AMRL-TR-66-24

RESEARCH ON ADVANCED COMPUTER METHODS FOR BIOLOGICAL DATA PROCESSING

D. N. STREETER, PhD
J. RAVIV, PhD

THOMAS J. WATSON RESEARCH CENTER
INTERNATIONAL BUSINESS MACHINES CORPORATION

APRIL 1966

AEROSPACE MEDICAL RESEARCH LABORATORIES
AEROSPACE MEDICAL DIVISION
AIR FORCE SYSTEMS COMMAND
WRIGHT-PATTERSON AIR FORCE BASE, OHIO

NOTICES

When US Government drawings, specifications, or other data are used for any purpose other than a definitely related Government procurement operation, the Government thereby incurs no responsibility nor any obligation whatsoever, and the fact that the Government may have formulated, furnished, or in any way supplied the said drawings, specifications, or other data, is not to be regarded by implication or otherwise, as in any manner licensing the holder or any other person or corporation, or conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

Requests for copies of this report should be directed to either of the addressees listed below, as applicable:

Federal Government agencies and their contractors registered with Defense Documentation Center (DDC):

DDC
Cameron Station
Alexandria, Virginia 22314

Non-DDC users (stock quantities are available for sale from):

Chief, Input Section
Clearinghouse for Federal Scientific & Technical Information (CFSTI)
Sills Building
5285 Port Royal Road
Springfield, Virginia 22151

Organizations and individuals receiving reports via the Aerospace Medical Research Laboratories' automatic mailing lists should submit the addressograph plate stamp on the report envelope or refer to the code number when corresponding about change of address or cancellation.

Do not return this copy. Retain or destroy.

800 - July 1966 - 773-50-1140

FOREWORD

This research was conducted by International Business Machines Corporation, Thomas J. Watson Research Center, P.O. Box 218, Yorktown Heights, New York, under Contract AF 33(615)-2047, for Aerospace Medical Research Laboratories, Wright-Patterson Air Force Base, Ohio. The work was initiated by the Mathematics and Analysis Branch, Biodynamics and Bionics Division of the Biophysics Laboratory in support of Project 7233, "Biological Information Handling Systems and their Functional Analogs," Task 723305, "Theory of Information Handling." The research was initiated 15 July 1964 and completed 14 July 1965 . The technical contract monitor was Hans L. Oestreicher, Ph.D, Chief, Mathematics and Analysis Branch.

This report contains the second phase of a program which started with the investigation of the usefulness of the Loeve-Karhunen method for biological data processing. The first phase was reported in AMRL-TR-65-114, The Loeve-Karhunen Expansion as a Means of Information Compression for Classification of Continuous Signals.

The following people are acknowledged for their participation in the research on this contract: Mr. Morris C. Grove, Programmer; Mr. Joseph E. Harry, Programmer; Dr. Walter L. Makous, Physiologist.

This technical report has been reviewed and is approved.

J. W. HEIM, PhD
Technical Director
Biophysics Laboratory
Aerospace Medical Research Laboratories

ABSTRACT

The purpose of the research carried out under this contract has been the development of mathematical methods and computer programs for the extraction of meaningful information from biological, primarily neurophysiological, measurements. Emphasis has been placed on statistical methods suitable for separating two or more random signals and which provide insight into the underlying mechanism by which the signals are generated. Loeve-Karhunen expansion and Discriminant Analysis methods are applied to the problem of time signal classification. Experiments are performed both on computer generated time signals and on electroencephalograms. Methods of coping with the singularity problem arising from a small sample size are investigated.

TABLE OF CONTENTS

SECTION I

INTRODUCTION

The purpose of the research reported in this and a previous volume[1] has been the development of new mathematical methods and computer programs for the extraction of meaningful information from biological, primarily neurophysiological, measurements. Emphasis has been placed on statistical methods suitable for separating two or more random signals and which provide insight into the underlying mechanism by which the signals are generated.

The previous efforts were largely concentrated on the Loeve-Karhunen decomposition of continuous signals. The work covered by this report is directed toward an evaluation, extension and modification of this method.

In attempting to extract meaningful information from biological measurements, several fundamental operations are often required. They are

Compression: i.e., the reduction of the amount of information which must be stored, consistent with the need for subsequent processing or reconstitution

Clustering: i.e., the separation of a set of measurements into subsets, the number of which may or may not be predetermined, on the basis of some measure of similarity

Classification: i.e., the operation of assigning a measurement to one of a number of existing classes.

In this research, "biological measurements" are assumed to be time signals, i.e., continuous scalar functions of time, having a known bandwidth. The continuous signals can be replaced, without loss of generality, by discrete time series produced by appropriate sampling techniques.

To provide insight into the choices of methodology and understanding of assumptions that must be satisfied before significant solutions can be expected, some of the differences between biological signals and those classes of signals more frequently encountered in classification problems should be considered.

Most classification problems which have been solved (e.g., communication problems, optical character recognition problems) deal with a finite number of deterministic signals which are contaminated primarily by additive noise. In contrast, many biological signals (e.g., the spontaneous electroencephalogram) have no deterministic component and, therefore, can only be described in statistical terms. If an attempt is made to define the statistical parameters by averaging a number of signals and those signals have been significantly contaminated by nonadditive noise (such as multiplicative noise, random time base compression or random phase shift), the averaging may eliminate the signal. If systematic distortions have been introduced in the measurement process, the signal-to-noise ratio may actually be worsened by averaging.

Another problem arising in processing biological signals is whether the desired information is at all present in the signal being processed. Often it will be necessary to assume the information resides therein. A negative result can therefore mean either that the information is not present or that the retrieval technique is deficient.

The sample size may be relatively small in biological processing problems, because of the difficulty in making measurements or finding sufficient numbers of specimens of a certain class. This limitation may affect the choice of decision procedures.

Most of the methods described in this report are based on linear operations, which are based on the minimization of some mean-square error criterion. These limitations have been imposed for the sake of mathematical and computational simplicity. Any explicit or implicit claim of optimality should be interpreted in the context of these limitations.

The potential advantages of using computers to compress, cluster, or classify biological data may seem too obvious to warrant discussion. However, for completeness the following points are made:

1. Compression

It appears entirely feasible to replace simple or multiple time signals, or pictures, by sets of numbers such that a) any processing can be based on these numbers in lieu of the original signals, or b) the original signals can be reconstituted at will to any predetermined degree of accuracy. The primary advantages are the convenience and economy of transmitting and storing a few numbers instead of time signals or pictures.

2. Clustering

Clustering routines are useful for locating and defining subsets within a measurement set on the basis of commonly held attributes which may not be apparent from visual inspection. The information derived about the nature, location, and quantity of these "clusters" may provide a basis for the formulation of hypotheses regarding the underlying mechanism by which the signals have been generated. In some cases it will be possible to test these hypotheses by carrying out classification experiments on additional data.

3. Classification

The automatic classification of biological data can be considered at several levels of difficulty. At the simpler level are applications in which discrimination is required only between good records and those containing gross errors or artifacts, or between records with or without some large, well-defined occurrence. Some of these rudimentary but important applications are well within the capabilities of the methods described herein. The classifying performance on the basis of progressively more subtle attributes will have to be evaluated on the particular data classes of interest. In the more complex problems, the methods presented here may be useful components of a more comprehensive, as yet undeveloped, methodology.

SECTION II

EVALUATION OF LOEVE-KARHUNEN METHODS AS APPLIED TO THE ANALYTICAL CLASSIFICATION OF BIOLOGICAL DATA

1. Integral Equation Formulation

Loeve-Karhunen methods are based on certain desirable properties, (i.e., orthogonality, completeness, efficiency) of the eigenfunctions of linear integral equations having symmetric kernels.[2] Karhunen[3] and Loeve[4] applied this theory to the decomposition and discrete representation of random time functions, in which case the kernel is the autocovariance function of the random process. The methods have found practical application in the design of communication and radar detection systems.[5,6]

These methods completely specify the design of a set of linear filters that are "optimal" (in a sense that will be described below) for analyzing the random process in question. These filters have weighting functions matched to the eigenfunctions corresponding to the largest eigenvalues. The number of filters required in any application can be readily determined as a simple function of the magnitudes of the eigenvalues. If it is necessary to analyze a random process, the statistics of which are stationary but unknown, the following procedure is indicated:

1. Record a statistically representative sample of the process
2. Compute the eignevalues and eignefunctions of a sample covariance matrix of the process
3. Adjust the filter bank,* in accordance with the above.

The filter bank, then will more efficiently analyze that random process than any band-pass filter bank, or any other linear filter having the same number of paths.

If the statistics of the process change slowly and smoothly with time, periodic recomputation and adjustment provide the basis for an efficient time varying analyzer.

---

*The term "filter bank" is used for simplicity, although it is anticipated that the filter will usually be simulated.

The efficiency of the L-K method with respect to the analysis or compression problem can be illustrated as follows: Consider a zero-mean random process $\{x(t)\}, 0 \le t \le T$.

Any sample function, $x(t)$, can be described without loss of information by k numbers, according to Shannon's Sampling Theorem, where $k = 2wT$, w = bandwidth of $\{x(t)\}$, and where the k numbers are the sampled values of the signal at time intervals $1/(2w)$.

Alternatively, any such signal can be described to any desired degree of accuracy by n numbers, (n to be determined as described below) according to Loeve-Karhunen theory,* where the numbers are given by

$$a_j = \int_0^T x(t)\, \psi_j(t)\, dt, \quad j = 1, 2, \ldots, n. \tag{1}$$

$\psi_j(t)$ is the eigenfunction corresponding to the j'th largest eigenvalue, λ_j, of the integral equation

$$\int_0^T K_x(t, \tau)\, \psi(\tau)\, d\tau = \lambda\, \psi(t), \quad 0 \le t \le T \tag{2}$$

where

$$K_x(t, \tau) = E\left[x(t)\, x(\tau)\right] \tag{3}$$

* This neglects, for the time being, the storage required for the eigenfunctions.

Then, x(t) is said to "optimally approximated" in the sense that the error of approximation,

$$\theta = E\left\{\int_0^T \left[x(t) - \sum_{j=1}^{n} a_j \psi_j(t)\right]^2 dt\right\} \tag{4}$$

is minimized for any value of n by choosing the $\phi_i(t)$'s and a_j's to be the solutions of equations (2) and (1), respectively.

The value of n can be readily determined as a function of the tolerable approximation error in terms of the eigenvalues since

$$\theta = \int_0^T E\left[x^2(t)\right] dt - \sum_{i=1}^{n} \lambda_i = \sum_{n+1}^{\infty} \lambda_i \tag{5}$$

Therefore, for a given random process, x(t), and a tolerable error of approximation, a readily determined number of eigenfunctions must be computed and stored. Using these eigenfunctions as a basis for analyzing ℓ signals subsequently received from the random process, the following comparison can be made:

"Shannon" representation	ℓ x k numbers
L-K representation	ℓ x n numbers + n x k

(6)

where the ℓ x n numbers are the coefficients determined by (1) and the n x k numbers are the sampled representation of the eigenfunctions.

2. Geometrical Interpretation

The approximation problem has the following geometrical interpretation. Considering the random process x(t) to be a random vector $\underline{x}$ in an m-dimensional vector space Γ, the approximation $\hat{x}(t) = \sum_{i=1}^{n} a_i \phi_i(t)$ falls in a n-dimensional subspace of Γ, which shall be denoted by Γ_n. The approximation problem is then the problem of finding the orientation of Γ_n which minimizes the approximation error $\theta = E\left[|\underline{x} - \hat{x}|^2\right]$

It is readily shown,[6] that θ is minimized by finding the orientation of Γ_n which on the average maximizes the squared length of the projection of $\underline{x}$.

