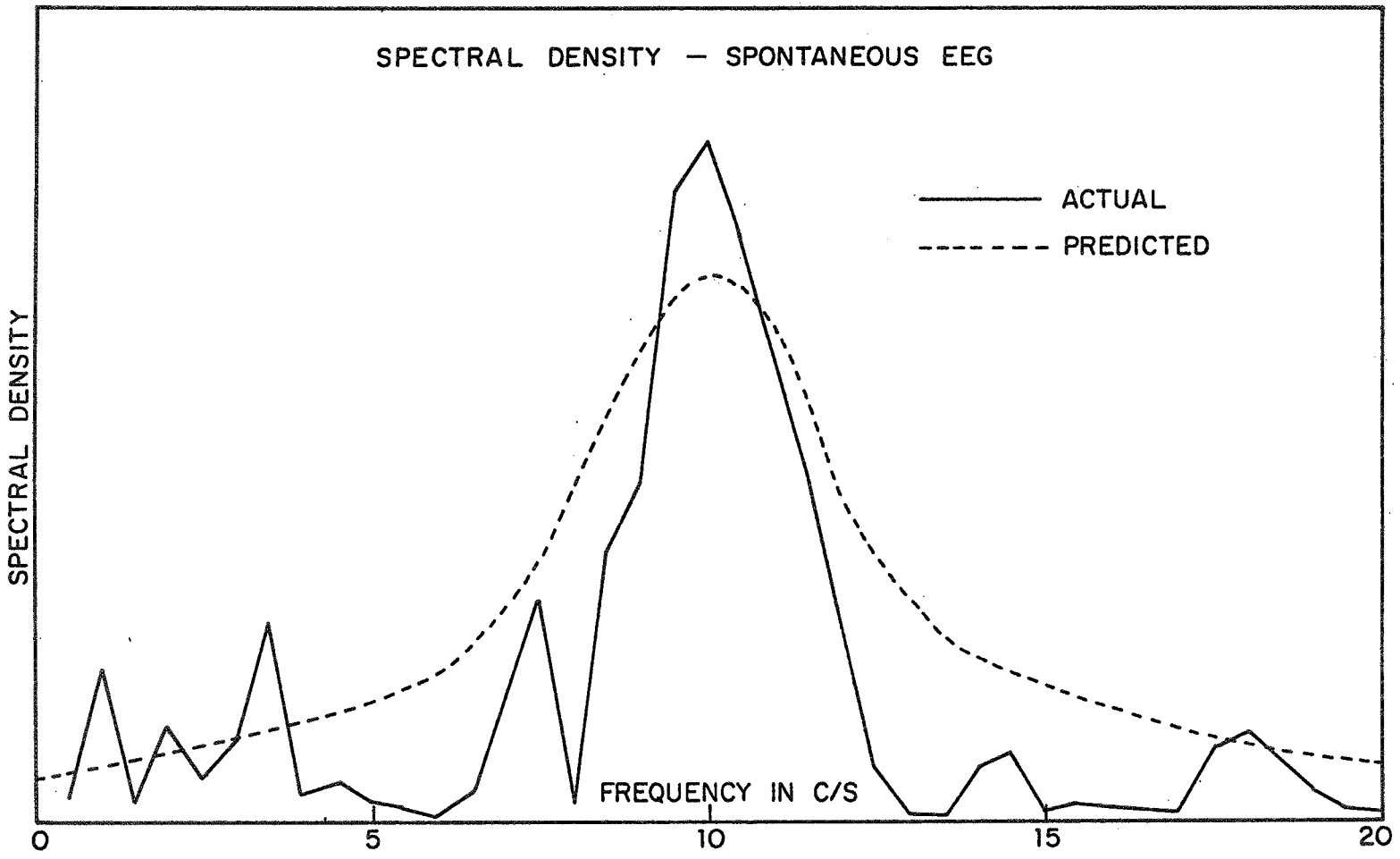

SPECTRAL DENSITY — SPONTANEOUS EEG
ACTUAL
PREDICTED
SPECTRAL DENSITY
FREQUENCY IN C/S
0
5
10
15
20

FIGURE 4-7

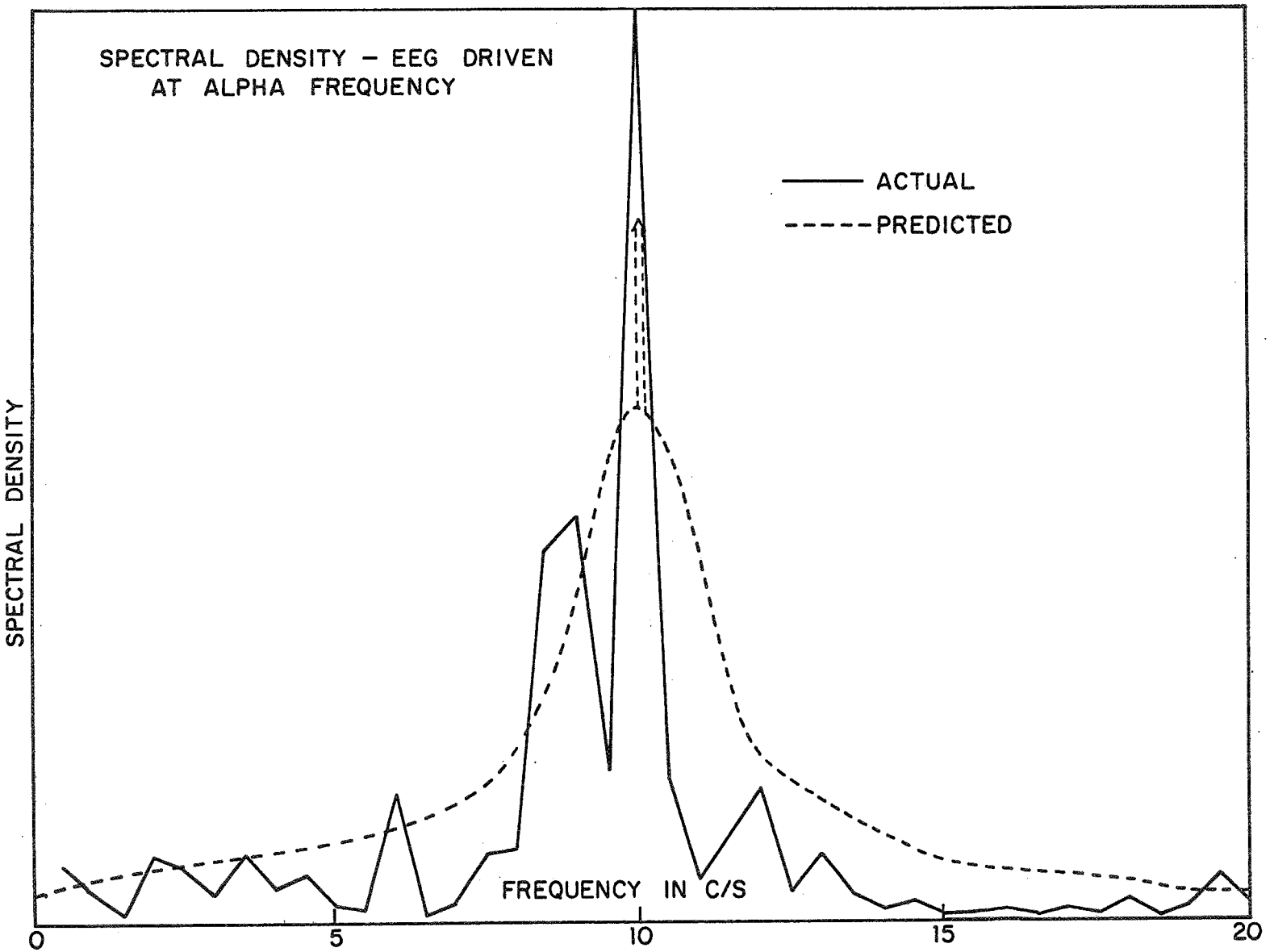

SPECTRAL DENSITY – EEG DRIVEN
AT ALPHA FREQUENCY
ACTUAL
PREDICTED
SPECTRAL DENSITY
FREQUENCY IN C/S
0
5
10
15
20

FIGURE 4-8

CHAPTER V

FEATURE REDUCTION AND CLASSIFICATION BASED ON THE MODEL

This chapter describes how an amplitude and phase analysis of the EEG record is carried out. The phase values are then used to develop a feature reduction and classification scheme based on the Bayes decision rule.

5.1 Amplitude and Phase Analysis of EEG

5.1.1 Concept of the analytic signal

Gabor [36] first introduced the concepts of the instantaneous amplitude and the instantaneous phase of an arbitrary, continuous, real signal x(t). He did this by defining a complex analytic signal z(t) associated with the real signal x(t). The signal x(t) is expressed as the real part of the analytic signal z(t). This is a generalisation of the relation for pure harmonic signals, namely, $\cos \omega t = \text{Re } e^{i\omega t}$.

The imaginary part y(t) of the analytic signal z(t) is termed the quadrature signal. (That is, the signal in quadrature to the in-phase signal x(t).) If the relation between the in-phase signal x(t) and the quadrature signal y(t) is specified, then the analytic signal z(t) can be uniquely determined from a knowledge of either x(t) or y(t).

It can be proved that, if x(t) and y(t) are defined to be a Hilbert transform pair [37], that is,

$$\left.\begin{aligned} y(t) &= \frac{1}{\pi} P \int_{-\infty}^{\infty} \frac{x(\tau)}{t-\tau} d\tau \\ x(t) &= -\frac{1}{\pi} P \int_{-\infty}^{\infty} \frac{y(t)}{t-\tau} d\tau \end{aligned}\right\} \qquad (5.1)$$

(where P denotes principal value)

then, the following desirable properties can be proved [38].

(i) If $x(t) = \cos \omega_0 t$ then $y(t) = \sin \omega_0 t$

(ii) The Fourier transforms of x(t) and y(t) are connected by

$$Y(\omega) = -iX(\omega) \operatorname{sgn} \omega.$$

In other words, x(t) and y(t) have the same power spectrum.

(iii) From (ii), $Z(\omega) = X(\omega) + iY(\omega)$

$$= (1 + \operatorname{sgn} \omega)X(\omega)$$

showing that the Fourier transform of z(t) vanishes for negative frequencies.

(iv) The one-sided nature of $Z(\omega)$ implies that, z(s) defined by inverse Fourier transformation

$$z(s) = \int_{0}^{\infty} Z(\omega)\, e^{i\omega s} dw$$

is an analytic function of the complex variable s.

(Hence the name 'analytic signal.')

Now the analytic signal can be expressed in phasor form as

$$z(t) = A(t)\, e^{i\varphi(t)}. \qquad (5.2)$$

A(t) and $\varphi(t)$ are then defined to be the amplitude and phase respectively of the real signal x(t) at the instant t.

5.1.2 Practical considerations

From equation (5.1) it is clear that generation of the quadrature signal y(t) involves a non-causal operation on the in-phase signal x(t). That is, computation of y(t) requires knowledge of future values $x(\tau)$, $\tau > t$. Theoretically, the entire EEG record is needed to compute the phase at any given instant. However, if we are willing to filter out frequencies outside a certain band from the EEG signal, the above requirement can be relaxed, as will be shown presently.

Such filtering is desirable from the practical viewpoint also. It is desirable to filter out high frequencies because any spurious high frequency 'noise' can throw off the phase value by multiples of 2π. It is also desirable to filter out very low frequencies to eliminate any bias which might have been introduced into the signal during recording. Since we are mainly interested in frequencies around the

alpha frequency (approximately 10 c/s), a bandpass of 0.2 - 50 c/s was used in practice. It is necessary that both the in-phase and quadrature signals be subjected to filtering in order to maintain the relationship between their Fourier transforms. If we work with the digitised EEG record, we have to construct two digital filters--an 'in-phase filter' and a 'quadrature filter.' Remembering the relationship between the Fourier transforms of x(t) and y(t), namely,

$$iY(\omega) = X(\omega) \operatorname{sgn} \omega,$$

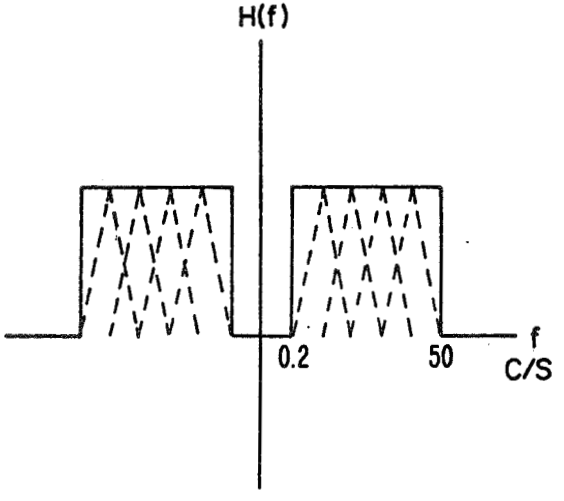
the ideal in-phase and quadrature filters would have characteristics as shown in Fig. 5.1.

In practice, the rectangular characteristics are approximated by a series of triangular filters [39, 40] as shown by broken lines in Fig. 5.1. If e(i) are the sampled values of the EEG signal, it can be shown [39] that the following operations represent the approximated in-phase and quadrature filters--

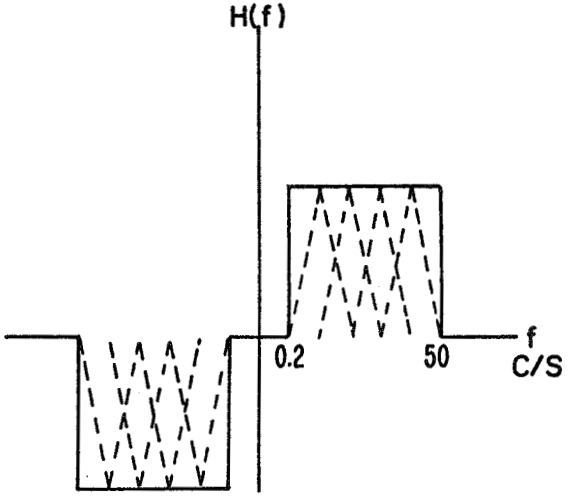
$$\left.\begin{aligned} x(k) &= \sum_{j=-m}^{m} a_j e(k+j) \\ y(k) &= \sum_{j=-m}^{m} b_j e(k+j) \end{aligned}\right\} \qquad (5.3)$$

where the numbers m, $\{a_j\}$, and $\{b_j\}$ depend on the number and shape of the triangular filters used and can be precomputed

CHARACTERISTICS OF FILTERS



IN-PHASE FILTER



QUADRATURE FILTER

FIGURE 5-1

and stored. It is thus seen that the generation of filtered in-phase and quadrature signals involves computation of two finite, moving, weighted averages. It is necessary to store (2m + 1) sampled values at a time. The filtered signals are obtained after a delay of $m\Delta t$, Δt being the sampling interval.

Once the sequences $\{x_k\}$ and $\{y_k\}$ are computed, the amplitude and phase are given by

$$A(k) = \sqrt{|x(k)|^2+|y(k)|^2}$$
$$\varphi(k) = \arctan\,(y(k),\ x(k)) \qquad (5.4)$$

It might be thought that the phase is determined only to within a multiple of 2π. However, if the sampling rate is chosen to be equal to the Nyquist rate for the highest frequency present in the filtered signal, then any two successive phase values $\varphi(k)$ and $\varphi(k + 1)$ can differ at most by π. This fact determines all phase values uniquely in relation to the initial phase value $\varphi(1)$, which is taken to be in the range $(-\pi, \pi)$.

Finally, the sequence $\theta(k)$, which is the random part in the phase, is obtained by subtraction of the linear part due to the alpha frequency ω_a.

5.2 Bayes Decision Rule Based on the Model

5.2.1 Derivation of the likelihood ratio

Let us suppose that an amplitude and phase analysis has been carried out on a length of EEG record equal to k periods of the alpha frequency. Let the sequence of amplitude and phase values be $\{A(i),\ i = 1, 2, \ldots, kN\}$ and $\{\theta(i),\ i = 1, 2, \ldots, kN\}$. Here N is the number of sampled values in one period of the alpha frequency. The k stimuli, if present, are assumed to have occurred just before the instants $1, N + 1, 2N + 1, \ldots, (k - 1)N + 1$.

Let the null hypothesis H^o be that the stimuli were not present; that is, the EEG record under examination is of the spontaneous type. A discrete version of the model for spontaneous EEG is

$$\theta_{i+1} = \theta_i + w_i \qquad i = 1, 2, \ldots, kN - 1$$

where w_i is a purely random Gaussian sequence with

$$E(w_i) = 0,\ E(w_i w_j) = q\delta_{ij}$$

and the initial uncertainty is Gaussian expressed by

$$E(\theta_1) = 0,\ E(\theta_1^2) = p_o .$$

At this point, we make the assumption that the amplitude sequence $\{A(i)\}$ and the phase sequence $\{\theta(i)\}$ are stochastically independent under either hypothesis. Thus,

$$p(A_1, A_2, \ldots, A_{kN}, \theta_1, \theta_2, \ldots, \theta_{kN} | H^i)$$

$$= p(A_1, A_2, \ldots, A_{kN} | H^i) p(\theta_1, \theta_2, \ldots, \theta_{kN} | H^i),$$

$$i = 0, 1. \tag{5.5}$$

Now from the Markovian nature of the process,

$$p(\theta_1, \theta_2, \ldots, \theta_{kN} | H^0)$$

$$= p(\theta_{kN} | \theta_{kN-1}, H^0) p(\theta_{kN-1} | \theta_{kN-2}, H^0) \ldots$$

$$p(\theta_2 | \theta_1, H^0) p(\theta^1 | H^0).$$

It is seen that

$$\log p(\theta_1, \theta_2, \ldots, \theta_{kN} | H^0)$$

$$= -\frac{1}{2q} \sum_{i=2}^{kN} (\theta_i - \theta_{i-1})^2 - \frac{\theta_1^2}{2p_0} - \left(\frac{kN-1}{2}\right) \log 2\pi q - \frac{1}{2} \log 2\pi p_0 . \tag{5.6}$$

Let the alternative hypothesis H^1 be that the stimuli were presented just before instants 1, N + 1, 2N + 1, . . . , (k - 1)N + 1. In this case we have the discrete version of

the model for EEG with repetitive stimuli as

$$\theta_{i+1} = \theta_i + w_i \qquad i = 1, 2, \ldots, N-1, N+1, N+2, \ldots, 2N-1, \ldots, (k-1)N+1, (k-1)N+2, \ldots, kN-1$$

with $E(w_i) = 0$, $E(w_i w_j) = q\delta_{ij}$,

so the θ sequence between any two successive stimuli is a random walk sequence with initial uncertainties given by

$$E(\theta_{jN+1}) = 0, \quad E(\theta^2_{jN+1}) = p_o, \qquad j = 0, 1, 2, \ldots, (k-1)$$

and the initial phase values are geometrically correlated as

$$E[\theta_{jN+1} \cdot \theta_{j'N+1}] = p_o \alpha^{|j-j'|}. \tag{5.7}$$

Now we can write

$$p(\theta_1, \theta_2, \ldots, \theta_{kN} | H^1) = p(\theta_{kN} | \theta_{kN-1}, \theta_{kN-2}, \ldots, \theta_1, H^1) \cdot p(\theta_{kN-1} | \theta_{kN-2}, \theta_{kN-3}, \ldots, \theta_1, H^1) \ldots p(\theta_3 | \theta_2, \theta_1, H^1) \cdot p(\theta_2 | \theta_1, H^1) \cdot p(\theta_1 | H^1). \tag{5.8}$$

All the conditional densities except those of θ_1, θ_{N+1}, θ_{2N+1}, . . . , $\theta_{(k-1)N+1}$ follow by the random walk model as before. The exceptional cases will have to be evaluated from the geometrical correlation model expressed by equation (5.7). These conditional densities are of the form

$$p(\theta_{jN+1} | \theta_{jN}, \theta_{jN-1}, \ldots, \theta_1, H^1), \quad j = 1, 2, \ldots, k-1.$$

As explained in Chapter IV, the model assumes that the stimuli split up the phase process into several Brownian processes, with the initial values correlated geometrically as in equation (5.7). Therefore, it is clear that the conditional density of θ_{jN+1} given, the past values depends only on other initial values $\theta_{(j-1)N+1}$, $\theta_{(j-2)N+1}$, . . . , θ_{N+1}, θ_1 and not on the intermediate values of the phase. Therefore,

$$p(\theta_{jN+1} | \theta_{jN}, \theta_{jN-1}, \ldots, \theta_1, H^1)$$

$$= p(\theta_{jN+1} | \theta_{(j-1)N+1}, \theta_{(j-2)N+1}, \ldots, \theta_{N+1}, \theta_1, H^1).$$

Under the geometrical correlation model, the conditional density on the right hand side can be proved to be normal with mean $\alpha\theta_{(j-1)N+1}$ and variance $(1 - \alpha^2)p_0$ (a proof is given in the Appendix).

Substituting everything into equation (5.8) it will be seen that

$$\log p(\theta_1, \theta_2, \ldots, \theta_{kN}|H^1) = -\frac{1}{2q}\sum_{j=1}^{k}\sum_{i=(j-1)N+2}^{jN}(\theta_i - \theta_{i-1})^2$$

$$-\frac{1}{2(1-\alpha^2)p_o}\sum_{j=1}^{k-1}(\theta_{jN+1} - \alpha\theta_{(j-1)N+1})^2 - \frac{\theta_1^2}{2p_o}$$

$$-\frac{k(N-1)}{2}\log 2\pi q - \frac{1}{2}\log 2\pi p_o$$

$$-\frac{k-1}{2}\log 2\pi(1-\alpha^2)p_o \qquad (5.9)$$

Now the likelihood ratio is

$$\frac{p(A_1, A_2, \ldots, A_{kN}|H^o)p(\theta_1, \theta_2, \ldots, \theta_{kN}|H^o)}{p(A_1, A_2, \ldots, A_{kN}|H^1)p(\theta_1, \theta_2, \ldots, \theta_{kN}|H^1)}.$$

Here we make the <u>assumption</u>

$$p(A_1, A_2, \ldots, A_{kN}|H^o) = p(A_1, A_2, \ldots, A_{kN}|H^1). \qquad (5.10)$$

That is, the stimuli do not affect the statistics of the amplitude, but only those of the phase. Then the log-likelihood ratio becomes, using (5.6) and (5.9),

$$\log p(\theta_1, \theta_2, \ldots, \theta_{kN}|H^o)$$

$$-\log p(\theta_1, \theta_2, \ldots, \theta_{kN}|H^1)$$

$$= -\frac{1}{2q}\sum_{j=1}^{k-1}(\theta_{jN+1} - \theta_{jN})^2$$

$$+\frac{1}{2(1-\alpha^2)p_o}\sum_{j=1}^{k-1}(\theta_{jN+1} - \alpha\theta_{(j-1)N+1})^2$$

$$+ \frac{k-1}{2} \log \frac{(1-\alpha^2)p_o}{q} \tag{5.11}$$

5.2.2 Sufficient statistics and feature reduction

Let

$$t_1 = \sum_{j=1}^{k-1} (\theta_{jN+1} - \theta_{jN})^2$$

$$t_2 = \sum_{j=1}^{k-1} \theta_{jN+1}^2$$

$$t_3 = \sum_{j=1}^{k-1} \theta_{(j-1)N+1}^2$$

$$t_4 = \sum_{j=1}^{k-1} \theta_{jN+1}\theta_{(j-1)N+1} .$$

Then the Bayes optimal separating surface, which is obtained by setting the log-likelihood ratio in equation (5.11) to zero, is

$$\boxed{\begin{aligned} &\left[-\frac{1}{2q}\right]\cdot t_1 + \left[\frac{1}{2(1-\alpha^2)p_o}\right]\cdot t_2 + \left[\frac{-\alpha}{(1-\alpha^2)p_o}\right]\cdot t_3 \\ &+ \left[\frac{\alpha^2}{2(1-\alpha^2)p_o}\right]\cdot t_4 + \frac{k-1}{2}\log\frac{(1-\alpha^2)p_o}{q} = 0. \end{aligned}} \tag{5.12}$$

Therefore, the decision regarding the classification of the given EEG record is based only on the four numbers (t_1, t_2, t_3, t_4) which are functions of the phase sequence $\{\theta(i), i = 1, 2, \ldots, kN\}$. From the optimality of the Bayes

decision rule, it can be claimed that, if the model holds exactly, complete knowledge of the sequence $\{\theta(i)\}$ cannot improve the average error probability. In this sense, the functions (t_1, t_2, t_3, t_4) can be termed sufficient statistics for classification--that is, they carry all the information needed for classification purposes.

The nonlinear map given by

$$\begin{pmatrix} \theta_1 \\ \theta_2 \\ \cdot \\ \cdot \\ \cdot \\ \theta_{kN} \end{pmatrix} \rightarrow \begin{pmatrix} t_1 \\ t_2 \\ t_3 \\ t_4 \end{pmatrix}$$

achieves considerable feature reduction. The following points are worth noting.

(i) If the model holds exactly, no discriminatory information is lost by the above mapping.

(ii) The dimension of the transformed pattern vector is four regardless of k, the length of the EEG record used.

(iii) The separating surface is linear in the transformed pattern space.

(iv) The sufficient statistics are functions only of the phase values immediately preceding and following the presentation of a stimulus. Since the dynamics of the phase process are the same under H^0 and H^1,

the intermediate phase values carry no discriminatory information.

(v) The sufficient statistics are in the form of cumulative sums, which can be continuously updated as more data becomes available.