Although the Loeve-Karhunen theory has been developed in the context of integral equations and continuous time signals, any digital implementation of the method necessarily demands that the continuous signal be replaced by a time-sampled representation. As noted above, this replacement can be accomplished with no loss of information if an upper bound on the bandwidth of the signal is known, a condition which certainly can be satisfied in any class of biological measurements of interest. Therefore, for all practical purposes Loeve-Karhunen Analysis can be considered equivalent to its discrete analog, which is well known in multivariate statistical theory as Principal Component Analysis.[7]

3. Matrix Formulations

Principal Component Analysis is concerned with the properties of the eigenvalues and eigenvectors of a scatter matrix which can be construed to be the autocovariance matrix of a time series. (In Ref. 1 this matrix is also referred to as the "density matrix" of the autocorrelation function.) This analogous representation can be stated as follows: Given a zero mean random process $\{x(t)\}$, $0 < t < T$, sampled at M equal intervals so that $x_{ki} = x_k(iT/M)$. An autocovariance matrix can be computed with elements

$$A_{ij} = \frac{1}{M} \sum_{k=1}^{n} x_{ki} x_{kj}. \tag{7}$$

Then the matrix equation

$$\underline{A}\ \underline{\psi} = \lambda\ \underline{\psi} \tag{8}$$

is analogous to (2) and the eigenvectors $\underline{\psi}$ and eigenvalues λ enjoy properties analogous to those described above.

Principal Components Analysis is more widely known than L-K Analysis and has been applied to the processing of biological signals.[8-11,15] A semantic difficulty exists in that a variety of names are used by various researchers in referring to identical or similar techniques. In addition to Loeve-Karhunen and Principal Components, Principal Factors, Discriminant Analysis and others are mentioned in the literature. In Section VIII an attempt will be made to clarify the relationship between the various techniques.

4. Experimental Results

A problem which continually arises when evaluating various techniques for analyzing and classifying signals of biological origin is the following: Are the errors caused by the shortcomings of the technique; or because the signals were incorrectly labeled, or because the effect of the attribute being studied is dominated by some other variable attribute, or insufficient training set, etc.

As an effort to decouple these errors, the strategy followed has been to generate random signals for preliminary checking of each evaluation procedure. The ability of the procedure to correctly classify these synthetic random signals gives no index of its effectiveness or genuine biological signals, but clearly is a necessary condition for the same.

The evaluation procedure followed is as shown schematically in Fig. 1. The ensembles $\{x_i^{A}(t)\}$ and $\{x_i^{B}(t)\}$ are synthesized in the computer (IBM 7094). Each sample function of these ensembles is measured in two ways: the measurements shown here are 1) the power content in various frequency bands $p_i^{(\nu)}$, and 2) the Loeve-Karhunen coefficients, $c_i^{(\nu)}$. We desire a figure of merit which

indicates the discriminating power of the alternate measurement sets relative to the input ensemble $\{x_i(t)\}$.* This figure of merit is defined as shown on Fig. 1, where μ and σ are measured along the Z axis. The Z axis has been located by a discriminant analysis technique to provide the optimal separation of the projections of the two sample clusters.

Such experiments were conducted on two groups of computer-generated time series and on spontaneous EEG's of one subject with eyes alternately open and closed. The results are shown in Table I.

TABLE I

EXPERIMENTAL RESULTS

Figure of merit	Computer-generated time series	Spontaneous EEG eyes open/eyes closed
Band-pass powers (six measurements)	8.7	3.4
L-K analysis (six measurements)	11.3	4.7
L-K analysis (one measurement)	9.2	2.8

†The power measurements were made with respect to the six frequency bands: 1.5- .5, .5-7.5, 7.5-9.5, 9.5-12.5, 12.5-17.5, and 17.5-25 cps; which is one of the many partitionings of the frequency spectrum which have been employed by EEG researchers.

*See Reference 25 for discussion of this figure of merit.

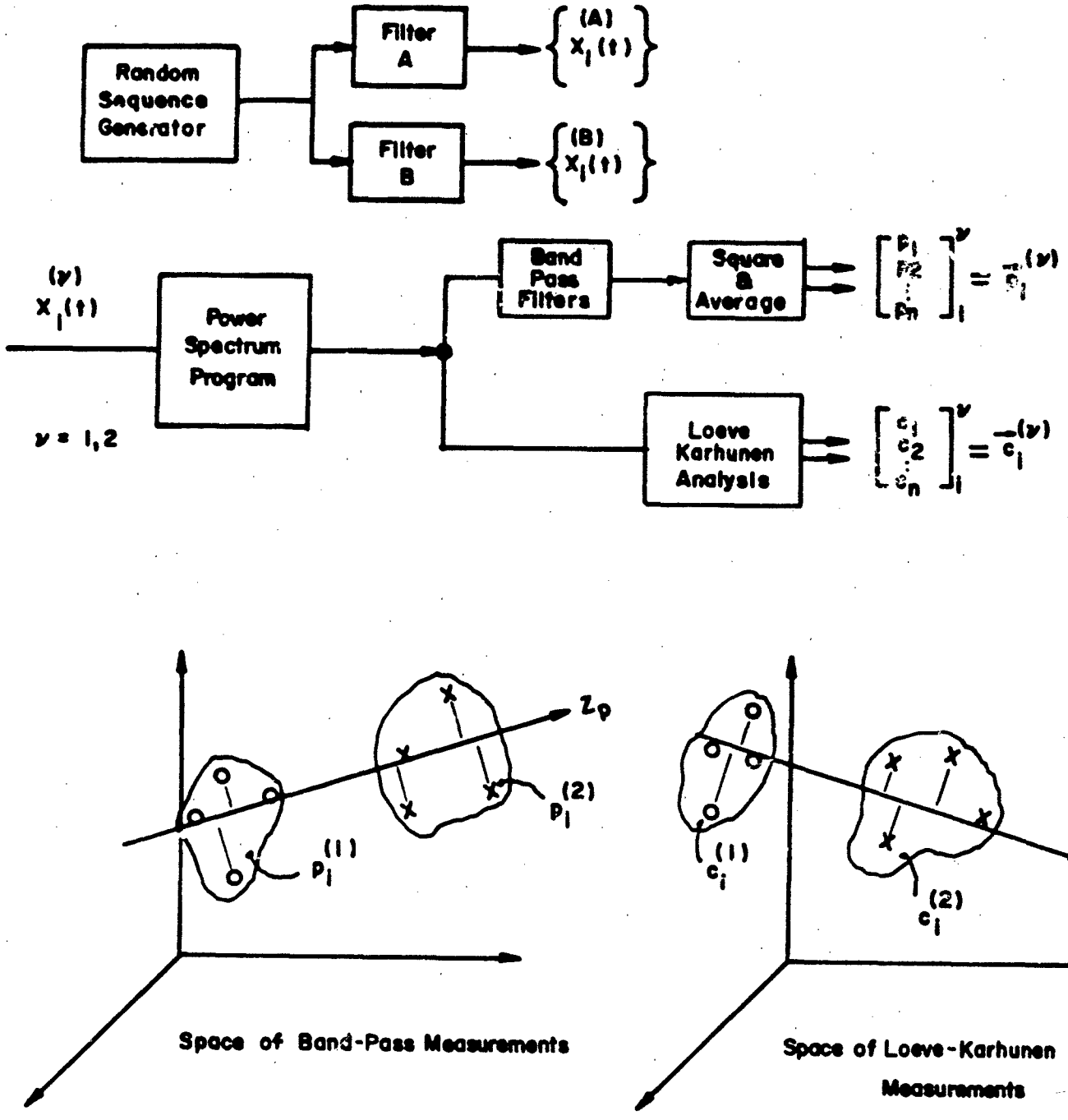


Fig. 1 The L-K Evaluation Procedure

5. Shape of the Eigenfunctions

The Loeve-Karhunen (or Principal Component) Analysis may be thought of as a generalized spectral representation of a random process. In this generalized representation the components are not limited to the family of sinusoids as in Fourier analysis but instead are, as has been described, chosen on the basis of economical approximation.

In addition to the efficiency of the L-K representation, another less well-known advantage may result from the unique properties of the method. This advantage follows from the fact that the method systematically finds the largest constituents of a random signal, subject to the constraint that their coefficients be uncorrelated. If then, an assumption of gaussianness can be justified, the eigenfunctions can be conceived as outputs of (statistically) independent mechanisms. Therefore, a research worker confronted with the problem of analyzing a complex waveform comprised of the sum of waveforms generated by a number of independent sources may be able to associate physical meanings with the various eigenfunctions.

Figures 2 and 3 show the first ten eigenfunctions of the computer-generated signals and the EEG's, respectively. It is interesting to note that the eigenfunctions of the EEG are not, in general, sinusoidal but do exhibit periodicities closely related to the well-known alpha, beta, and theta frequencies which have been postulated on the basis of Fourier analyses. Figure 4 shows the autocorrelograms from which the functions of Figure 3 were derived.

6. Inadequacy of L-K for Classification

Despite the advantage indicated previously relative to the conventional methods, Loeve-Karhunen analysis is clearly suboptimal with respect to the classification problem, since it does not make use of available information concerning the correlations within the several classes. In order to clarify this point, consider the hypothetical set of data represented in Figure 5 in which each point represents a sample time function of the state indicated, and three dimensions of an m-dimensional vector space are shown.

As described above, the effect of Loeve-Karhunen analysis is to define a subspace of any dimension $n < m$ such that the sample points, when projected on this subspace, are dispersed maximally