(vi) From equation (5.12) it is seen that the separating hyperplane moves parallel to itself as more data comes in, but its orientation remains unchanged. (k, which affects the threshold, appears only in the constant term.) Once the orientation is determined by learning the unknown parameters (see 5.2.3) the decision can be taken after observing any desired length of the EEG record, depending on the reliability required.

5.2.3 Estimation of unknown parameters

Three unknown parameters (p_o, q, α) appear in the model. The coefficients of the separating hyperplane (5.12) are functions of these parameters. Two essentially different approaches can be taken in the estimation of these parameters.

(i) Since we are really interested in determining the separating surface, rather than the parameters themselves, we can take advantage of the linearity of the separating surface in the transformed pattern space. There are several algorithms [1] available to determine the separating hyperplane from training patterns of known classification.

Several short lengths of EEG record of known classification would be needed. Only the sufficient statistics (t_1, t_2, t_3, t_4) would be used from each of them to fit a separating hyperplane in the transformed pattern space.

(ii) The problem can be treated in the general framework of estimation (identification) of parameters in dynamical systems. For example, equation (5.6) can be used to obtain maximum likelihood estimates [41] $\hat{p}_o$ and $\hat{q}$ from the phase values of a spontaneous EEG record as follows. Solving simultaneously

$$\frac{\partial}{\partial p_o} \log p(\theta_1, \theta_2, \ldots, \theta_{kN} | H^o) = 0 \quad \text{and}$$

$$\frac{\partial}{\partial q} \log p(\theta_1, \theta_2, \ldots, \theta_{kN} | H^o) = 0$$

we obtain

$$\hat{p}_o = \theta_1^2 \quad \text{and}$$

$$\hat{q} = \frac{1}{kN-1} \sum_{i=2}^{kN} (\theta_i - \theta_{i-1})^2$$

as maximum likelihood estimates of p_o and q.

Similarly, from equation (5.9), phase values of an EEG record with repetitive stimuli can be used to obtain maximum likelihood estimates of p_o, q, α. Solving simultaneously,

$$\frac{\partial}{\partial p_o} \log p(\theta_1, \theta_2, \ldots, \theta_{kN} | H^1) = 0$$

$$\frac{\partial}{\partial q} \log p(\theta_1, \theta_2, \ldots, \theta_{kN} | H^1) = 0 \quad \text{and}$$

$$\frac{\partial}{\partial \alpha} \log p(\theta_1, \theta_2, \ldots, \theta_{kN} | H^1) = 0$$

we obtain for $\hat{q}$

$$\hat{q} = \frac{1}{k(N-1)} \sum_{j=1}^{k} \sum_{i=(j-1)N+2}^{jN} (\theta_i - \theta_{i-1})^2$$

$\hat{\alpha}$ is to be obtained from

$$\hat{\alpha} \frac{V+\hat{\alpha}^2 D-2\hat{\alpha}C}{k} = (1 - \hat{\alpha}^2)(C - \hat{\alpha}D) \quad \text{and}$$

$$\hat{p}_o = \frac{C-\hat{\alpha}D}{\hat{\alpha}}$$

where

$$V = \sum_{j=0}^{k-1} \theta^2_{jN+1}$$

$$D = \sum_{j=1}^{k-2} \theta^2_{jN+1}$$

$$C = \sum_{j=1}^{k-1} \theta_{(j-1)N+1} \theta_{jN+1}$$

5.3 Results

The method was tested on EEG data obtained from two

different subjects. Subject A had a greater percentage of alpha activity than subject B. Here the percent alpha activity is defined as the percentage of the record for which the instantaneous frequency lies within a specified narrow band around the alpha frequency. The Nyquist rate for the filtered signal is 100 per second; therefore, there would be 10 phase values in one period of the alpha frequency (N = 10).

For subject A, the estimated parameter values were

$$\hat{p}_o = 3.0 \text{ radian}^2$$

$$\hat{q} = 0.3 \text{ radian}^2$$

$$\hat{\alpha} = 0.8.$$

The separating hyperplane obtained from these values agreed reasonably well with the one obtained by a regression method [42]. Fig. 5.2 shows the error rate vs. the length of the EEG record on which the decisions were based (solid line); it is compared with the error rate obtained by the nonparametric method in Chapter III (broken line). The error rate was less than 5% when the decisions were based on 20 periods of the alpha frequency. The training and test sets were separated by about two minutes of EEG record. For subject B, the estimated parameter values were

$$\hat{p}_o = 3.4 \text{ radian}^2$$

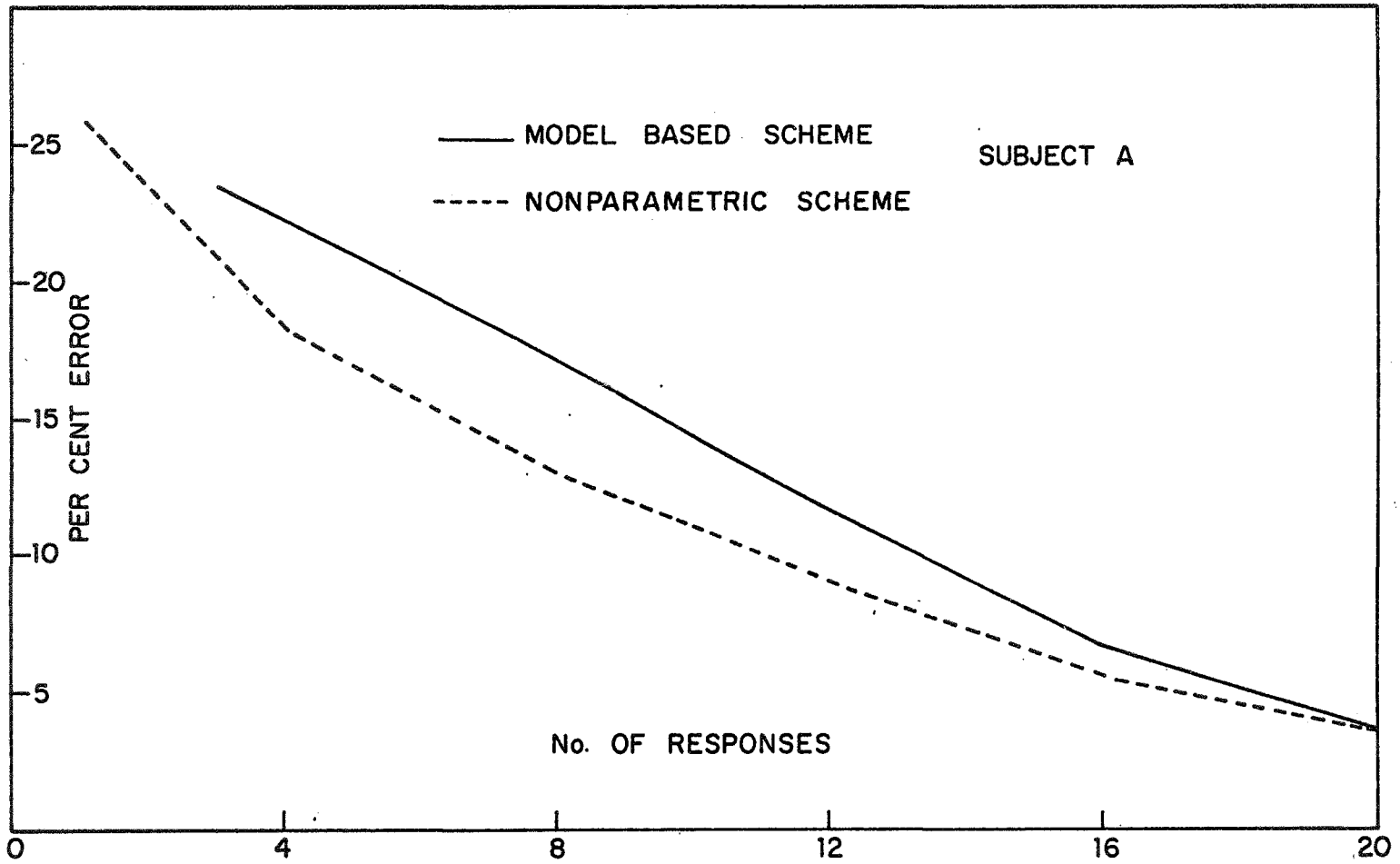

MODEL BASED SCHEME
NONPARAMETRIC SCHEME
SUBJECT A
PER CENT ERROR
25
20
15
10
5
No. OF RESPONSES
0
4
8
12
16
20

FIGURE 5-2

$\hat{q} = 0.35$ radian^2

$\hat{\alpha} = 0.65$.

The separating hyperplanes obtained by the two methods were again in reasonable agreement. Fig. 5.3 shows the error rate and compares it with that obtained by the method in Chapter III. In this case, the decisions have to be based on 40 periods of the alpha frequency before the error rate falls to 5%.

It is seen that for very short lengths of EEG record, the nonparametric method performs better. Perhaps this is because the nonparametric method makes fewer assumptions on the data--only periodic stationarity in the wide sense is required. However, as the length of the EEG record increases, even this assumption becomes harder to meet in practice and probably becomes as much 'off the mark' as the assumptions made by the model. This would explain why the error rates produced by the two methods approach each other in the 5% range.

Both the methods perform better on data obtained from subject A than on data obtained from subject B. Also, for short lengths of EEG record, the discrepancy in the performances of the two methods is larger for subject B. The model-based approach is more sensitive to large deviations from the alpha frequency than the nonparametric method. However, in the 5% range of error rate, which is of practical

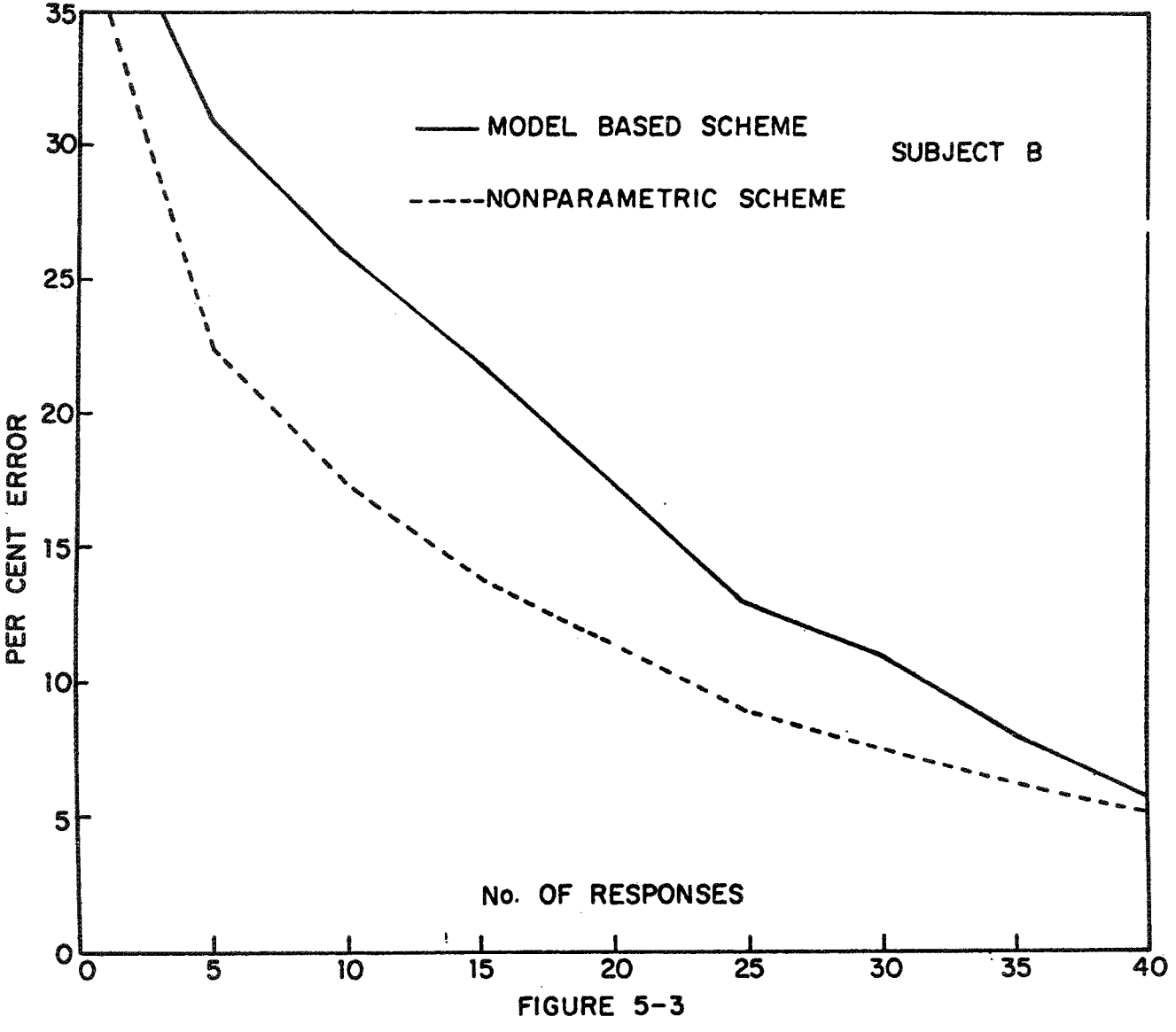

MODEL BASED SCHEME
NONPARAMETRIC SCHEME
SUBJECT B
PER CENT ERROR
No. OF RESPONSES
0
5
10
15
20
25
30
35
40

FIGURE 5-3

interest, the methods perform equally well. The computational simplicity of the model-based approach gives it a decisive advantage.

CHAPTER VI

CONCLUSIONS, POSSIBLE DIRECTIONS OF FURTHER RESEARCH

6.1 Conclusions

The feature reduction methods available so far in the literature may be considered to be of the statistical type. These methods consider the different pattern classes as statistical distributions in the N-dimensional feature space. In general, those features which have a high statistical correlation with the variable describing the category of each pattern are attempted to be singled out. Specifically, those features which minimise the probability of misclassification in a statistical sense are sought.

Let us first consider the difficulties inherent in the statistical approach. The probability of misclassification is hard to compute in all but the simplest cases. Therefore, efforts were directed towards formulating other statistical measures, like entropy and distance measures, which could conceivably be related to the probability of misclassification, and at the same time would be easy to compute. Blackwell [10], however, showed that no single measure can be explicitly related to the probability of misclassification. For this reason, the author feels that it would be wrong to expect any greatly simplified feature reduction method to be developed from the statistical approach.

Moreover, the statistical approach is, at present, restricted to considering a subset or linear combinations of the totality of all available features. This fact makes efficient feature reduction very much dependent on efficient feature selection. If the original features are carelessly chosen, there is just no way of picking a 'good' subset or a 'good' linear combination. This points to the basic flaw in statistical feature reduction--there is no link between feature selection and feature reduction.

To be sure, the approach has its advantages. The nonparametric methods, in particular, are completely general and can be put to work on almost any body of data to be classified without regard to where it came from. The author, however, doubts whether such complete generality is necessary or even desirable in <u>any given specific application</u>. Take, for example, the EEG data. There has been a great deal of research done on EEG by researchers in various fields from different points of view. When we come across a pattern classification problem involving EEG data, are we to ignore the results of this past research and take a purely statistical approach? Probably not. As another example, we can consider the problem of identifying the occurrence of different kinds of seismic phenomena from observed seismological data. Here again it would be wiser to take into account the mechanism which generates the data than to take a purely statistical approach.

This leads us to what may be termed the model-based approach. The relevant, known results on the body of data under consideration can be used to formulate a plausible mathematical model. (Random process models are especially suited to the case where the pattern features form a temporal or spatial sequence.) It is important that the mathematical model have some physical basis. Or else, the mathematical model would not be much better than fitting a statistical distribution to the data. If the model is reasonably good, it would tell us what features or attributes to look for (the phase values in the case of the EEG signal) and thus achieve feature selection. In addition, if the model is reasonably simple, it would admit a vector of sufficient statistics whose dimension is low enough so feature reduction is also accomplished.

6.2 Possible Directions of Further Research

(i) It was seen in Chapter V that the error rate obtained by the model-based approach is somewhat higher than that obtained from the non-parametric method, especially for short lengths of EEG record. It is possible that, by modelling the phase process in a different way, a better fit to the EEG data might be obtained. For instance, we can introduce damping into the Brownian process via

$$\dot{\theta} + a\theta = w \qquad (a > 0).$$

The phase dispersion can then be made to approach an asymptotic value, instead of increasing linearly. By having a larger damping in the case of EEG driven at the alpha frequency, a smaller phase dispersion can be realised. It would be interesting to find the predicted statistical behavior of the EEG under this model and its performance in classification.

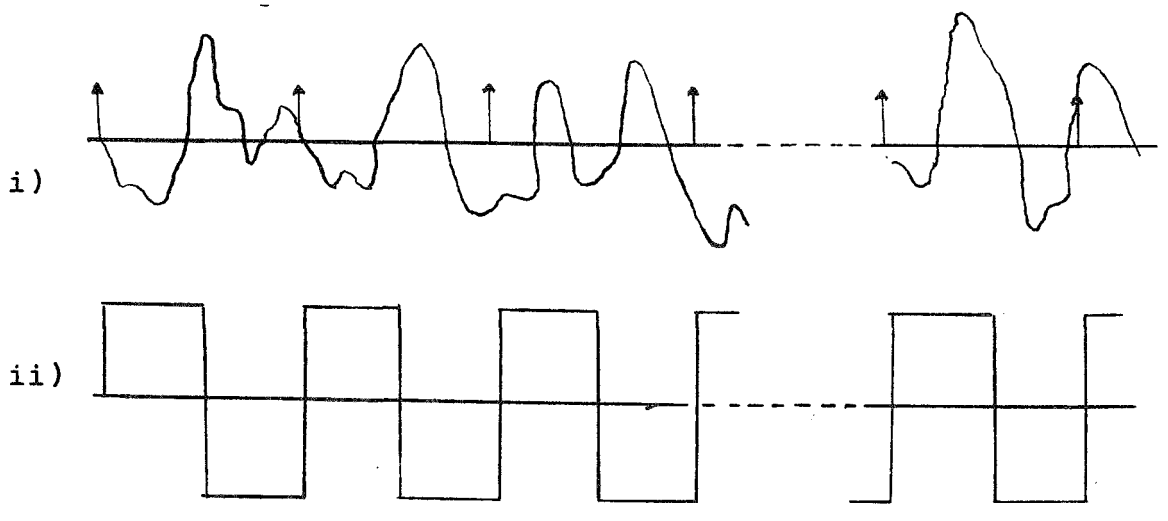
(ii) In this thesis, we have concerned ourselves with only one type of photic stimulation, namely, equally spaced stimuli at the alpha frequency. A possible line of further research is to modify, and investigate the behavior of, the model for other kinds of stimuli. Firstly, the effect of repetitive stimuli at frequencies other than the alpha frequency has to be investigated. The stimuli which are presented to the brain in a realistic situation are, however, not repetitive in nature. As the next step in generalisation, we can consider stimuli presented at random instants, which have some assumed statistical behavior. Finally, we can consider stimuli which are not instantaneous, but instead are continuously modulated.

APPENDIX

A.1 Description of Data

The EEG record of ten minutes duration from each of two subjects was obtained through NASA-ERC, Cambridge, Massachusetts. The recording was done on each person in a single sitting from a pair of electrodes located in the left occipital-parietal area. The two kinds of EEG signal are generated as follows. Stroboscopic light is flashed into the eye of the subject through closed eyelids, and it is tuned to the frequency of his alpha rhythm, which is approximately 10 c.p.s. Thus a flash occurs once every 100 milliseconds approximately. The stroboscopic light is periodically blocked so it does not reach the eye of the subject. Thus the entire EEG record is split up into groups of two kinds of responses--spontaneous and driven at the frequency of the alpha rhythm. The on and off periods each last for about 25 seconds. For general experimental conditions see Anliker [43].

The EEG data was first recorded in analog form on a 3-track FM tape as follows.



iii)

i) is the EEG record itself.

ii) is a square wave at approximately 10 c.p.s. At every leading edge of it, a stroboscopic flash occurs.

iii) is an 'on-off' waveform which indicates whether or not the light from the stroboscopic flash is reaching the eye of the subject.

To facilitate digital computer work, each of these waveforms was discretized by sampling every millisecond. It is seen that waveform (ii) serves only as a timing reference. Between two successive stroboscopic stimuli, we would have approximately 100 sampled values. In practice, it was found that the number sometimes exceeded 100 due to drifts in the stroboscopic frequency. For the work described in Chapter III, it is necessary to have pattern vectors of uniform dimension. Therefore, only the first 100 values were retained. Waveform (iii) contains only 'on-off' information. A number 1 or 0, depending on whether the subject sees the stroboscopic flash or not, was augmented to the 100 dimensional vector. This vector of 101 numbers contains all the information for the work in Chapter III.

For the amplitude and phase analysis, it is necessary to retain all sampled values in the sequence of occurrence. The timing reference information and the 'on-off' information are then stored separately.

For the benefit of anyone wishing to make use of the data, they are available on 9-track magnetic tapes for use on IBM 360/65. The format used is (4XI4, 4X15I4, 5(/4X17I4)).

A.2 Properties of $(d_{n+1} - d_n)$

In Chapter III, the following properties were claimed for the distance measure defined as

$$d = (\underline{\mu}^0 - \underline{\mu}^1)^T (\Sigma^0 + \Sigma^1)^{-1} (\underline{\mu}^0 - \underline{\mu}^1)$$

(the reader is referred to Chapter III for notation).

(i) Adding one more feature can never worsen the discriminating capacity of the features already chosen.

$$d_{n+1} - d_n \geq 0.$$

(ii) If x_{n+1} is a linear combination of $x_1, x_2, \ldots, x_n$ plus a noise term, then

$$d_{n+1} - d_n = 0.$$

(iii) If the two classes are singularly distributed in $(n + 1)$ dimensional space on separate, parallel hyperplanes, then

$$d_{n+1} - d_n = \infty.$$

(iv) $$\mu_{n+1} - C^T\Sigma^{-1}\mu = E\,[x_{n+1} \,|\, x_1 = x_2 = \cdots = x_n = 0]$$

$$\sigma_{n+1} - C^T\Sigma^{-1}C = \mathrm{Var}\,[x_{n+1} \,|\, x_1, x_2, \ldots, x_n].$$

The proofs are given below.

(i) $$\det \begin{pmatrix} \Sigma & C \\ C^T & \sigma_{n+1} \end{pmatrix} = \det \begin{pmatrix} \Sigma & C \\ C^T - C^T\Sigma^{-1}\Sigma & \sigma_{n+1} - C^T\Sigma^{-1}C \end{pmatrix}$$

by elementary row operations

$$= \det \begin{pmatrix} \Sigma & C \\ 0 & \sigma_{n+1} - C^T\Sigma^{-1}C \end{pmatrix}$$

$$= (\sigma_{n+1} - C^T\Sigma^{-1}C) \cdot \det \Sigma$$

expanding in terms of the last row.

Now,

$\det \begin{pmatrix} \Sigma & C \\ C^T & \sigma_{n+1} \end{pmatrix} \geq 0$, $\det \Sigma \geq 0$, since both are covariance matrices. Assuming Σ nonsingular,

$$\sigma_{n+1} - C^T\Sigma^{-1}C \geq 0. \qquad \text{Q.E.D.}$$

(ii) Let $x_{n+1} = \underline{\beta}^T\underline{x} + \epsilon$

where ϵ has mean $\bar{\epsilon}$, variance υ, and uncorrelated with $\underline{x}$.

Then $\mu^o_{n+1} = \underline{\beta}^T\underline{\mu}^o + \bar{\epsilon}$

$$\mu^1_{n+1} = \underline{\beta}^T\underline{\mu}^1 + \bar{\epsilon}.$$

Therefore, $\mu_{n+1} = \mu^o_{n+1} - \mu^1_{n+1}$

$$= \underline{\beta}^T(\underline{\mu}^o - \underline{\mu}^1)$$

$$= \underline{\beta}^T\underline{\mu} \qquad \text{(A.1)}$$

$$\sigma^o_{n+1} = \underline{\beta}^T\Sigma^o\underline{\beta} + \upsilon$$

$$\sigma^1_{n+1} = \underline{\beta}^T\Sigma^1\underline{\beta} + \upsilon.$$

Therefore, $\sigma_{n+1} = \sigma^o_{n+1} + \sigma^1_{n+1}$

$$= \underline{\beta}^T(\Sigma^o + \Sigma^1)\underline{\beta} + 2\upsilon$$

$$= \underline{\beta}^T\Sigma\underline{\beta} + 2\upsilon \qquad \text{(A.2)}$$

$$C^o = E_{H^o}\ (\underline{x} - \underline{\mu}^o)(x_{n+1} - \mu^o_{n+1})$$

$$= E_{H^o}\ (\underline{x} - \underline{\mu}^o)[(\underline{x} - \underline{\mu}^o)^T\underline{\beta} + \epsilon - \bar{\epsilon}]$$

$= \Sigma^o \underline{\beta}$ since $\underline{x}$ and ϵ are uncorrelated.

Similarly $C^1 = \Sigma^1 \underline{\beta}$.

Therefore, $C = C^o + C^1 = \Sigma\underline{\beta}$. (A.3)

From (A.1), (A.2), and (A.3),

$$\mu_{n+1} - C^T\Sigma^{-1}\mu = \underline{\beta}^T\underline{\mu} - \underline{\beta}^T\Sigma\Sigma^{-1}\underline{\mu} = 0$$

$$\sigma_{n+1} - C^T\Sigma^{-1}C = \underline{\beta}^T\Sigma\underline{\beta} + 2\upsilon - \underline{\beta}^T\Sigma\Sigma^{-1}\Sigma\underline{\beta} = 2\upsilon.$$

Therefore, $d_{n+1} - d_n = \frac{(\mu_{n+1} - C^T\Sigma^{-1}\mu)^2}{\sigma_{n+1} - C^T\Sigma^{-1}C} = 0$. Q.E.D.