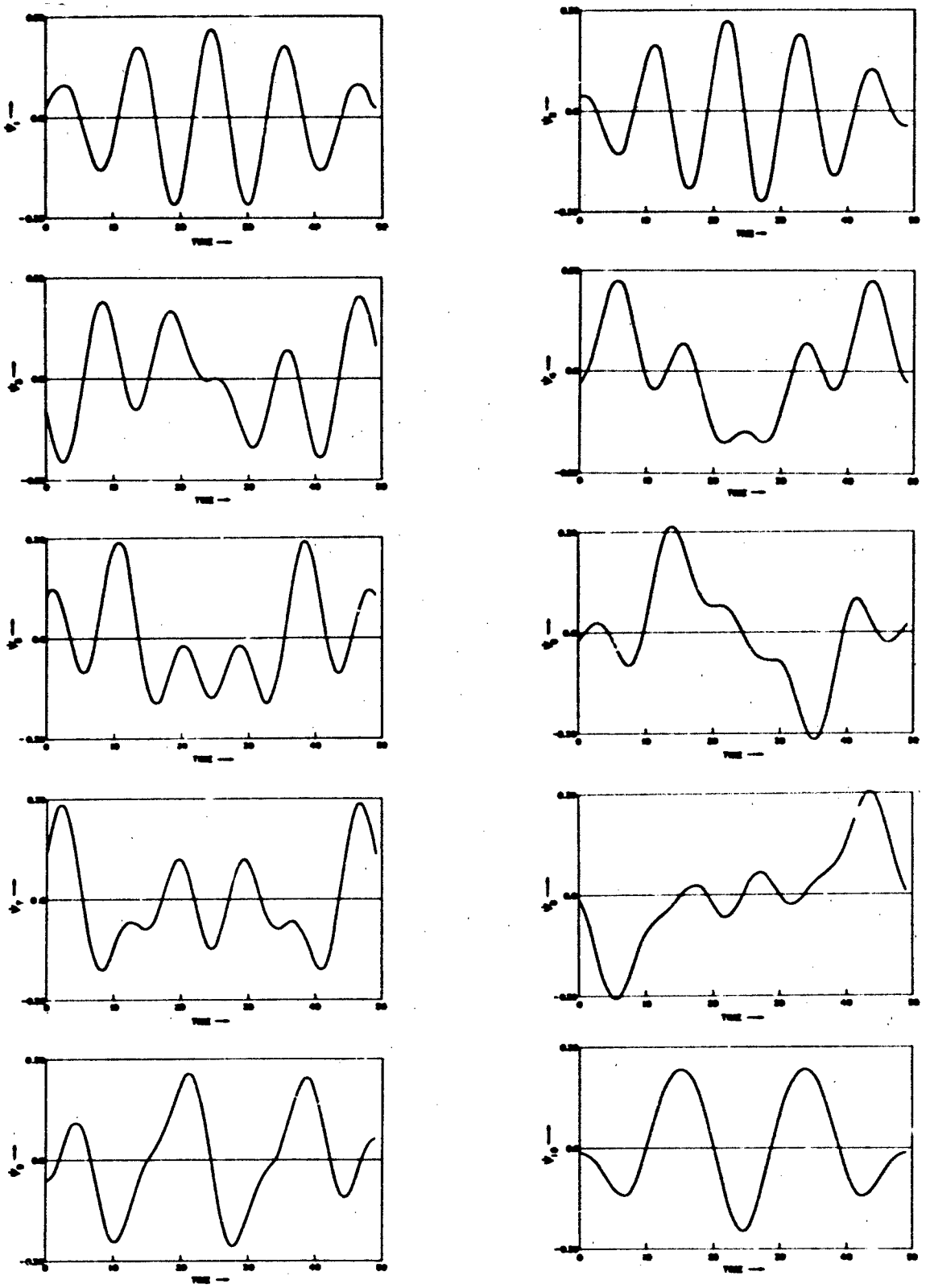


Fig. 2 Eigenfunctions of computer-generated time signals

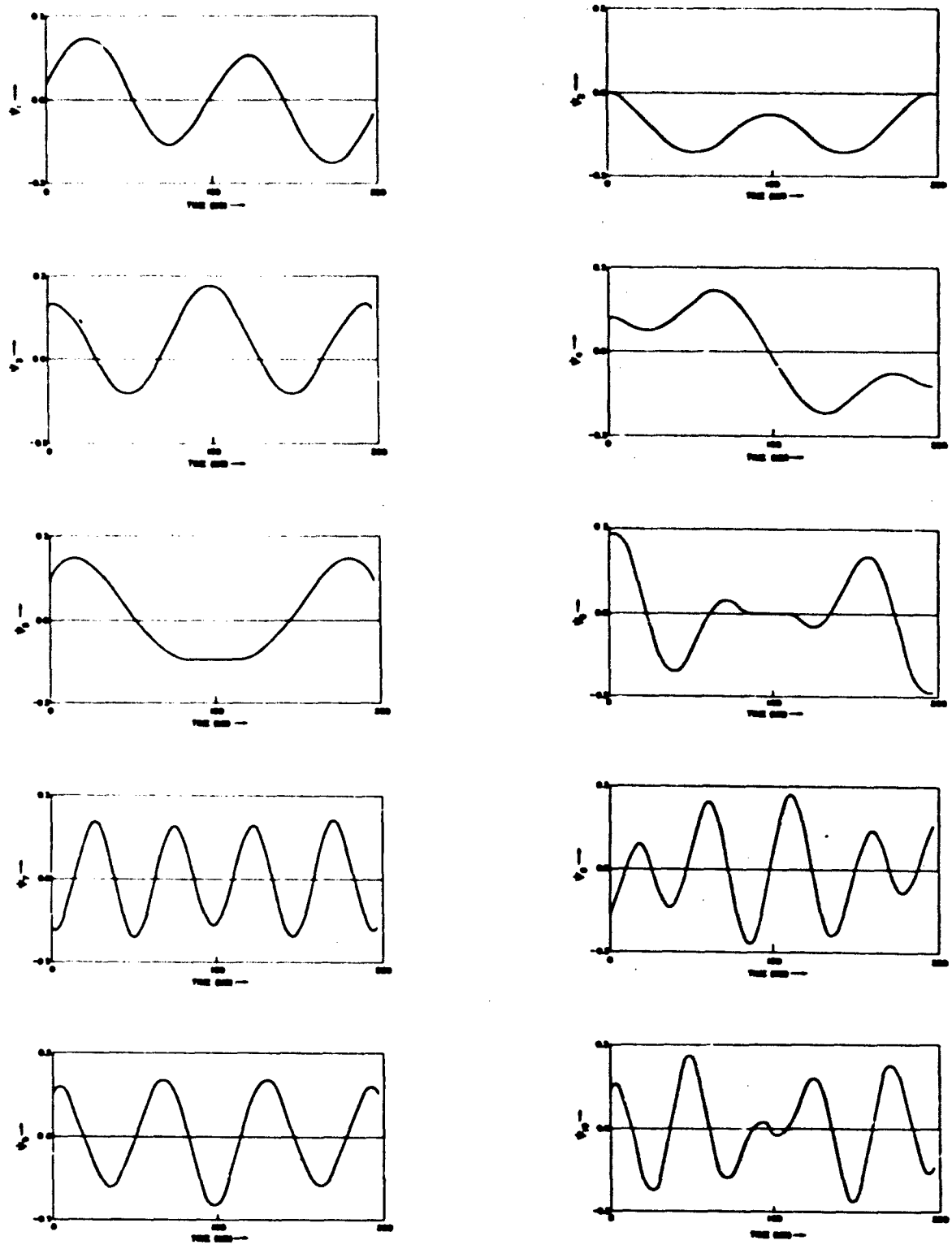


Fig. 3 Eigenfunctions of spontaneous electroencephalograms

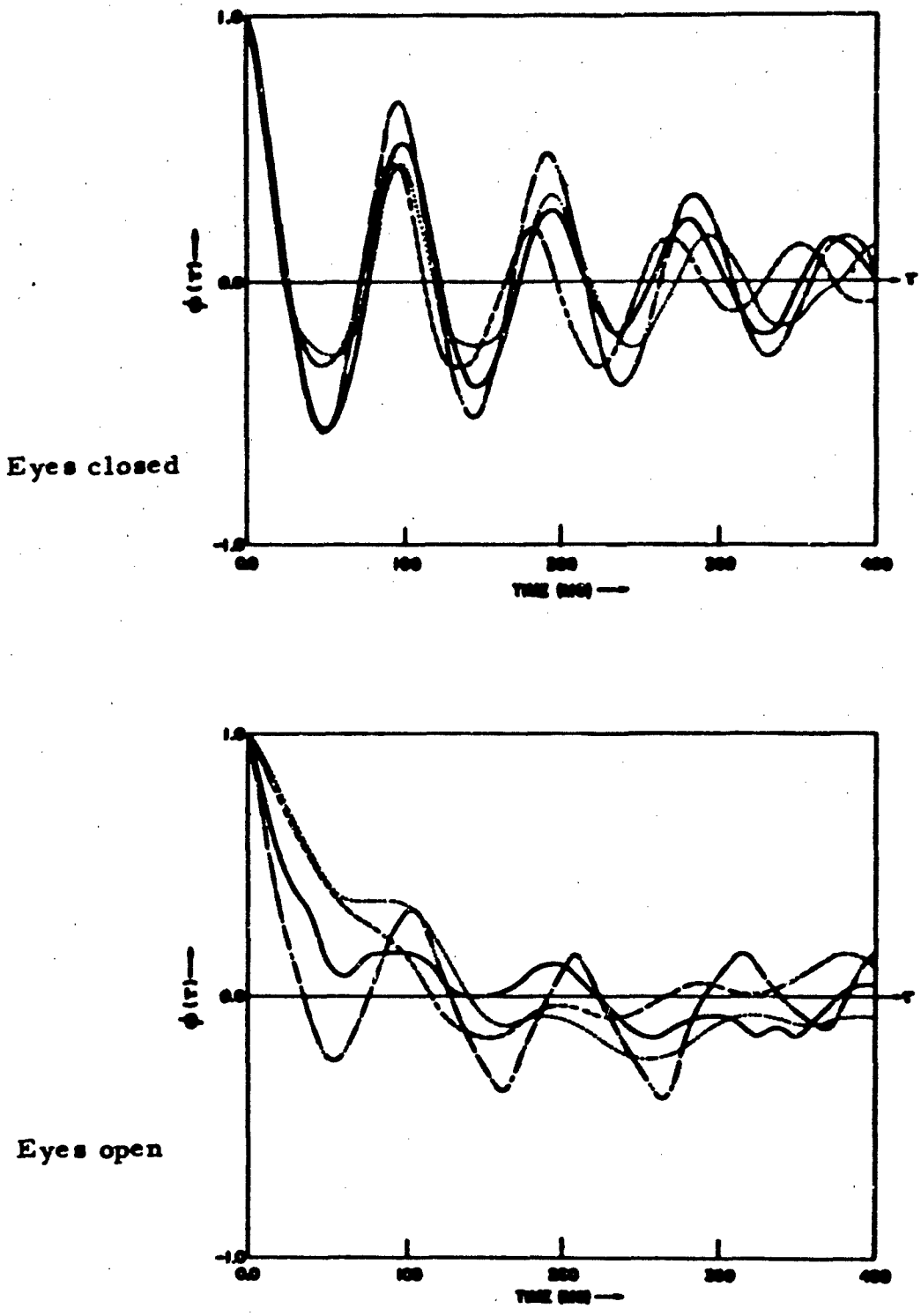
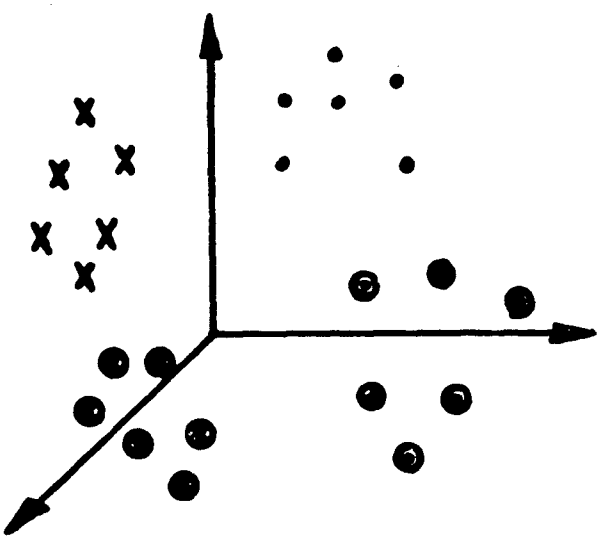


Fig. 4 Autocorrelograms of spontaneous electroencephalograms

in the mean-square sense, without regard to any inherent grouping of the points. It follows, therefore, that L-K treatment of the data as shown in Figure 5 will be the same whether the purpose of the processing is to infer the state of the eyes, or the level of alertness. It is evident that an optimal method would necessarily be a function of the use for which it is intended. Various means of accomplishing this are discussed in the following section.



	Eyes Open	Eyes Closed
Alert	X	•
Drowsy	●	●

Fig. 5 Geometrical Representation of Data

SECTION III

DISCRIMINANT ANALYSIS

1. Discussion

Discriminant Analysis differs from L-K physically, in that it assumes the existence of a set of labeled data, i.e., data whose correct class membership is known. This class information is exploited in that the sample covariance of the subgroups are computed as well as that of the aggregate. Then, formally, where L-K or Principal Components is based on the solutions of the simple eigenvalue problem

$$[\underline{A} - \lambda \underline{I}]\ \underline{\phi} = 0\ ,$$

Discriminant Analysis is based on the solutions of a generalized eigenvalue of the form

$$[\underline{B} - \lambda \underline{C}]\ \underline{\phi} = 0\ ,$$

where B and C are functions of the aggregate and subgroup covariances respectively, the precise definitions of which will be presented in Section VIII.

The solution of the simple problem defines a subspace which maximizes the scatter* of the sample points. The solution of the generalized problem defines a subspace which maximizes the scatter of the groups, keeping the within group scatter constant.

In contrast to the simple problem, the solution of the generalized problem requires matrix inversion. Therefore a singularity problem arises when the sample size is less than the dimensionality of the original space.

We have evaluated experimentally the following three different means of circumventing the singularity problem. The use of the Moore-Penrose Generalized Inverse for inverting the singular within sample scatter matrix; the 'H' inverse, a method suggested

---

* See Section VIII 1 for the definition of "scatter."

by T. J. Harley and described in Section VIII 3; and finally the method suggested by us of using the L-K method to reduce dimensionality and then applying D. A. The D. A. program has been provided with the option of taking the 'H' inverse and the program described as program G accomplishes L-K followed by D. A.

On the basis of the experiments we have performed, L-K followed by D. A. gave best results. As explained in Section VIII 2, D. A. can also be accomplished via prewhitening followed by K-L.

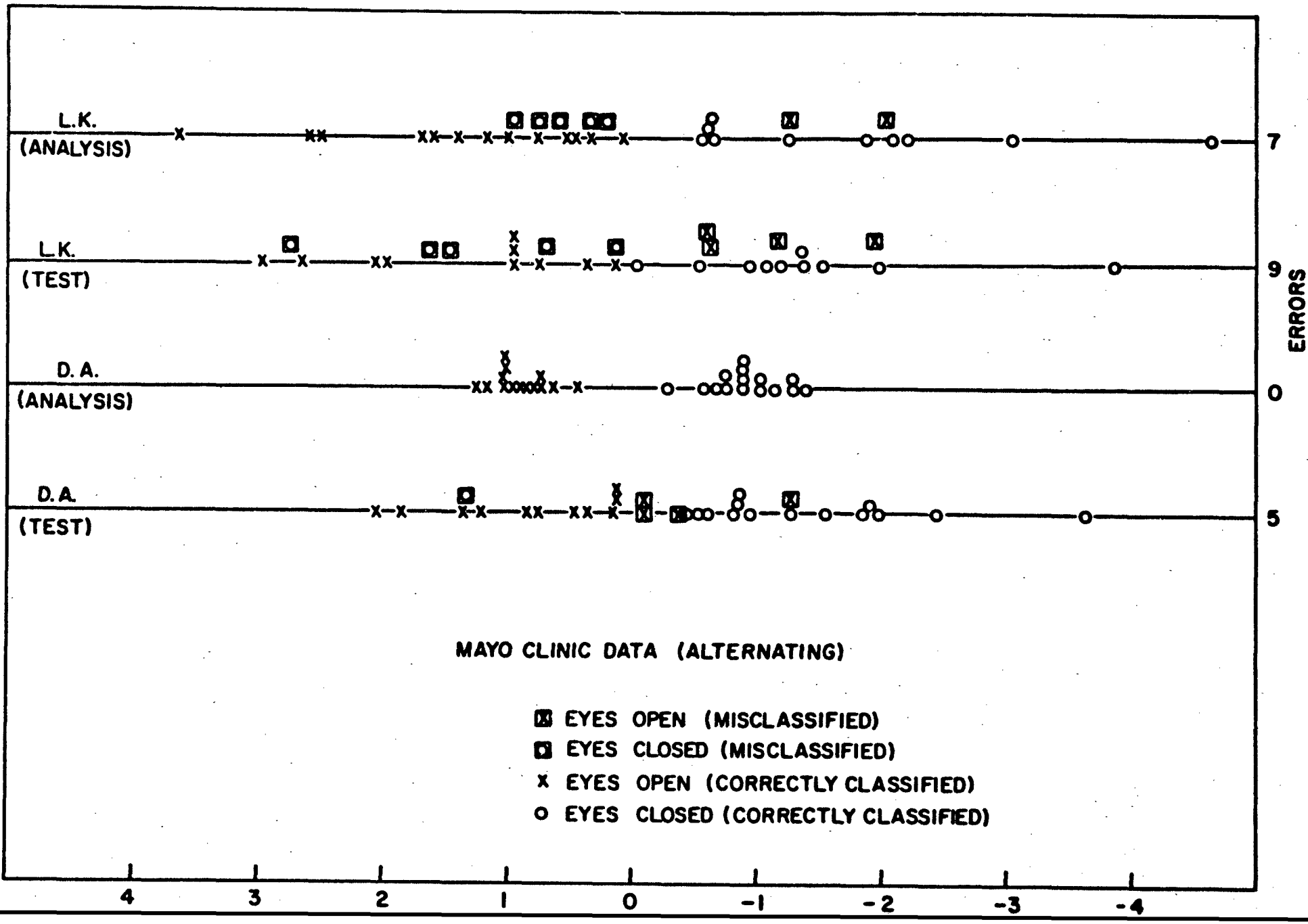
2. Experimental Results

The following experiment was conducted on a set of 60 visually evoked EEGs recorded at the Mayo Clinic. Thirty samples were recorded with subject's eyes open, 30 with eyes closed. Each set of 30 was divided so that samples 1, 3, 5, ..., 29 comprise an analysis set and samples 2, 4, 6, ..., 30 comprise a test set, where the numbers indicate the time sequence of occurrence. (This division of the samples was chosen to minimize the effect of any nonstationarity of the data.) Using various methods, then, the analysis data were used to define the directions of the corresponding lines on which sample points could be projected from the original 95 point vector space. The separation of the test data when projected on one or another of these lines, provides a measure of the efficacy of the respective methods by which the directions are defined.

Figure 6 shows the result of projecting the analysis and test data, first on the line by L-K Analysis, and then on the line defined by Discriminant Analysis. Because of the singularity problem, the Discriminant Analysis is performed on the data after it has been reduced from 95 to 25 dimensional space L-K.

Figures 8 through 11 represent the data used in this experiment and the marked signals, i.e., records 2, 10, 12, 15 for eyes open test data and record 5 for the eyes closed test data, represent the errors indicated on Figure 6.

*The computer programs are described and listed in reference 27. Copies of the programs may also be obtained from Mathematics and Analysis Branch, Aerospace Medical Research Laboratories, Wright-Patterson Air Force Base, Ohio.


L.K.
(ANALYSIS)
7
L.K.
(TEST)
9
D.A.
(ANALYSIS)
0
D.A.
(TEST)
5
ERRORS
MAYO CLINIC DATA (ALTERNATING)
EYES OPEN (MISCLASSIFIED)
EYES CLOSED (MISCLASSIFIED)
EYES OPEN (CORRECTLY CLASSIFIED)
EYES CLOSED (CORRECTLY CLASSIFIED)
4
3
2
1
0
-1
-2
-3
-4

The results demonstrate the characteristic tendency of the L-K projection to be maximally dispersed without necessarily effectively separating the classes of interest. The D.A., on the other hand, demonstrates the tendency to separate the classes of interest.

Figure 7 illustrates the case when L-K analysis would indicate a line of projection which does not effectively separate the two classes but D.A. does find the appropriate line of interest. If the two cluster centers, in this example, were far apart L-K would indicate an appropriate line of projections.

A third experiment was conducted which verified the formal equivalence between D.A., and prewhitening followed by L-K.