(iii) Under hypothesis H^o let

$$x_{n+1} = \underline{\beta}^T\underline{x} + \epsilon^o$$

where ϵ^o has mean $\bar{\epsilon}^o$, variance υ and uncorrelated with $\underline{x}$.

As in (ii) it can be shown that

$$\mu^o_{n+1} = \underline{\beta}^T\underline{\mu}^o + \overline{\epsilon^o}$$

$$\sigma^o_{n+1} = \underline{\beta}^T\Sigma^o\underline{\beta} + \upsilon$$

$$C^o = \Sigma^o\underline{\beta}.$$

Under hypothesis H^1, let

$$x_{n+1} = \underline{\beta}^T \underline{x} + \epsilon^1$$

where ϵ^1 has mean $\overline{\epsilon^1}$, variance υ, and uncorrelated with $\underline{x}$. Then,

$$\mu^1_{n+1} = \underline{\beta}^T \underline{\mu}^1 + \overline{\epsilon^1}$$

$$\sigma^1_{n+1} = \underline{\beta}^T \Sigma^1 \underline{\beta} + \upsilon$$

$$C^1 = \Sigma^1 \beta .$$

Thus, $\mu_{n+1} = \mu^o_{n+1} - \mu^1_{n+1}$

$$= \underline{\beta}^T \underline{\mu} + \overline{\epsilon^o} - \overline{\epsilon^1}$$

$$\sigma_{n+1} = \sigma^o_{n+1} + \sigma^1_{n+1}$$

$$= \underline{\beta}^T \Sigma \underline{\beta} + 2\upsilon$$

$$C = C^o + C^1 = \Sigma \underline{\beta}$$

$$d_{n+1} - d_n = \frac{(\mu_{n+1} - C^T \Sigma^{-1} \mu)^2}{\sigma_{n+1} - C^T \Sigma^{-1} C}$$

$$= \frac{(\overline{\epsilon^o} - \overline{\epsilon^1})^2}{2\upsilon} .$$

Now if we let $\upsilon \to 0$ with $\overline{\epsilon^o} \neq \overline{\epsilon^1}$, we can make the two pattern classes be singularly distributed on separate, parallel hyperplanes.

$$\lim_{\upsilon \to 0} (d_{n+1} - d_n) = \infty . \quad \text{Q.E.D.}$$

(iv) It is a well-known property of multivariate normal distributions that

$$E\,[x_{n+1} \,|\, \underline{x} = \underline{\xi}] = \mu_{n+1} + C^T \Sigma^{-1} (\underline{\xi} - \underline{\mu})$$

$$\text{Var}\,[x_{n+1} \,|\, \underline{x} = \underline{\xi}] = \sigma_{n+1} - C^T \Sigma^{-1} C.$$

Setting $\underline{\xi} = \underline{0}$, we get the desired result.

A.3 Characteristic Functions

If $p(x_1, x_2, \ldots, x_n)$ is the probability density function of a random vector $(x_1, x_2, \ldots, x_n)$ then

$$M(v_1, v_2, \ldots, v_n) = E\, e^{i(v_1 x_1 + v_2 x_2 + \cdots + v_n x_n)}$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} p(x_1, x_2, \ldots, x_n)$$

$$e^{i(v_1 x_1 + v_2 x_2 + \cdots + v_n x_n)} \cdot dx_1\, dx_2 \cdots dx_n$$

is called its characteristic function. It is the multi-dimensional Fourier transform of the probability density function.

The characteristic function may be used to obtain moments of any order.

$$E\left[x_1^{r_1} x_2^{r_2} \cdots x_n^{r_n}\right] = \frac{1}{i^{r_1+r_2+\cdots+r_n}} \cdot \frac{\partial^{r_1+r_2+\cdots+r_n} M(v_1, v_2 \cdots v_n)}{\partial v_1^{r_1} \partial v_2^{r_2} \cdots \partial v_n^{r_n}}$$

$$v_1 = v_2 = \cdots = v_n = 0.$$

It can also be used to obtain the statistics of certain non-linear functions of random variables. For example, if x is a scalar r.v.,

$$E \cos x = \frac{1}{2} E\left(e^{ix} + e^{-ix}\right) = \frac{1}{2} M(1) + \frac{1}{2} M(-1).$$

From the defining equation, it is seen that $M(-1) = M^*(1)$.

Therefore, $E \cos x = \text{Re } M(1)$

or, $E \text{ Re } e^{ix} = \text{Re } E\, e^{ix}$ showing commutativity of E

and Re operators in the case of complex functions of real r.v.

As a further example, if (x_1, x_2) is a random vector,

$$E \sin x_1 \sin x_2 = \frac{1}{2} E\left[\cos(x_1 - x_2) - \cos(x_1 + x_2)\right]$$

$$= \frac{1}{4} E\left[e^{i(x_1-x_2)} + e^{-i(x_1-x_2)} - e^{i(x_1+x_2)} - e^{-i(x_1+x_2)}\right]$$

$$= \frac{1}{4} [M(1, -1) + M(-1, 1) - M(1, 1) - M(-1, -1)]$$

$$= \frac{1}{2} \text{Re} [M(1, -1) - M(1, 1)],$$

since $M(-1, 1) = M^*(1, -1)$ and $M(-1, -1) = M^*(1, 1)$

A.4 Variability of Estimates

In Chapter IV, the expected values of the average signal and the autocorrelogram as predicted by the model were shown to agree closely with the predicted behavior. Here the variability of these estimates is derived. In each case, it is shown that the variability goes to zero, as the length of the EEG record used goes to infinity.

If z is a complex r.v.,

$$\text{Var } z = E\,|z - E(z)|^2$$

$$= E\,|z|^2 - |E(z)|^2.$$

The average signal (spontaneous EEG)

$$z = \frac{1}{N} \sum_{j=0}^{N-1} x(jt_a + t).$$

It was proved in Chapter IV that

$$E(z) = \frac{1}{N} e^{-\frac{1}{2}qt_a} \left(\frac{1-e^{-\frac{1}{2}Nqt_a}}{1-e^{-\frac{1}{2}qt_a}} \right) e^{-\frac{1}{2}(p_o+qt)} e^{i\omega_a t}.$$

Now,

$$|z|^2 = zz^*$$

$$= \frac{1}{N^2} \sum_{j=0}^{N-1} \sum_{k=0}^{N-1} x(jt_a + t)x^*(kt_a + t)$$

$$= \frac{1}{N^2} \sum_{j=0}^{N-1} \sum_{k=0}^{N-1} e^{i\omega_a (jt_a+t)+i\theta(jt_a+t)} \; e^{-i\omega_a (kt_a+t)-i\theta(kt_a+t)}$$

$$= \frac{1}{N^2} \sum_{j=0}^{N-1} \sum_{k=0}^{N-1} e^{i\theta(jt_a+t)-i\theta(kt_a+t)} \qquad \text{since } \omega_a t_a = 2\pi.$$

In the sum, there are N terms for which j = k and hence these terms are all equal to unity. By symmetry, the rest of the sum is equal to twice the triangular sum for which j > k. Thus,

$$|z|^2 = \frac{1}{N^2} \left(N + 2 \sum_{j=0}^{N-1} \sum_{k=0}^{j-1} e^{i\theta(jt_a+t)-i\theta(kt_a+t)} \right). \qquad \text{(A.4)}$$

$\theta(jt_a + t)$ and $\theta(kt_a + t)$ are jointly Gaussian with

$$E\,[\theta(jt_a + t)]^2 = p_o + q(jt_a + t)$$

$$E\,[\theta(kt_a + t)]^2 = p_o + q(kt_a + t)$$

$$E\,\theta(jt_a + t)\theta(kt_a + t) = p_o + q(kt_a + t) \quad \text{for } k < j.$$

Therefore,

$$E\ e^{i\theta(jt_a+t)-i\theta(kt_a+t)} = M(1, -1)$$

$$= e^{-\frac{1}{2}(j-k)qt_a}. \qquad (A.5)$$

Now,

$$\sum_{j=0}^{N-1}\sum_{k=0}^{j-1} e^{-\frac{1}{2}(j-k)qt_a} = \sum_{j=0}^{N-1} e^{-\frac{1}{2}jqt_a} \sum_{k=0}^{j-1} e^{\frac{1}{2}kqt_a}$$

$$= \sum_{j=0}^{N-1} e^{-\frac{1}{2}jqt_a} \cdot \left(\frac{1-e^{\frac{1}{2}jqt_a}}{1-e^{\frac{1}{2}qt_a}}\right)$$

$$= \frac{1}{1-e^{\frac{1}{2}qt_a}} \left[\frac{1-e^{-\frac{1}{2}Nqt_a}}{1-e^{-\frac{1}{2}qt_a}} - N\right]. \qquad (A.6)$$

From (A.4), (A.5) and (A.6)

$$E\ |z|^2 = \frac{1}{N}\frac{e^{\frac{1}{2}qt_a}+1}{e^{\frac{1}{2}qt_a}-1} + \frac{1}{N^2}\frac{2(1-e^{-\frac{1}{2}Nqt_a})}{(1-e^{\frac{1}{2}qt_a})(1-e^{-\frac{1}{2}qt_a})}$$

$$\text{Var } z = E\ |z|^2 - |E(z)|^2$$

$$= \frac{1}{N}\frac{e^{\frac{1}{2}qt_a}+1}{e^{\frac{1}{2}qt_a}-1} + \frac{1}{N^2}\frac{2(1-e^{-\frac{1}{2}Nqt_a})}{(1-e^{\frac{1}{2}qt_a})(1-e^{-\frac{1}{2}qt_a})}$$

$$- \frac{1}{N^2} e^{-qt_a} \left(\frac{1-e^{-\frac{1}{2}Nqt_a}}{1-e^{-\frac{1}{2}qt_a}}\right)^2 e^{-(p_o+qt)}.$$

It is seen that var $z \to 0$ as $N \to \infty$, implying mean square convergence.

<u>The average signal (EEG with repetitive stimuli)</u>

In this case,

$$E\,[\theta(jt_a + t)]^2 = p_o + qt$$

$$E\,[\theta(kt_a + t)]^2 = p_o + qt$$

$$E\;\theta(jt_a + t)\cdot\theta(kt_a + t) = p_o\,\alpha^{|j-k|}.$$

Therefore,

$$E\;e^{i\theta(jt_a+t)-i\theta(kt_a+t)} = e^{-(p_o+qt)}\;e^{+p_o\alpha^{|j-k|}}$$

and

$$E|z|^2 = \frac{1}{N^2}\sum_{j=0}^{N-1}\sum_{k=0}^{N-1} e^{p_o\alpha^{|j-k|}}\cdot e^{-(p_o+qt)}. \qquad (A.7)$$

From Chapter IV,

$$|E(z)|^2 = e^{-(p_o+qt)}.$$

Therefore,

$$\text{Var } z = e^{-(p_o+qt)}(S_N - 1) \quad \text{where } S_N \text{ is the double sum in (A.7).}$$

It will now be proved that $\lim_{N\to\infty} S_N = 1$, so we have mean square convergence.

$$S_N = \frac{1}{N^2}\sum_{j=0}^{N-1}\sum_{k=0}^{N-1} e^{p_o\alpha^{|j-k|}} \qquad 0 < \alpha < 1.$$

Given any infinitesimal $\epsilon_1 > 0$, we can always find $N_1(\epsilon_1)$ such that

$$1 \leq e^{p_o \alpha^{|j-k|}} \leq 1 + \epsilon_1 \quad \text{for } |j - k| \geq N_1(\epsilon_1).$$

(The left hand inequality is obviously true. The right hand inequality implies

$$p_o \alpha^{|j-k|} \leq \ln (1 + \epsilon_1)$$

$$\rightarrow \ln p_o + (j - k) \ln \alpha \leq \ln \ln (1 + \epsilon_1)$$

$$\rightarrow |j - k| \geq \frac{\ln \ln (1+\epsilon_1) - \ln p_o}{\ln \alpha}$$

(observing that $\ln \alpha$ is negative).

The right side is a positive number and we have to merely pick $N_1(\epsilon_1)$ as the nearest integer above it.)

Now split the double sum into two parts--T_{1N} for which $|j - k| \geq N_1(\epsilon_1)$ and T_{2N} for which $|j - k| < N_1(\epsilon_1)$ so that

$$S_N = \frac{1}{N^2} (T_{1N} + T_{2N}).$$

Any term in T_{2N} is bounded above by e^{p_o}, the bound being achieved for $j - k = 0$.