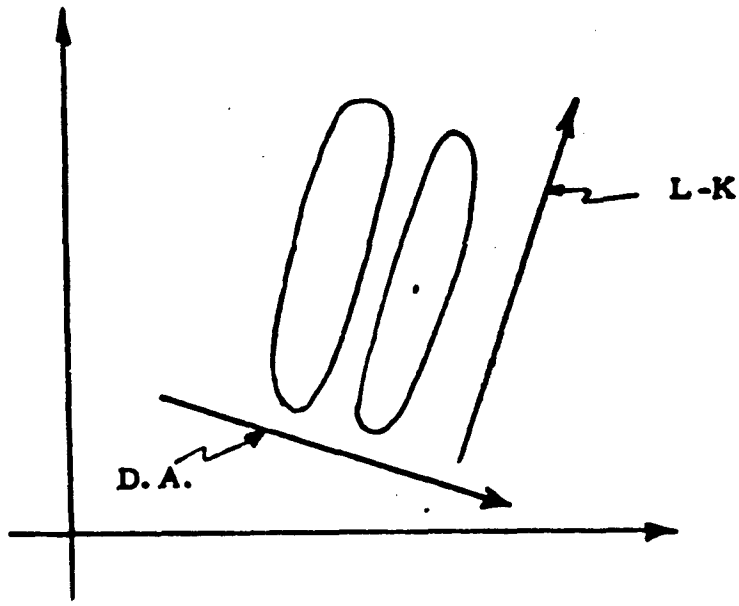


Fig. 7 Example of L-K Inadequacy

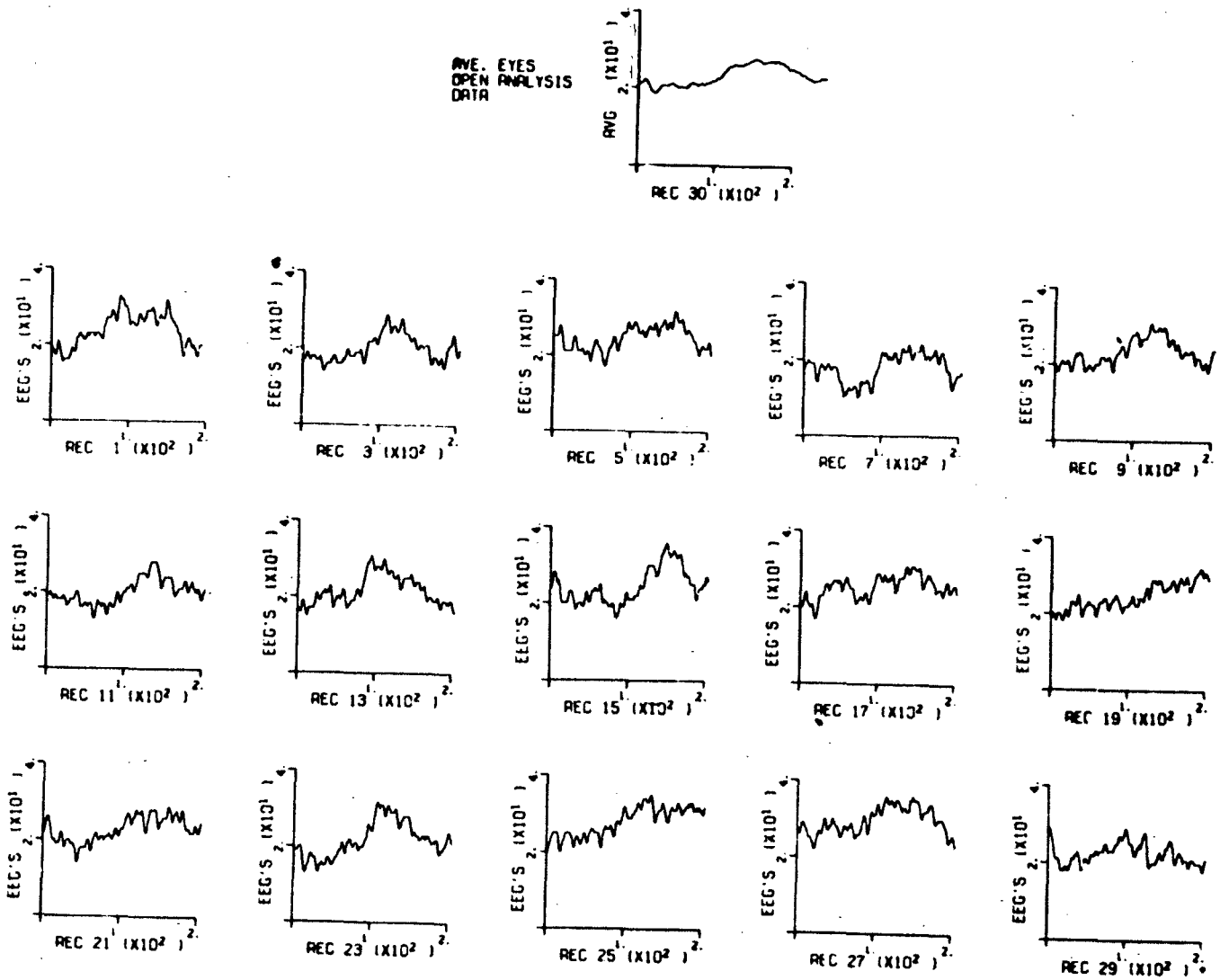


Fig. 8 Eyes open, analysis.

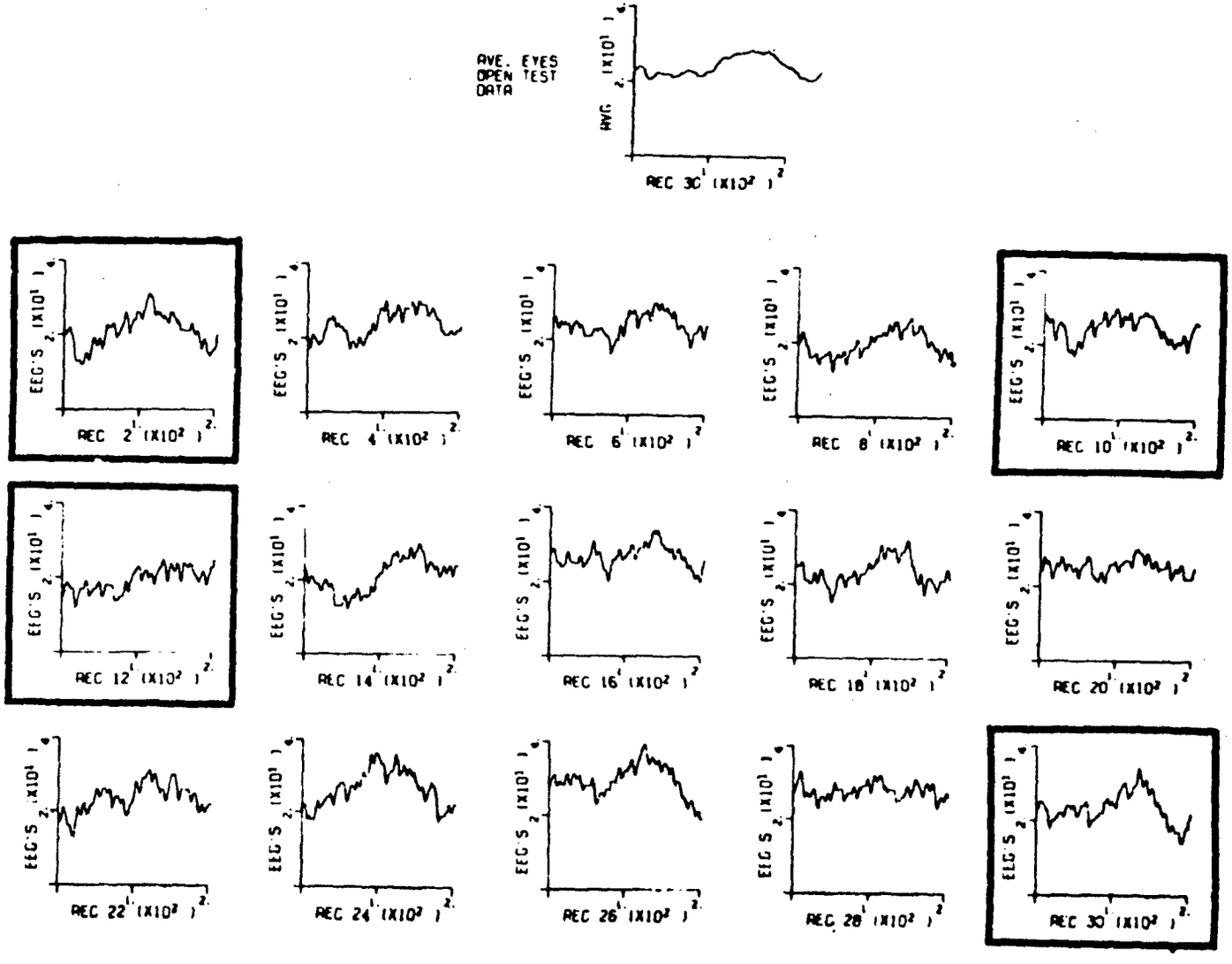


Fig. 9 Eyes open, test.

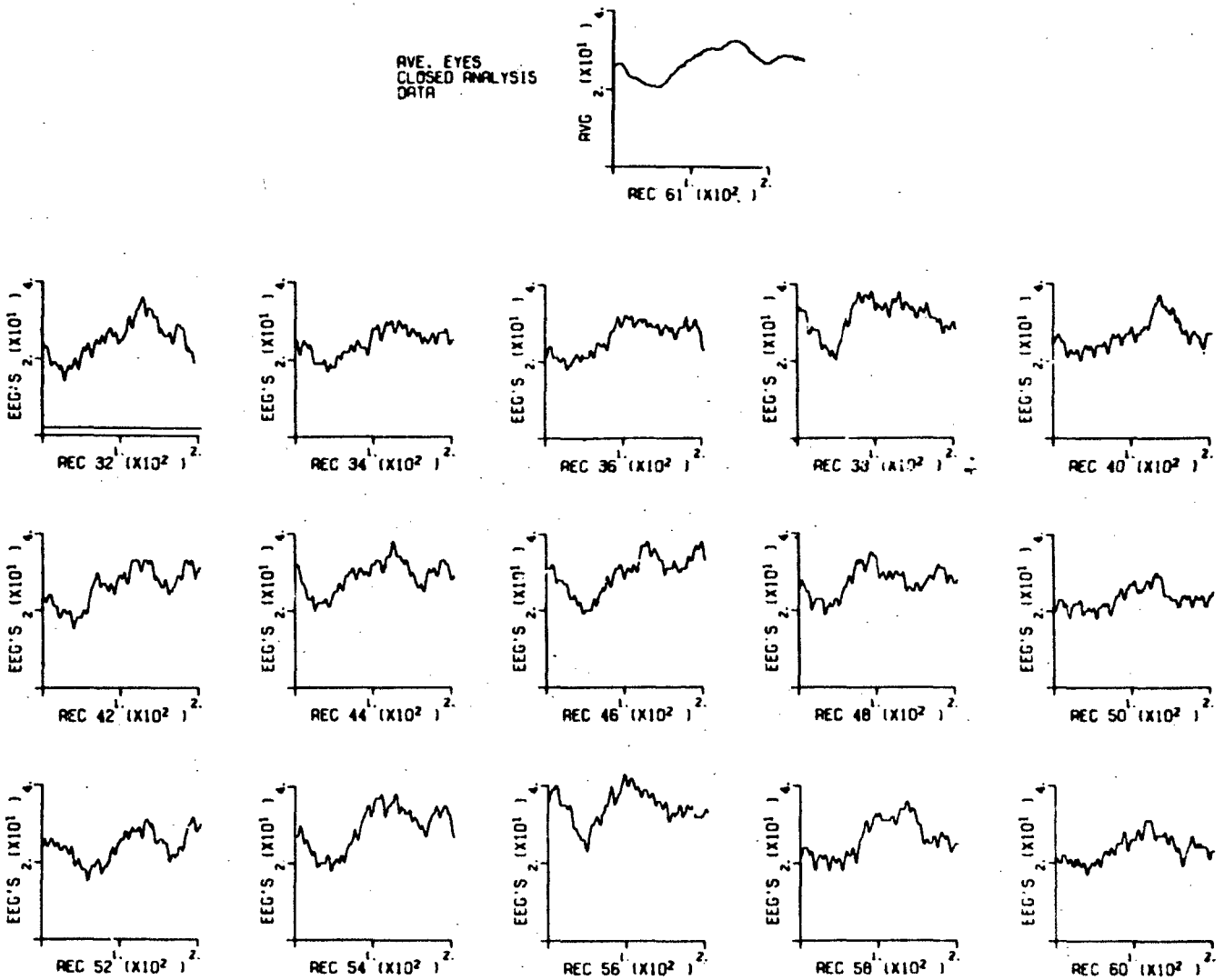


Fig. 10 Eyes closed, analysis.