Therefore,

$$1 \leq \text{any term in } T_{1N} \leq 1 + \epsilon_1 \leq \text{any term in } T_{2N} \leq e^{Po}.$$

The number of terms in T_{2N} is of $0(N)$. Therefore, given any infinitesimal $\epsilon_2 > 0$, we can find $N_2(\epsilon_2)$ such that

$$0 \leq \frac{1}{N^2} T_{2N} \leq \epsilon_2 \quad \text{for } N \geq N_2(\epsilon_2). \tag{A.8}$$

The number of terms in T_{1N} is $0(N^2)$. Therefore,

$$1 \leq \frac{1}{N^2} T_{1N} \leq 1 + \epsilon_1 \quad \text{for } N \geq N_1(\epsilon_1). \tag{A.9}$$

Adding inequalities (A.8) and (A.9),

$$1 \leq S_N \leq 1 + \epsilon_1 + \epsilon_2 \quad \text{for } N \geq \text{Max } [N_1(\epsilon_1), N_2(\epsilon_2)].$$

Letting $\epsilon_1, \epsilon_2 \to 0$ and $N \to \infty$, we get

$$\lim_{N \to \infty} S_N = 1 \qquad \text{Q.E.D.}$$

Autocorrelogram (spontaneous EEG)

$$R_N(\tau) = \frac{1}{Nt_a} \int_0^{Nt_a} x(t)x^*(t + \tau)dt$$

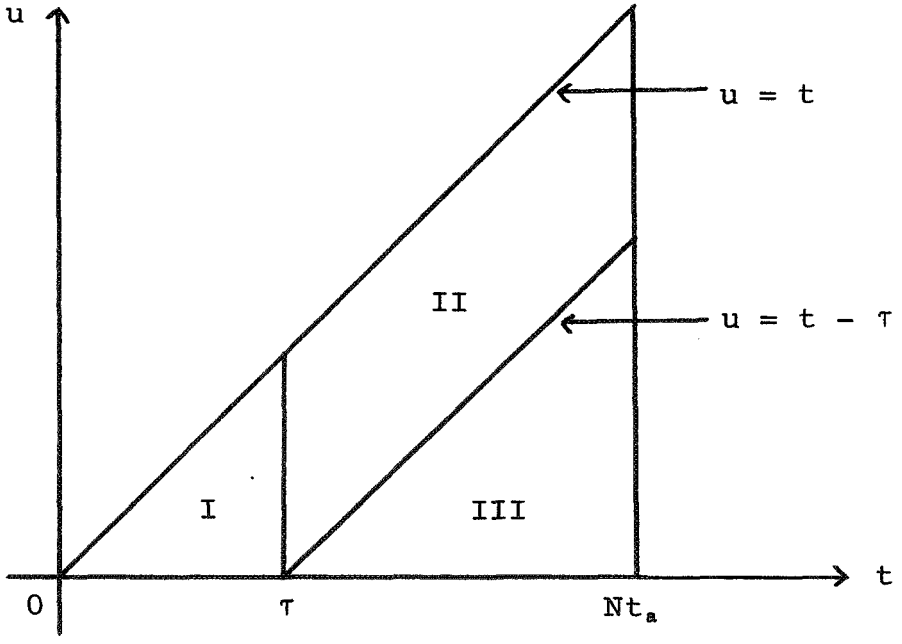
$$|R_N(\tau)|^2 = \frac{1}{N^2 t_a^2} \int_0^{Nt_a} x(t)x^*(t + \tau)dt \cdot \int_0^{Nt_a} x^*(u)x(u + \tau)du$$

$$= \frac{1}{N^2 t_a^2} \int_0^{Nt_a} \int_0^{Nt_a} e^{i[\omega_a t + \theta(t)]} e^{-i[\omega_a u + \theta(u)]}.$$

$$e^{i[\omega_a(u+\tau)+\theta(u+\tau)]} \cdot e^{-i[\omega_a(t+\tau)+\theta(t+\tau)]} dudt$$

$$= \frac{1}{N^2 t_a^2} \int_0^{Nt_a} \int_0^{Nt_a} e^{i\theta(t)-i\theta(t+\tau)-i\theta(u)+i\theta(u+\tau)} dudt.$$

In order to evaluate $E\,|R_N(\tau)|^2$, we need the fourth order characteristic function of random variables $\theta(t)$, $\theta(t+\tau)$, $\theta(u)$, $\theta(u+\tau)$ which are jointly Gaussian. It will have to be evaluated at (1, -1, -1, 1) and depends on the relative positions of the instants t, $t+\tau$, u, $u+\tau$. First, it is observed that, by symmetry, the integral is twice that over the triangular region shown in the figure below. The triangle itself is split up into regions I, II, III, the integrals over which are denoted by I_1, I_2, I_3.



Thus $E|R_N(\tau)|^2 = \frac{2}{N^2 t_a^2}(I_1 + I_2 + I_3)$. (A.9)

In regions I and II,

$$u < t < u + \tau < t + \tau.$$

Therefore, the covariance matrix of the four random variables $\theta(t)$, $\theta(t + \tau)$, $\theta(u)$, $\theta(u + \tau)$ is given by

$$\begin{bmatrix} p_o + qt & p_o + qt & p_o + qu & p_o + qt \\ p_o + qt & p_o + qt + q\tau & p_o + qu & p_o + qu + q\tau \\ p_o + qu & p_o + qu & p_o + qu & p_o + qu \\ p_o + qt & p_o + qu & p_o + qu + q\tau & p_o + qu \end{bmatrix}$$

Thus it can be seen that

$$M(1, -1, -1, 1) = e^{-q(t-u)} \quad \text{after cancellation of terms.}$$

Therefore,

$$I_1 = \int_o^\tau \int_o^\tau e^{-q(t-u)} du dt$$

$$= \frac{\tau}{q} - \frac{1}{q^2}(1 - e^{-q\tau}) \quad \text{after evaluation.}$$

$$I_2 = \int_\tau^{Nt_a} \int_{t-\tau}^{t} e^{-q(t-u)} du dt$$

$$= \frac{1}{q}(Nt_a - \tau)(1 - e^{-q\tau}) \quad \text{after evaluation.}$$

In region III,

$$u < u + \tau < t < t + \tau.$$

The covariance matrix of the four random variables is then given by

$$\begin{bmatrix} p_o + qt & p_o + qt & p_o + qu & p_o + qu + q\tau \\ p_o + qt & p_o + qt + q\tau & p_o + qu & p_o + qu + q\tau \\ p_o + qu & p_o + qu + q\tau & p_o + qu & p_o + qu \\ p_o + qu + q\tau & p_o + qu + q\tau & p_o + qu & p_o + qu + q\tau \end{bmatrix}$$

Thus,

$$M(1, -1, -1, 1) = e^{-q\tau}$$

$$I_3 = \int_{\tau}^{Nt_a} \int_{0}^{t-\tau}$$

$$= \frac{1}{2} (Nt_a - \tau)^2 e^{-q\tau}.$$

Substituting into (A.9),

$$E\ |R_N(\tau)|^2 = \frac{1}{N^2 t_a^2} \left[\frac{2\tau}{q} - \frac{2}{q^2} (1 - e^{-q\tau}) + \frac{2}{q} (Nt_a - \tau)(1 - e^{-q\tau}) \right.$$

$$\left. + (Nt_a - \tau)^2 e^{-q\tau} \right].$$

As $N \to \infty$, <u>for fixed τ</u>,

$$\lim_{N \to \infty} E\ |R_N(\tau)|^2 = e^{-q\tau}.$$

From Chapter IV $|E\ R_N(\tau)|^2 = e^{-q\tau}$.

Therefore, $\lim_{N\to\infty}$ Var $R_N(\tau) = 0$, again proving mean square convergence.

A.5 Evaluation of a Conditional Density

$(x_1, x_2, \ldots, x_N)$ are real r.v.s with zero mean and covariance matrix $p_o \Sigma_N$ where

$$\Sigma_N = \begin{pmatrix} 1 & \alpha & \alpha^2 & \alpha^3 & \ldots & \alpha^{N-1} \\ \alpha & 1 & \alpha & \alpha^2 & \ldots & \alpha^{N-2} \\ \alpha^2 & \alpha & 1 & \alpha & \ldots & \alpha^{N-3} \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ \alpha^{N-1} & \alpha^{N-2} & \ldots & \ldots & \alpha & 1 \end{pmatrix}$$

It is required to find the conditional density

$$p(x_N | x_1, x_2, \ldots, x_{N-1}).$$

It is first necessary to find det Σ_N and Σ_N^{-1}. Expanding in terms of the first row,

$$\det \Sigma_N = \det \begin{pmatrix} 1 & \alpha & \alpha^2 & \cdots & \alpha^{N-2} \\ \alpha & 1 & \alpha & \cdots & \alpha^{N-3} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \alpha^{N-2} & \alpha^{N-3} & \cdots & \cdots & 1 \end{pmatrix}$$

$$- \alpha \cdot \det \begin{pmatrix} \alpha & \alpha & \cdots & \alpha^{N-2} \\ \alpha^2 & 1 & \cdots & \alpha^{N-3} \\ \alpha^3 & \alpha & & \alpha^{N-4} \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ \alpha^{N-1} & \alpha^{N-3} & \cdots & 1 \end{pmatrix}$$

since all other cofactors are zero

$$= [\det \Sigma_{N-1} - \alpha^2 \det \Sigma_{N-1}]$$

$$= (1 - \alpha^2) \det \Sigma_{N-1} .$$

By induction it follows that

$$\det \Sigma_N = (1 - \alpha^2)^{N-1} \det \Sigma_1 = (1 - \alpha^2)^{N-1} .$$

Σ_N^{-1} is the tridiagonal matrix

$$\frac{1}{1-\alpha^2}\begin{pmatrix} 1 & -\alpha & 0 & 0 & \cdots & 0 \\ -\alpha & 1+\alpha^2 & -\alpha & 0 & \cdots & 0 \\ 0 & -\alpha & 1+\alpha^2 & -\alpha & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & -\alpha & 1+\alpha^2 & -\alpha \\ 0 & 0 & \cdots & 0 & -\alpha & 1 \end{pmatrix}$$

which can be verified by direct multiplication.

$$\text{Now } p(x_N \mid x_1, x_2, \ldots, x_{N-1}) = \frac{p(x_1, x_2, \ldots, x_N)}{p(x_1, x_2, \ldots, x_{N-1})}$$

$$p(x_1, x_2, \ldots, x_N) = \frac{1}{(2\pi)^{\frac{N}{2}} (p_o^N \det \Sigma_N)^{\frac{1}{2}}} e^{-\frac{Q_N}{2p_o(1-\alpha^2)}}$$

$$p(x_1, x_2, \ldots, x_{N-1}) = \frac{1}{(2\pi)^{\frac{N-1}{2}} (p_o^{N-1} \det \Sigma_{N-1})^{\frac{1}{2}}} e^{-\frac{Q_{N-1}}{2p_o(1-\alpha^2)}}$$

where $Q_N = (x_1, x_2, \ldots, x_N)(\Sigma_N)^{-1} \begin{pmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ x_N \end{pmatrix}$

$$= (x_1^2 + x_2^2 + \cdots + x_N^2) + \alpha^2 (x_2^2 + x_3^2 + \cdots + x_{N-1}^2)$$

$$- 2\alpha(x_1 x_2 + x_2 x_3 + \cdots + x_{N-1} x_N).$$

It is seen that

$$Q_N - Q_{N-1} = x_N^2 + \alpha^2 x_{N-1}^2 - 2\alpha x_{N-1} x_N$$

$$= (x_N - \alpha x_{N-1})^2 .$$

Therefore,

$$p(x_N | x_1, x_2, \ldots, x_{N-1}) = \frac{1}{\sqrt{2\pi p_o (1-\alpha^2)}} \exp - \frac{(x_N - \alpha x_{N-1})^2}{2p_o (1-\alpha^2)}$$

which is Gaussian with mean αx_{N-1} and variance $p_o(1 - \alpha^2)$.

BIBLIOGRAPHY

1. Ho, Y.C. and Agrawala, A. K. "On Pattern Classification Algorithms--Introduction and Survey." Proc. IEEE, LVI, No. 12 (December, 1968), 2101-14.
2. Fu, K.S. and Chien, Y.T. "A Dynamic Programming Approach to Sequential Pattern Recognition." IEEE Trans. EC (December, 1967), 790-803.
3. Abramson, N.M. Information Theory and Coding. New York: McGraw-Hill, 1963.
4. Kolmogorov, A.N. and Fomin, S.V. Elements of the Theory of Functions and Functional Analysis. Rochester: Graylock Press, 1957.
5. Grettenberg, T.L. "Signal Selection in Communication and Radar Systems." IEEE Trans. IT (October, 1963), 265-75.
6. Kailath, T. "The Divergence and Bhattacharyya Distance Measures in Signal Selection." IEEE Trans. COM (February, 1967), 52-60.
7. Karlin, S. and Bradt, R.N. "On the Design and Comparison of Dichotomous Experiments." Ann. Math. Stat., XXVII (1956), 389-409.
8. Kovalevsky, V.A. (ed.). Character Readers and Pattern Recognition. Spartan Books, Macmillan & Co. Ltd., 1968.
9. Bhattacharyya, A. "On a Measure of Divergence between Two Statistical Populations Defined by their Probability

Distributions." Bull. Calcutta Math. Soc., XXXV (1943), 99-109.