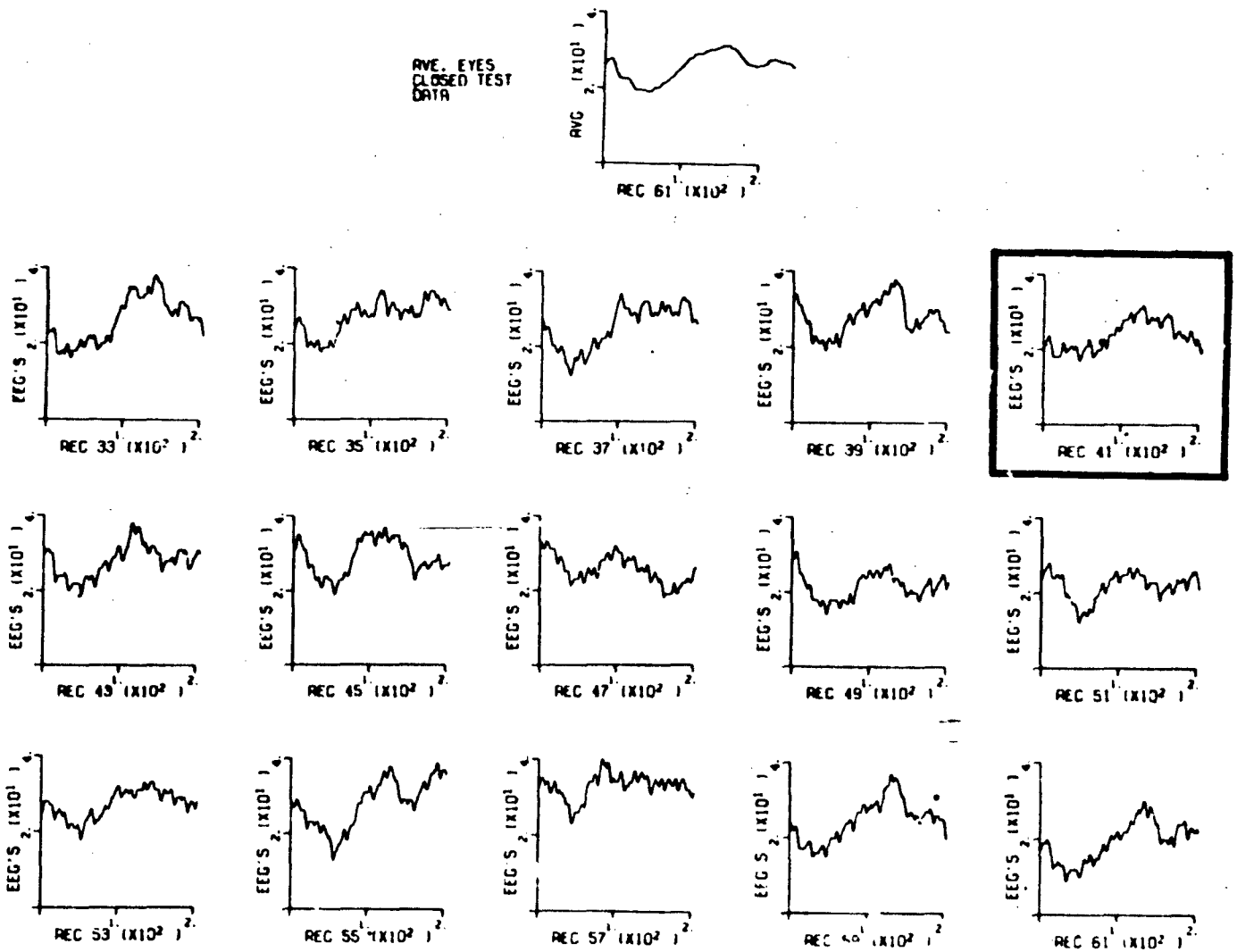


Fig. 11 Eyes closed, test.

SECTION IV

EVALUATION OF DECISION PROCEDURES - BAYES VS MINIMUM DISTANCE

The Bayes Decision Procedure has certain desirable properties, such as the ability to nonlinearly partition the vector space and to provide minimum, expected loss decisions under certain conditions. However, its implementation requires either the fitting of multi-dimensional density function, or the discretization of the measurement space and the empirical determination of $M(NK + 1)$ probabilities, where M is the number of classes, N is the number of measurements, and K is the number of quantization levels. It is difficult to predict the variance of these tabulated probabilities, but it may be noted that many hundreds of samples are usually used in making such a tabulation in character recognition work.

Assuming that such sample sizes may often be prohibitive in biological work, an attempt has been made to evaluate Bayes relative to other decision methods for small samples, and to provide an alternate decision program for such cases.

The Bayes decision program was tested on the six-dimensional sample points produced by the power band and the L-K measurements of the computer-generated time series. As a consequence of the small number of samples available for training (five for each class in each case), only two of ten test samples could be classified correctly in one case and none of ten in the other; i.e., no decision in 18 of 20 test samples. Although this sample size is extremely small, the point is that simple decision procedures based on distance measures can discriminate reliably under such conditions. For example, in this case, the sign of the first L-K coefficient is a sufficient criterion, with ample margin of safety, to classify all the test samples. The results indicate that an alternative decision procedure is needed for problems involving small sample size. The discriminant analysis procedure described in the previous section is one such approach. This approach, when used in conjunction with some algorithm for setting threshold and a means of circumventing the singularity problem, is recommended for small sample decision problems.

SECTION V

EVALUATION OF POWER SPECTRUM PROGRAM

In the program supplied with Reference 1, the power spectrum is obtained by computing the sample autocorrelation function and then taking its Fourier transform. This program was used many times during this year's experimentation, both on synthetic time series and on spontaneous EEGs. Investigation of certain spurious components of the computed spectra indicated the advisability of making the following restriction and revision:

1) The autocorrelation function ρ_y (y) should not be computed for lags greater than 10% of the interval X;

2) A "gate function" g (y) , providing more smoothing than does g (y) = 1, $0 < y < X$, must be used. Good results have been obtained with the "hamming" function[12].

With these changes, the program gave generally satisfactory results except that the spectra thus derived are still extremely sensitive to sampling error in the autocorrelogram. This sensitivity has been commented on by other investigators[11,12]. An alternative method of power spectrum determination exists which does not require the autocorrelation function, and which therefore may yield more consistent results[13]. This "direct" method was used, successfully, to compute the spectra of some EEG data. However, it was decided not to carry out any comparative evaluation of the two methods but, rather, to attempt to develop a testing procedure which eliminates the need for computing the power spectrum altogether. The results of this effort are reported in the following section.

SECTION VI

TIME-DOMAIN PROCESSING OF STOCHASTIC SIGNALS

Certain classes of neurological signals are stochastic: i.e., unpredictable in detail, with only certain statistical properties of the signal knowr Spontaneous EEGs exemplify such signals. The methods of orthogonal analysis and classification developed under this contract do not apply naturally to stochastic signals, since the method of generating coefficient assumes the signal to have deterministic content with respect to some point of time reference. A standard procedure of transforming stochastic signals, if stationary, into deterministic signals is to compute the power spectrum or auto-correlation function, and then to use that as the signal to be analyzed rather than the original time function. (This was the procedure suggested in Ref. 1.) This artifice has the disadvantages of consuming computer time, introducing additional error, and producing eigen-functions which are not interpretable in terms of the time functions.

A time-domain procedure which eliminates these shortcomings has been derived and programmed (see Sections VIII 6). This procedure can be implemented in real time via a bank of filters matched to the eigenfunctions used in conjunction with delay lines, multipliers and integrators; or via a relatively simple digital program.

SECTION VII

EXPERIMENTAL FACILITY FOR COLLECTING EEG DATA

In order to obtain empirical data upon which these analyses might be tested, equipment was designed and built to elecit responses from the human visual cortex by photic stimulation. By placing small, high powered collimating and Maxwellian lenses close to a glow modulator tube (Sylvania R1166), it is possible to produce a uniform field of 72,000 feet. Lamberts covering a 60^{o} solid angle (Fig. 13). This is sufficiently bright that a beam splitter reducing the brightness to 24,000 feet. Lamberts might be used to provide the option of superimposing another, independent field on the first, and to permit monitoring of the two fields by photocell. The second field is at present being used to supply a fixation point consisting of a 15-45 Pinlite (Kay Electric), reduced in intensity and reddened by a series, variable resistor.

The advantages of the Maxwellian view are threefold: (a) it directs into the eye all of the light incident within its area instead of scattering it as a reflecting surface or ground glass would do; (b) by concentrating most of the rays of light on the center of the pupil, it increases their effectiveness by virtue of the Stiles-Crawford effect; (c) but most important, it focuses an image of the source within the 2 millimeter area in the center of the pupil within which the pupil cannot constrict; thereby eliminating variations of retinal illuminance resulting from pupillary oscillations. It does present a problem in measuring the equivalent luminance, however, especially in this case where the lens is only about 7 millimeters from the cornea. The solution adopted here was to occlude half of the collimating lens and substitute in its place light incident through the second half of the beam splitter. Once the two have been matched, the second field can be measured by conventional means.

The luminous flux of a glow modulator is a nearly linear function of its anode current, but along with this change in flux is a change in hue. Consequently, for visual experiments it is best not to vary the anode current for strictly comparable results. This problem was circumvented in the present system by delivering the light in 90 microsecond pulses of uniform shape and amplitude. By Bloch's law the effect of a given amount of light energy is independent of its distribution over time, up to a certain critical duration. Thus the apparent

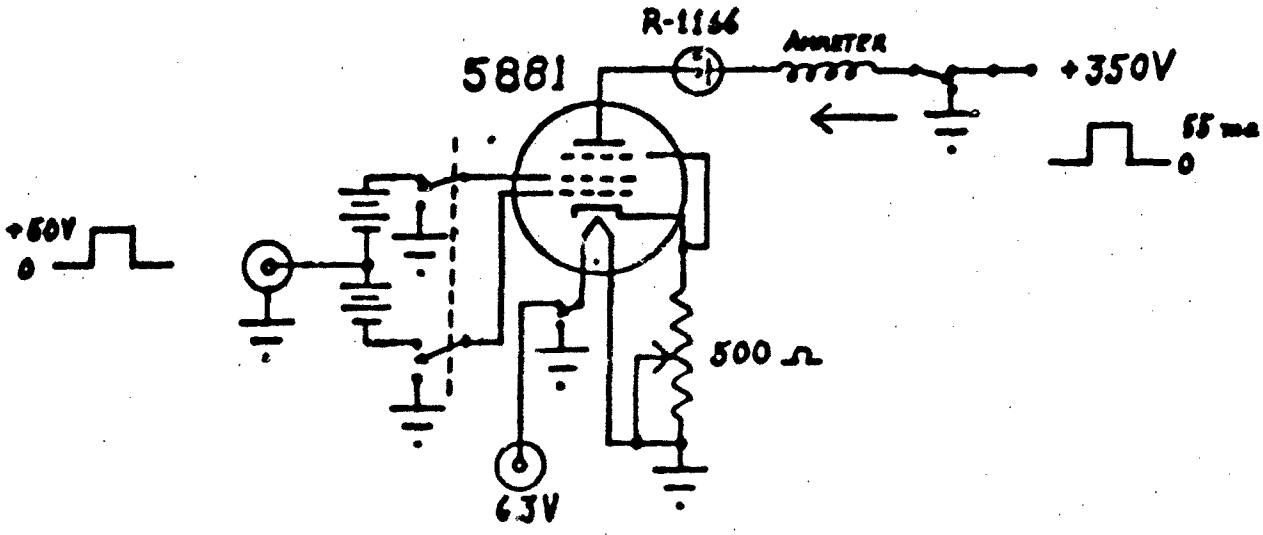


Fig. 12 Control Circuit for Glow Modulator.

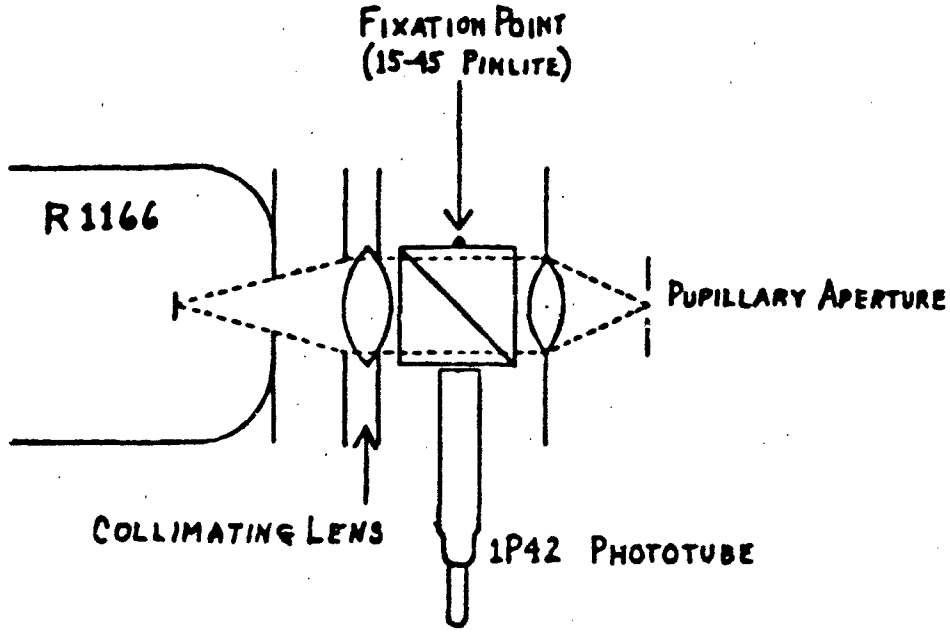


Fig. 13 Optical Stimulus

brightness can be varied by varying the number of pulses delivered within this critical duration, which varies from 10 to 100 milliseconds depending upon conditions. Thus, the equivalent luminance can be varied over a range of two to three log units by this means, and the range can be shifted about by the variable resistor connected to the cathode of the pentode in Fig. 12. The range can also be extended downward another 0.6 log units by delivering the pulses to the grids of the pentode with a device having faster rise time than the Tektronix 161 pulse generator used here. The tise time of the light output of the glow modulator itself is under 20 microseconds, and the time constant of decay is less than 2 microseconds. (This is measured with a 1P42 photocell in series with a Tektronix 545 oscilloscope, oscilloscope, shunted by a 10K resistor, a very fast system with spectral response approximately that of the human eye.)