10. Blackwell, D. "Comparison of Experiments." Proceedings of the Second Berkeley Symposium on Probability and Statistics, Vol. I. Berkeley: University of California Press, 1951, pp. 93-102.

11. Watanabe, S. "Karhunen-Loéve Expansion and Factor Analysis, Theoretical Remarks and Applications." Information Theory, Statistical Decision Functions, Random Processes; Transactions of the 4th Prague Conference (1965), 635-60.

12. Watanabe, S. et al. "Evaluation and Selection of Variables in Pattern Recognition." Computer and Information Sciences--II. Washington, D.C.: Academic Press, 1967, pp. 91-103.

13. Chien, Y.T. and Fu, K.S. "On the Generalised Karhunen-Loéve Expansion." IEEE Trans. IT (July, 1967), 518-20.

14. Tou, J.T. and Heydorn, R.P. "Some Approaches to Optimum Feature Extraction." Computer and Information Sciences--II. Washington, D.C.: Academic Press, 1967, pp. 57-89.

15. Patrick, E.A. and Fischer, F.P. "Nonparametric Feature Selection." IEEE Trans. IT (September, 1969), 577-83.

16. Rémond, A. et al. "The Alpha Average I--Methodology and Description." Electroencephalography and Clinical Neurophysiology, XXVI (1969), 245-65.

17. Estrin, T. and Uzgalis, R. "Computerised Display of Spatio-Temporal EEG patterns." IEEE Trans. BME (July, 1969), 192-6.

18. Blackman, R.B. and Tukey, J.W. The Measurement of Power Spectra. New York: Dover, 1958.

19. Rosenblith, W. A. (ed.). Processing Neuroelectric Data. Cambridge: The M.I.T. Press, 1962.

20. Matoušek, M. "Frequency Analysis in Routine Electroencephalography." Electroencephalography and Clinical Neurophysiology, XXIV (1968), 365-73.

21. Johnson, L. et al. "Spectral Analysis of the EEG of Dominant and Non-dominant Alpha Subjects During Waking and Sleeping." Electroencephalography and Clinical Neurophysiology, XXVI (1969), 361-70.

22. Regan, D. "A High Frequency Mechanism which Underlies Visual Evoked Potentials." Electroencephalography and Clinical Neurophysiology, XXV (1968), 231-7.

23. Kitajima, H. "On the Cerebral Evoked Response in Man as a Function of the Intensity of Flicker Stimulation." Electroencephalography and Clinical Neurophysiology, XXII(1967), 325-36.

24. Kawabata, N. "Scalp Responses to Photic Stimulation by Time Sequence Patterns." Electroencephalography and Clinical Neurophysiology, XXV (1968), 449-54.

25. Bracewell, R. The Fourier Transform and Its Applications. New York: McGraw-Hill, 1965.

26. Mahalanobis, P.C. "On the Generalised Distance in Statistics." Proc. National Inst. Sci. (India), XII (1936), 49-55.

27. Anderson, T.W. An Introduction to Multivariate Statistical Analysis. New York: Wiley, 1958.

28. Marill, T. and Green, D.M. "On the Effectiveness of Receptors in Recognition Systems." IEEE Trans. IT (January, 1963), 11-17.

29. Bodewig, E. Matrix Calculus. Amsterdam: North Holland Publishing Co., 1956.

30. Peterson, D.W. and Mattson, R.L. "A Method of Finding Linear Discriminant Functions for a Class of Performance Criteria." IEEE Trans. IT (July, 1966).

31. Bryson, A.E. and Ho, Y.C. Applied Optimal Control. Waltham: Blaisdell, 1969.

32. Wiener, N. Nonlinear Problems in Random Theory. Cambridge: The M.I.T. Press, 1966.

33. Joseph, J.P. et al. "The Alpha Average II: Quantitative Study and the Proposition of a Theoretical Model." Electroencephalography and Clinical Neurophysiology, XXVI (1969), 350-60.

34. Wax, N. (ed.). Selected Papers on Noise and Stochastic Processes. New York: Dover, 1954.

35. Papoulis, A. Probability, Random Variables, and Stochastic Processes. New York: McGraw-Hill, 1965.

36. Gabor, O. "Theory of Communication." J-IEE, XLIII,

Part 3 (November, 1946), 429-57.

37. Titchmarsh, E.C. Theory of Fourier Integrals. Oxford: The Clarendon Press, 1937.

38. Voelcker, H.B. "Toward a Unified Theory of Modulation Part I: Phase-envelope Relations." Proc. IEEE, LIV, No. 3 (March, 1966), 340-53.

39. Goodman, N.R. "Measuring Amplitude and Phase." Journal of the Franklin Institute (December, 1960), 437-50.

40. Dixon, W.J. (ed.). BMD--Biomedical Computer Programs. Berkeley and Los Angeles: University of California Press, 1968.

41. Deutsch, R. Estimation Theory. Englewood Cliffs: Prentice-Hall, Inc., 1965.

42. Ho, Y.C. and Kashyap, R.L. "A Class of Iterative Procedures for Linear Inequalities." J. SIAM Control, IV, No. 1 (1966), 112-15.

43. Anliker, J. E. "Simultaneous Changes in Visual Separation Threshhold and Voltage of Cortical Alpha Rhythm." Science, Vol. 153, No. 3733 (July 15, 1966), 316-318.

ACKNOWLEDGEMENTS

The author is indebted to Professor Y. C. Ho for guidance and encouragement throughout this work. He was always willing to engage in a helpful discussion and many a time channelled the author's thoughts in the right direction.

The author would like to express his appreciation to Dr. J. E. Anliker of the Biotechnology Division, NASA-ERC. He provided the EEG data which stimulated this work. Moreover, he gave the author an insight into the behavior of the phase in the two types of EEG signal and pointed out that the phase would be an important attribute for discrimination. This led the author to formulate the random process model described in the report.

The research was largely supported by NASA through Contract No. NGL 22-007-143.

The author had discussions with many of his colleagues, notably Mr. A. K. Agrawala and Mr. J. L. Poage.

Joint Services Distribution List

Asst Director/Research (Rm 3C128)
Office of the Secretary of Defense
Pentagon
Washington, D. C. 20301

Technical Library
DDR and E
Room 3C-122, The Pentagon
Washington, D. C. 20301

Chief, R and D Division (340)
Defense Communications Agency
Washington, D. C. 20305

Director for Materials Sciences
ARPA
Room 3D179, The Pentagon
Washington, D.C. 20301

Major Richard J. Gowen
Tenure Associate Professor
Dept of Electrical Engineering
USAF Academy, Colorado 80840

Defense Documentation Center
Attn: DDC-TCA
Cameron Station
Alexandria, Virginia 22314 (20)

M. A. Rothenberg (STEPD-SD(S))
Scientific Director
Desert Test Center
Bldg 100, Soldiers' Circle
Fort Douglas, Utah 84113

Mr. H. E. Webb, Jr (EMBIS)
Rome Air Development Center
Griffiss Air Force Base, New York 13440

Central Intelligence Agency
Attn: CRS/ADD/PUBLICATIONS
Washington, D.C. 20505

Hq. USAF (AFRDDD)
The Pentagon
Washington, D.C. 20330

Hq. USAF (AFRDDG)
The Pentagon
Washington, D.C. 20330

Hq. USAF (AFRDSD)
The Pentagon
Washington, D.C. 20330
Attn: LTC C. M. Waespy

Colonel E. P. Gaines, Jr.
ACDA/FO
1901 Pennsylvania Avenue N.W.
Washington, D.C. 20451

Lt. Col. H. W. Jackson (SREE)
Chief, Electronics Division
Directorate of Engineering Sciences
Air Force Office of Scientific Research
Arlington, Virginia 22209 (5)

Dr. I. R. Mirman
Hq AFSC (SCGP)
Andrews Air Force Base
Washington, D. C. 20331

Commanding General
USACDC Institute of Land Combat
Attn: Technical Library, Rm 636
2461 Eisenhower Avenue
Alexandria, Virginia 22314

Rome Air Development Center
Attn: Documents Library (EMTLD)
Griffiss Air Force Base
New York 13440

MIT Lincoln Laboratory
Attn: Library A-082
P. O. Box 73
Lexington, Mass. 02173

Dr. L. M. Hollingsworth
AFCRL (CRN)
L. G. Hanscom Field
Bedford, Massachusetts 01730

VELA Seismological Center
300 North Washington Street
Alexandria, Virginia 22314

Hq. ESD (ESTI)
L. G. Hanscom Field
Bedford, Massachusetts 01730 (2)

Prof. R. H. Rediker
Electrical Engineering, Professor
MIT
Building 13-3050
Cambridge, Massachusetts 02139

AFAL (AVT) Dr. H. V. Noble
Electronics Technology Division
Air Force Avionics Laboratory
Wright-Patterson AFB, Ohio 45433

Director
Air Force Avionics Laboratory
Wright-Patterson AFB
Ohio 45433

AFAL (AVTA/R. D. Larson)
Wright-Patterson AFB
Ohio 45433

Director of Faculty Research
Dept. of the Air Force
U. S. Air Force Academy
Colorado Springs, Colorado 80840

Academy Library (DFSLB)
USAF Academy
Colorado Springs, Colorado 80840

Director
Aerospace Mechanics Sciences
Frank J. Seiler Research Lab. (OAR)
USAF Academy
Colorado Springs, Colorado 80840

Director, USAF PROJECT RAND
Via: Air Force Liaison Office
The RAND Corporation
Attn: Library D
1700 Main Street
Santa Monica, California 90406

HQ SAMSO (SMTAE/Lt. Belate)
AF Unit Post Office
Los Angeles, California 90045

Miss R. Joyce Harman
Project MAC, Room 810
545 Main Street
Cambridge, Mass. 02139

AUL3T-9663
Maxwell AFB, Alabama 36112

AFETR Technical Library
(ETV, MU-135)
Patrick AFB, Florida 32925

ADTC (ADBPS-12)
Eglin AFB, Florida 32542

Mr. B. R. Locke
Technical Adviser, Requirements
USAF Security Service
Kelly Air Force Base, Texas 78241

Hq. AMD (AMR)
Brooks AFB, Texas 78235

USAFSAM (SMKOR)
Brooks AFB, Texas 78235

Commanding General
Attn: STEWS-RE-L, Technical Library
White Sands Missile Range
New Mexico 88002 (2)

Hq. AEDC (AETS)
Arnold AFS, Tennessee 37389

USAF
European Office of Aerospace Research
APO, New York 09667

Director
Physical and Engineering Sciences Div.
3045 Columbia Pike
Arlington, Virginia 22204

Commanding General
U. S. Army Security Agency -
Attn: IARD-T
Arlington Hall Station
Arlington, Virginia 22212

Commanding General
U. S. Army Material Command
Attn: AMCRD-TP
Washington, D.C. 20315

Commanding Officer
Harry Diamond Laboratories
Attn: Dr. Berthold Altman (AMXDO-TI)
Connecticut Avenue & Van Ness St. N. W.
Washington, D. C. 20438

Chief
Missile Electronic Warfare Tech Area
(AMSEL-WL-M)
U. S. Army Electronics Command
White Sands Missile Range
New Mexico 88002

Commanding Officer (AMXRD-BAT)
U. S. Army Ballistics Research Lab.
Aberdeen Proving Ground
Aberdeen, Maryland 21005

Technical Director
U. S. Army Limited War Laboratory
Aberdeen Proving Ground
Aberdeen, Maryland 21005

Commanding Officer
U. S. Army Engineer Topographic Labs
Attn: STINFO Center
Fort Belvoir, Virginia 22060

U. S. Army Munitions Command
Attn: Science & Technology Info. Branch
Building 59
Picatinny Arsenal, SMUPA-RT-S
Dover, New Jersey 07801

U. S. Army Mobility Equipment Research
and Development Center
Attn: Technical Document Center
Building 315
Fort Belvoir, Virginia 22060

Commanding Officer (AMSEL-BL-WS-R)
Atmospheric Sciences Laboratory
U. S. Army Electronics Command
White Sands Missile Range
New Mexico 88002

Dr. Herman Robl
Deputy Chief Scientist
U. S. Army Research Office (Durham)
Box CM, Duke Station
Durham, North Carolina 27706

Richard O. Ulsh (CRDARD-IP)
U. S. Army Research Office (Durham)
Box CM, Duke Station
Durham, North Carolina 27706

Technical Director
(SMUFA-A2000-107-1)
Frankford Arsenal
Philadelphia, Pa. 19137

Redstone Scientific Info. Center
Attn: Chief, Document Section
U. S. Army Missile Command
Redstone Arsenal, Alabama 35809

Commanding General
U. S. Army Missile Command
Attn: AMSMI-RR
Redstone Arsenal, Alabama 35809

Commanding General
U. S. Army Strategic Comm. Command
Attn: SCC-CG-SAE
Fort Huachuca, Arizona 85613