The glow modulator and accessory equipment are mounted on a three-coordinate manipulator for easy positioning of the image within the pupil. The subject is provided with adjustable bit-board and two-point head rest.

The evoked response is recorded from the scalp one inch above and one inch to the right of the occipital protuberance, with the left ear lobe as reference. The data logging system has been described in the draft report submitted to the contracting agency at the termination of the previous contract year.

SECTION VIII

MATHEMATICAL REPORT

1. Discriminant Analysis: Its Theoretical Justification and Relation to Loeve-Karhunen Analysis

The basic assumption underlying probabilistic classification into classes is that there exists for each class a multivariate probability distribution, the member $(x_1^i, \ldots, x_k^i)$ of class i being considered a sample from a population which is distributed in a k-dimensional space according to a certain probability distributions. Our approach to the problem is that the only knowledge we have about the distributions is that which can be inferred from training samples and no assumptions are made about the functional forms of the distributions. Let us consider linear classification procedures first, and limiting the discussion to two classes, define linear procedures formally (the classification to a finite number of classes will be accomplished by a repetition of the pairwise procedure an appropriate number of times). Let $\underline{b}$ ($\neq 0$) be a column vector (of k components) and c a scalar. An observation (signal) $\underline{x}$ is classified as coming from the first population if $\underline{b}' \cdot \underline{x} \leq c$ and as from the second if $\underline{b}' \cdot \underline{x} > c$ (the symbol $'$ stands for transpose). The vector $\underline{b}$ and the constant c are to be chosen to provide maximum discrimination, in some specified sense, between the two classes. An example of such a procedure which was originally considered by Fisher[21] and the method of statistical analysis which was developed from the solution of the problem is called discriminant analysis. A good exposition of this solution is given by Wilks.[7] Discriminant analysis solves the following problems: Suppose we have two samples from k-dimensional distributions, these can be represented geometrically as two sample clusters in Euclidean k-space. (Sample 1 contains n_1 points and sample 2 contains n_2 points.) We want to project these two sample clusters orthogonally onto a line so that the variation between the two projected samples is as large as possible, relative to the variation within the two projected samples. The terms "variation between" and "variation within" as they are defined in the solution are presented below. Using the notation in Wilks,[7] suppose we are given a sample of size n from a k-dimensional space $(x_{1\xi}, \ldots, x_{k\xi})$ $\xi = 1, \ldots, n$ and let $\bar{x}_i = \frac{1}{n} \sum_{\xi=1}^{n} x_{i\xi}$

(i= 1 ,..., k)then the matrix U with elements $u_{ij} = \sum_{\xi=1}^{n} (x_{i\xi} - \bar{x}_i)(x_{j\xi} - \bar{x}_j)$ will be called the internal scatter matrix and its determinant.

$|U|$ the internal scatter of the sample. It will be noted that U is non-singular if and only if the n point in the sample does not lie on a hyper-plane of less than k-dimensions. Now suppose

$(x_{1\xi_1}^{(1)}, \ldots, x_{k\xi_1}^{(1)})$ $\xi_1 = 1, \ldots, n_1$ and $(x_{1\xi_2}^{(2)}, \ldots x_{k\xi_2}^{(2)})$ $\xi_2 = 1, \ldots,$

n_2 are two samples. Let $\underline{x}^{(\gamma)} = (\bar{x}_1^{(\gamma)}, \ldots, \bar{x}_k^{(\gamma)})$, $\gamma = 1, 2$ be the vectors of means and $U^{(1)}$, $U^{(2)}$ the internal scatter matrices of the two samples respectively. Let $\underline{\bar{x}} = (\bar{x}_1 \ldots, \bar{x}_k)$ be the vector of sample composed of the two samples pooled together. And let $U^W = U^{(1)} + U^{(2)}$, the within-samples scatter matrix for the two samples. Note that geometrically U^W is the scatter matrix for the k-dimensional cluster one obtains by rigidly translating one sample cluster with respect to the other (without rotation) until the means of both samples coincide, and then pooling the two sample clusters together as a single cluster. Finally let $U^B = U - U^W$. For an arbitrary vector $(b_1, \ldots b_k)$ let

$$z_{\xi\gamma}^{(\gamma)} = \sum_{i=1}^{k} b_i x_{i\xi\gamma}^{(\gamma)}, \quad \xi_\gamma = 1, \ldots, n_\gamma \quad \gamma = 1, 2 \tag{9}$$

or using vector notations $z_{\xi\gamma}^{(\gamma)} = \underline{b}' \cdot \underline{x}_{\xi\gamma}^{(\gamma)}$.

Thus, $z_1^{(1)}, \ldots, z_{n_1}^{(1)}$ and $z_1^{(2)}, \ldots, z_{n_2}^{(2)}$, except for scaling, are

one dimensional samples obtained respectively by projecting the original k-dimensional samples onto a line whose direction cosines in the original k-dimensional space are proportional to $(b_1, \ldots, b_k)$. See Figure 14 for k = 2.

Let $\bar{z}^{(1)}$ and $\bar{z}^{(2)}$ be means of the two samples of z's and $\bar{z}$ the mean of the pooled samples.

$$\text{Let } S_w = \sum_{\gamma=1}^{2} \sum_{\xi\gamma=1}^{n_\gamma} (z_{\xi\gamma}^{(\gamma)} - \bar{z}^{(\gamma)})^2 \tag{10}$$

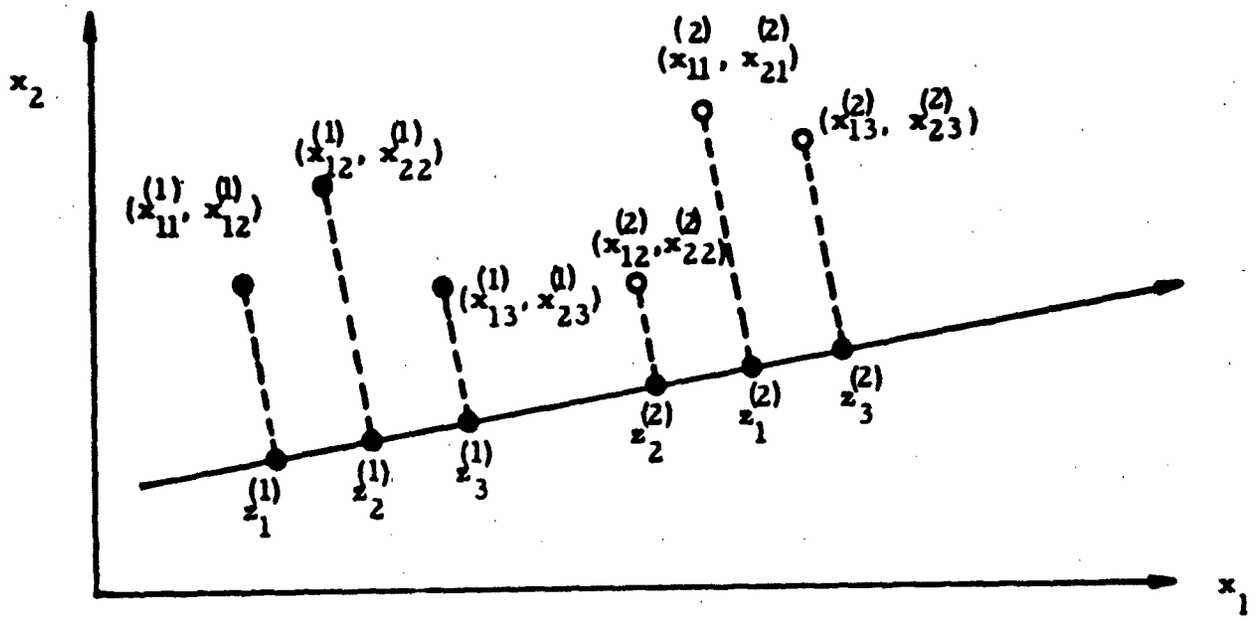


Fig. 14 A Projection of Two Group Clusters

$$S_B = \sum_{\gamma=1}^{2} n_\gamma \left(\bar{z}^{(\gamma)} - \bar{z}\right)^2 \tag{11}$$

Note that if S is the scatter of the grand sample obtained by pooling the two samples of z's then $S = S_w + S_B$. S_w is the within sample component, i.e., the sum of the squares of the n_1 possible segments $(z^{(1)}_{\xi_1} - \bar{z}^{(1)})$ generated by taking $\bar{z}^{(1)}$ and each $z^{(1)}_{\xi_1}$ in one sample summed with the sum of the squares of the n_2 possible segments $(z^{(2)}_{\xi_2} - \bar{z}^{(2)})$ generated by taking $\bar{z}^{(2)}$ and each $z^{(2)}_{\xi_2}$ in the other sample, and S_B is the between-sample component of S. It is also clear that

$$S_w = \underline{b}' U^w \underline{b},\ S_B = \underline{b}' U^B \underline{b} \text{ and } S = \underline{b}' U \underline{b}. \tag{12}$$

The problem is to determine $\underline{b}' = (b_1, \ldots, b_k)$ so as to maximize S_B (or equivalently to minimize $S_w/(S_w + S_B)$ for a fixed value of S_w).

The basic results concerning the solution of this problem may be stated as follows: The value of $(b_1, \ldots, b_k)$ say $(b_1^*, \ldots, b_k^*)$ which minimize the ratio

$$Q = \frac{S_w}{S_w + S_B} \tag{13}$$

so that S_w has a fixed value $D \neq 0$, is the solution of the equation

$$(U^B - \lambda_1 U^w)\,\underline{b} = 0 \tag{14}$$

where λ_1 is the nonzero root of the characteristic equation

$$|U^B - \lambda U^w| = 0, \quad \text{thus } \underline{b}^* = (U^w)^{-1} \cdot (\underline{\bar{x}}^{(1)} - \underline{\bar{x}}^{(2)}). \tag{15}$$

Let us note that if $U^w = I$ (I is the identity matrix) Q takes the following form

$$Q = \frac{S_w}{S} = \frac{\underline{b}'U^w\underline{b}}{S} = \frac{\underline{b}'I\underline{b}}{S} = \frac{\sum_{i=1}^{K} b_i^2}{S} \qquad (16)$$

minimizing Q while keeping $S_w = D \neq 0$ is equivalent to maximizing $\frac{S}{\sum_{i=1}^{k} b_i^2}$ keeping $\sum_{i=1}^{k} b_i^2 = D \neq 0$ and we can set D = 1.