Commanding Officer
Army Materials and Mechanics
Research Center
Attn: Dr. H. Priest
Watertown Arsenal
Watertown, Mass. 02172

Commandant
U. S. Army Air Defense School
Attn: Missile Science Div. C&S Dept.
P. O. Box 9390
Fort Bliss, Texas 79916

Commandant
U. S. Army Command and General
Staff College
Attn: Acquisitions, Lib. Div.
Fort Leavenworth, Kansas 66027

Mr. Norman J. Field, AMSEL-RD-S
Chief, Office of Science & Technology
Research & Development Directorate
U. S. Army Electronics Command
Fort Monmouth, New Jersey 07703

Mr. Robert O. Parker, AMSEL-RD-S
Executive Secretary, JTAC/JSEP
U. S. Army Electronics Command
Fort Monmouth, New Jersey 07703

Commanding General
U. S. Army Electronics Command
Fort Monmouth, New Jersey 07703
Attn: AMSEL-SC
DL
GG-DD
XL-D
XL-DT
BL-FM-P
CT-D
CT-R
CT-S
CT-L (Dr. W. S. McAfee)
CT-O
CT-I
CT-A
NL-D (Dr. H. Bennett)
NL-A
NL-C
NL-P
NL-P-2
NL-R
NL-S
KL-D
KL-I
KL-E
KL-S
KL-SM
KL-T
VL-D
VL-P
WL-D

Dr. Alvin D. Schnitzler
Institute for Defense Analyses
Science and Technology Division
400 Army-Navy Drive
Arlington, Virginia 22202

Director (NV-D)
Night Vision Laboratory, USAECOM
Fort Belvoir, Virginia 22060

Commanding Officer
Atmospheric Sciences Laboratory
U. S. Army Electronics Command
White Sands Missile Range
New Mexico 88002

Code 8050
Maury Center Library
Naval Research Laboratory
Washington, D. C. 20390

Dr. A. G. Jordan
Head of Dept. of Electrical Engineering
Carnegie-Mellon University
Pittsburgh, Pennsylvania 15213

Project Manager
NAVCON
Attn: Harold H. Bahr (AMCPM-NS-TM)
Building 439
U. S. Army Electronics Command
Fort Monmouth, New Jersey 07703

Director, Electronic Programs
Attn: Code 427
Dept. of the Navy
Washington, D.C. 20360 (3)

Commander
Naval Security Group Command
Naval Security Group Headquarters
Attn: G43
3801 Nebraska Avenue
Washington, D.C. 20390

Director
Naval Research Laboratory
Washington, D.C. 20390
Attn: Code 2027 (6)
Dr. W. C. Hall, Code 7000 (1)
Dr. A. Brodzinsky (1)
Supt. Elec. Div.

Dr. G. M. R. Winkler
Director, Time Service Division
U. S. Naval Observatory
Washington, D.C. 20390

Naval Air Systems Command
AIR 03
Washington, D.C. 20360 (2)

Naval Ship Systems Command
SHIP 031
Washington, D.C. 20360

Naval Ship Systems Command
SHIP 035
Washington, D.C. 20360

U. S. Naval Weapons Laboratory
Dahlgren, Virginia 22448

Naval Electronic Systems Command
ELEX 03, Rm 5534 Main Navy Bldg.
Dept. of the Navy
Washington, D.C. 20360 (2)

Government Documents Dept.
University of Iowa Libraries
Iowa City, Iowa 52240

Commander
U. S. Naval Ordnance Laboratory
Attn: Librarian
White Oak, Maryland 21502 (2)

Director
Naval Research Laboratory
Attn: Library, Code 2029 (ONRL)
Washington, D.C. 20390 (6)

Hollander Associates
P. O. Box 2276
Fullerton, California 92633

Illinois Institute of Technology
Dept. of Electrical Engineering
Chicago, Illinois 60616

The University of Arizona
Dept. of Electrical Engineering
Tucson, Arizona 85721

Utah State University
Dept. of Electrical Engineering
Logan, Utah 84321

Case Institute of Technology
Engineering Division
University Circle
Cleveland, Ohio 44106

Carl E. Baum, Capt.
AFWL (WLRE)
Kirtland AFB, New Mexico 87117

Lenkurt Electric Co., Inc.
1105 County Road
San Carlos, California 94070
Attn: Mr. E. K. Peterson

Dr. F. R. Charvat
Union Carbide Corporation
Materials Systems Division
Crystal Products Dept.
8888 Balboa Avenue
P. O. Box 23017
San Diego, California 92123

Director
U. S. Army Advanced Materiel
Concepts Agency
Washington, D. C. 20315

Electromagnetic Compatibility
Analysis Center
(ECAC), Attn: ACOAT
North Severn
Annapolis, Maryland 21402

Dept. of Electrical Engineering
Rice University
Houston, Texas 77001

Research Laboratories for the
Engineering Sciences
School of Engineering and Applied
Science
University of Virginia
Charlottesville, Virginia 22903

Dept. of Electrical Engineering
Clippinger Laboratory
Ohio University
Athens, Ohio 45701

Lehigh University
Dept. of Electrical Engineering
Bethelehem, Pennsylvania 18015

Professor James A. Cadzow
Dept. of Electrical Engineering
State Univ. of New York at Buffalo
Buffalo, New York 14214

Director
Office of Naval Research Branch Office
495 Summer Street
Boston, Mass. 02210

Commander (ADL)
Naval Air Development Center
Johnsville, Warminster, Pa. 18974
Attn: NADC Library

Commander (Code 753)
Naval Weapons Center
Attn: Technical Library
China Lake, California 93555

Commanding Officer
Naval Weapons Center
Corona Annex
Attn: Library
Corona, California 91720

Commanding Officer (56322)
U. S. Naval Missile Center
Point Mugu, California 93041

W. A. Eberspacher, Assoc. Head
Systems Integration Division
Code 5340A
U. S. Naval Missile Center
Point Mugu, California 93041

Commander
Naval Electronics Laboratory Center
Attn: Library
San Diego, California 92152 (2)

Deputy Director and Chief Scientist
Office of Naval Research Branch Office
1030 East Green Street
Pasadena, California 91101

Library (Code 2124)
Technical Report Section
Naval Postgraduate School
Monterey, California 93940

Glen A. Myers (Code 52 Mv)
Assoc. Prof. of Electrical Engineering
Naval Postgraduate School
Monterey, California 93940

Commanding Officer (Code 2064)
Navy Underwater Sound Laboratory
Fort Trumbull
New London, Conn. 06320

Commanding Officer
Naval Avionics Facility
Indianapolis, Indiana 46241

Dr. H. Harrison, Code RRE
Chief, Electrophysics Branch
National Aeronautics and Space Admn.
Washington, D.C. 20546

NASA Lewis Research Center
Attn: Library
21000 Brookpark Road
Cleveland, Ohio 44135

Los Alamos Scientific Laboratory
Attn: Reports Library
P. O. Box 1663
Los Alamos, New Mexico 87544

Mr. M. Zane Thornton, Chief
Network Engineering, Communications
and Operations Branch
Lister Hill National Center for
Biomedical Communications
8600 Rockville Pike
Bethesda, Maryland 20014

U. S. Post Office Dept.
Library - Room 6012
12th & Pennsylvania Ave, N.W.
Washington, D.C. 20260

Director
Research Lab of Electronics
MIT
Cambridge, Mass. 02139

Mr. Jerome Fox
Research Coordinator
Polytechnic Institute of Brooklyn
333 Jay Street
Brooklyn, New York-11201

Director
Columbia Radiation Laboratory
Columbia University
538 West 120th Street
New York, New York 10027

Director
Coordinated Science Laboratory
University of Illinois
Urbana, Illinois 61801

Director
Stanford Electronics Laboratories
Stanford University
Stanford, California 94305

Director
Microwave Physics Laboratory
Stanford University
Stanford, California 94305

Director
Electronics Research Laboratory
University of California
Berkeley, California 94720

Director
Electronic Sciences Laboratory
University of Southern California
Los Angeles, California 90007

Director
Electronics Research Center
The University of Texas at Austin
Engineering-Science Bldg 110
Austin, Texas 78712

Division of Engineering and Applied Physics
130 Pierce Hall
Harvard University
Cambridge, Mass. 02138

Dr. G. J. Murphy
The Technological Institute
Northwestern University
Evanston, Illinois 60201

Dr. John C. Hancock, Head
School of Electrical Engineering
Purdue University
Lafayette, Indiana 47907

Dept. of Electrical Engineering
Texas Technological College
Lubbock, Texas 79409

Aerospace Corporation
P. O. Box 95085
Los Angeles, California 90045
Attn: Library Acquisitions Group

Professor Nicholas George
California Institute of Technology
Pasadena, California 91109

Aeronautics Library
Graduate Aeronautical Laboratories
California Institute of Technology
1201 E. California Blvd
Pasadena, California 91109

The Johns Hopkins University
Applied Physics Laboratory
Attn: Document Librarian
8621 Georgia Avenue
Silver Spring, Maryland 20910

Hunt Library
Carnegie-Mellon University
Schenley Park
Pittsburgh, Pa. 15213

Dr. Leo Young
Stanford Research Institute
Menlo Park, California 94025

Chairman, Electrical Engineering
Arizona State University
Tempe, Arizona 85281

Engineering & Mathematical
Sciences Library
University of Calif. at L. A.
405 Hilgred Avenue
Los Angeles, California 90024

Sciences-Engineering Library
University of California
Santa Barbara, California 93106

Prof. Joseph E. Rowe
Chairman, Dept. of Electrical Engin
The University of Michigan
Ann Arbor, Michigan 48104

Dr. W. R. LePage, Chairman
Syracuse University
Dept. of Electrical Engineering
Syracuse, New York 13210

Yale University
Dept. of Engineering and Applied Science
New Haven, Conn. 06520

Airborne Instruments Laboratory
Deerpark, New York 11729

Raytheon Company
Research Division Library
28 Seyon Street
Waltham, Massachusetts 02154

Dr. Sheldon J. Welles
Electronic Properties Information Center
Mail Station E-175
Hughes Aircraft Company
Culver City, California 90230

Dr. Robert E. Fontana
Dept. of Electrical Engineering
Air Force Institute of Technology
Wright-Patterson AFB, Ohio 45433

Dr. John R. Ragazzini, Dean
School of Engineering and Science
New York University
University Heights
Bronx, New York 10453

Sylvania Electronic Systems
Applied Research Laboratory
Attn: Documents Librarian
40 Sylvan Road
Waltham, Mass. 02154

Unclassified

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)*
Division of Engineering and Applied Physics
Harvard University
Cambridge, Massachusetts

2a. REPORT SECURITY CLASSIFICATION
Unclassified

2b. GROUP

3. REPORT TITLE

ON FEATURE REDUCTION WITH APPLICATION TO ELECTROENCEPHALOGRAMS

4. DESCRIPTIVE NOTES *(Type of report and inclusive dates)*
Technical Report No. 615

5. AUTHOR(S) *(First name, middle initial, last name)*
Karkal P. S. Prabhu

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
September 1970	157	-

8a. CONTRACT OR GRANT NO.
N00014-67-A-0298-0006

b. PROJECT NO.

c.

d.

9a. ORIGINATOR'S REPORT NUMBER(S)
Technical Report No. 615

9b. OTHER REPORT NO(S) *(Any other numbers that may be assigned this report)*
NASA NGL 22-007-143

10. DISTRIBUTION STATEMENT

This document has been approved for public release and sale; its distribution is unlimited. Reproduction in whole or in part is permitted by the U. S. Government.

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY
	NASA

13. ABSTRACT

This report deals with the feature reduction problem in pattern classification, with application to electroencephalograph (EEG) signals. The problem under consideration is that of discriminating between two kinds of signals--spontaneous EEG and EEG driven by photic stimuli at the alpha frequency. Since an EEG record represents a large amount of data, efficient feature reduction methods are required to pick out a few features which are significant for discrimination purposes.

The first two chapters are of an introductory nature describing statistical feature reduction methods given in the literature and some relevant facts about EEG signals. The third chapter develops a nonparametric feature reduction procedure based on a distance measure. The features used are sampled values of the EEG. A feature of the method is that the computations involved in feature reduction also yield the best separating hyperplane at each stage.

The fourth chapter develops a random process model for the two kinds of EEG signals. The signal is essentially represented as a sinusoid at the alpha frequency with random amplitude and phase modulation. It is seen that the statistical properties predicted by the model agree closely with the observed results. In the fifth chapter, the model is employed for feature reduction and pattern classification. The model provides a four dimensional vector of sufficient statistics, which contains all the information necessary for discrimination purposes. The sufficient statistics are functions of the phase values of the EEG. They are in the form of cumulative sums which can be updated as more data becomes available. Moreover, the Bayes optimal separating surface is linear in terms of these sufficient statistics.

The error rates obtained by the two methods are compared. It is seen that in the 5% range of error rate, which is of practical interest, the two methods perform equally well. The computational simplicity of the model-based method gives it a decisive advantage.

14. KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Feature reduction Pattern classification Electroencephalograph						