We see that when $U^w = I$ the problem reduces to finding a line direction $\underline{b}' = (b_1, \ldots, b_k)$ which will maximize the scatter of the projected points of the two samples pooled together. The solution to this problem is Principal Component Analysis and $\underline{b}'^* = (b_1^*, \ldots, b_k^*)$ is the eigenvector corresponding to the largest eigenvalue in the solution of the equation $(U - \lambda I)\underline{b} = 0$. We notice, as mentioned before, that this last equation is identical to the equation which is solved to get the eigenvectors of the Loeve-Karhunen expansion. If the Loeve-Karhunen method would have been used to obtain the optimal line to project our samples on, for the purpose of discrimination, this line would coincide with the one obtained by the method of discriminant analysis when the within-sample scatter matrix U^w is the identity matrix. Looking at it geometrically, what discriminant analysis tries to do is to project the two samples onto a line so that the means of the two one-dimensional samples of points are as far as possible relative to the within sample scatter of these two one-dimensional samples. The Loeve-Karhunen method tries to project the pooled two samples on a line so as to maximize the distance among all the projected points. Since $U^w = I$ corresponds geometrically to a spherical within-group scatter, it is clear that in this case making all the projected points in the pooled sample as far as possible from each other is equivalent to maximizing the distance among the means of the two projected samples. See Figure 15.

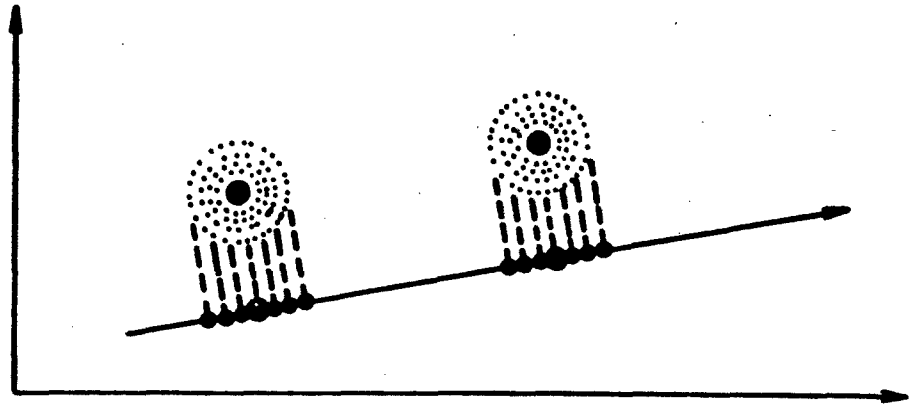


Figure 15 - A Projection of Spherical Clusters

2. The Use of Prewhitening Filters

The arguments presented in Section VIII-1 suggest that solving the discriminant analysis problem is equivalent to first making a change of coordinates, such that in the new set of coordinates the within sample scatter matrix is spherical, and then solving the Loeve-Karhunen problem i.e., finding the eigenvector corresponding to the largest eigenvalue. We shall derive the appropriate change of coordinates to accomplish this.

If we consider each sample vector as $(x_{1\xi_\gamma} - \bar{x}_1^{(\gamma)}, x_{2\xi_\gamma}^{(\gamma)} - \bar{x}_2^{(\gamma)}, \ldots, x_{k\xi_\gamma}^{(\gamma)} - \bar{x}_{k\xi_\gamma}^{(\gamma)}) = \underline{x}_{\xi_\gamma}^{(\gamma)'}$ $\gamma = 1, 2, \xi_\gamma = 1, \ldots n_\gamma$, i.e., subtract from each sample vector the sample mean, and denote the $(n_1 + n_2)$ x k matrix

$$= \begin{bmatrix} x_{11}^1, \ldots, x_{21}^1, \ldots, & x_{k1}^1 \\ \vdots & \vdots \\ x_{1n_1+1}^2 & x_{k\ n_1+1}^2 \\ \vdots & \vdots \\ x_{1\ n_1+n_2}^2, \ldots, \ldots, & x_{k\ n_1+n_2}^2 \end{bmatrix} \qquad (17)$$

then $U^w = X'X$.

If we make a change of coordinates such that the vector $\underline{x}_{\xi\gamma}^{(\gamma)}$ is transformed to the vector $\underline{y}_{\xi\gamma}^{(\gamma)}$ by the transformation $\underline{y}_{\xi\gamma}^{(\gamma)} = (U^w)^{-\frac{1}{2}}\ \underline{x}_{\xi\gamma}^{(\gamma)}$ then $Y = X[(U^w)^{-\frac{1}{2}}]'$ and if we denote the within scatter matrix in the new coordinate system as V^w we have

$$V^w = Y'Y = (U^w)^{-\frac{1}{2}} X'X[(U^w)^{-\frac{1}{2}}]' = (U^w)^{-\frac{1}{2}}(U^w)\ (U^w)^{-\frac{1}{2}}[(U^w)^{-\frac{1}{2}}]' = I \qquad (18)$$

We conclude that the appropriate change of coordinates for making the within-group scatter spherical is the transformation $\underline{y}_{\xi\gamma}^{(\gamma)} = (U^w)^{-\frac{1}{2}} \underline{x}_{\xi\gamma}^{(\gamma)}$.

In the language of communication engineers, the linear transformation $[U^W]^{-\frac{1}{2}}$ is called a prewhitening filter because of the terminology of white noise used to denote noise which has a covariance matrix I.

The fact that D. A. can be accomplished by prewhitening followed by L-K can also be demonstrated as follows:

The optimal line of projection $\underline{b}^*$ is found by the method of D. A. as the eigenvector corresponding to the largest (for two groups the non-zero) eigenvalue of the characteristic equation

$$|(U^B - \lambda U^W)| = 0 \tag{19}$$

or the solution of

$$(U^B - \lambda_1 U^W)\,\underline{b} = 0 \tag{20}$$

and then we obtain

$$z_{\xi\gamma}^{(\gamma)} = \underline{b}^{*\prime} \cdot \underline{x}_{\xi\gamma}^{(\gamma)} \tag{21}$$

where both $\underline{x}$ and $\underline{b}^*$ are column vectors. Since $U^B = U - U^W$, Eq. 20 can be rewritten as $(U - (1 + \lambda_1)U^W)\underline{b} = 0$.

or

$$\left[(U^W)^{-\frac{1}{2}} U (U^W)^{-\frac{1}{2}} - (1 + \lambda_1) I\right] (U^W)^{\frac{1}{2}}\,\underline{b} = 0 \tag{22}$$

setting $(1 + \lambda) = \lambda_1'$ and $(U^W)^{\frac{1}{2}}\underline{b} = \underline{c}$, we obtain

$$\left[(U^W)^{-\frac{1}{2}} U (U^W)^{-\frac{1}{2}} - (\lambda_1')\, I\right]\,\underline{c} = 0 \tag{23}$$

Since by prewhitening we multiply each incoming signal by $(U^W)^{-\frac{1}{2}}$ the internal scatter matrix of the grand sample becomes $(U^W)^{-\frac{1}{2}} U (U^W)^{-\frac{1}{2}}$ if we now apply L-K (or equivalently principle components) analysis and pick the eigenvector corresponding to the largest eigenvalue of the characteristic equation

$$|(U^W)^{-\frac{1}{2}} U (U^W)^{-\frac{1}{2}} - \lambda' I)| = 0 \tag{24}$$

we obtain the optimal line to project on, $\underline{c}^*$, as the solution of the equation

$$\left[(U^W)^{-\frac{1}{2}} U (U^W)^{-\frac{1}{2}} - \lambda_1'\, I\right]\,\underline{c} = 0 \tag{25}$$

and we see that except for (possibly) a scale factor we have

$$\underline{c}^{*} = (U^{w})^{\frac{1}{2}} \underline{b}^{*} \quad . \tag{26}$$

We note that since each of the incoming vectors was multiplied by $(U^{w})^{-\frac{1}{2}}$ the z's (except, possibly, for a scale factor) will become

$$z_{\xi\gamma}^{(\gamma)} = \underline{c}^{*'} \cdot ((U^{w})^{-\frac{1}{2}} \underline{x}_{\xi\gamma}^{(\gamma)}) = b^{*'}(U^{w})^{\frac{1}{2}} \cdot (U^{w})^{-\frac{1}{2}} \underline{x}_{\xi\gamma}^{(\gamma)} = \underline{b}^{*'} \cdot \underline{x}_{\xi\gamma}^{(\gamma)} \tag{27}$$

i.e., the same as Eq. 21.

It is appropriate to point out that if we think of one of the classes as being "noise" and the other as being "signal plus noise" (again using communication engineering terminology) and identify $\sum_{i=1}^{k} b_i x_i$ as the output of a discrete filter, then the ratio Q maximized in the discriminant analysis problem is seen to be "signal-to-noise ratio". Thus, coefficients b_1^{*} define the filter which maximizes the signal-to-noise ratio.

3. The Singularity Problem

For linear classification, discriminant analysis is used to reduce the dimensionability of the sample space to one dimension and a threshold is used to classify the sample points. The optimal direction to project the sample points $\underline{b}^{*}$ is given in terms of the inverse of the within samples scatter matrix. The problem of singularity of this matrix arises when discriminant analysis is applied to a problem with a high dimensional sample space and a small sample size. The internal scatter matrix is nonsingular if and only if the n points in the sample do not lie on a hyperplane of less than k-dimensions, k being the dimensionability of the sample space. If $n_1 + n_2 \leq k + 2$, n_1 and n_2 being the number of samples for group one and group two respectively, the matrix U^{w} is singular, i.e., its inverse does not exist.

In biological classification problems the sample size may be small relative to the dimensionality of the sample space. Since the Loeve-Karhunen method is optimal, in the mean square sense, for reduction of dimensionality, it can be used for tackling small sample problems

in a high dimensional space. The approach would be to reduce the dimensionality of the sample space using the Loeve-Karhunen method to a point where $n_1 + n_2 - 2$ is larger than the dimensionality of the sample space and then apply discriminant analysis.

There are a few other methods which have been suggested for the solution of the small sample problem which results in singularity. One such method is to use the Moore-Penrose Generalized Inverse[24] for inverting the singular within-sample scatter matrix. This method effectively restricts our search for best projection line to a subspace of the sample space, which is orthogonal to the subspace in which there is no scatter. Thus, one can eliminate "good" directions this way and examples can be constructed in which this is just the wrong thing to do. Another suggestion is to replace the zero eigenvalues of U^w by the average eigenvalue thus making the matrix nonsingular[22]. Or more precisely, if Q is the matrix whose columns are the eigenvectors $\underline{\theta}_i$

$$Q = \left[\underline{\theta}_1 \vdots \underline{\theta}_2 \vdots \cdots \vdots \underline{\theta}_n\right]$$

then

$$A^H = Q D_m^{-1} Q'$$

where

$$D_m = \left(\frac{m-2}{n+m-2}\right) D + \left(1 - \frac{(m-2)}{(n+m-2)}\right)\left(\frac{\text{trace } D}{n}\right) (I)$$

$$D = \text{DIAG}(\lambda_1, \cdots \lambda_n)$$

trace D = sum of the elements on the main diagonal of the matrix D

m = total number of samples

n = dimension of space

I = identity matrix

A^H will be referred to as 'H' inverse ('H' standing for T. J. Harley who suggested it).

4. Principal Component Analysis[7]

The Loeve-Karhunen expansion theorem states that a random process in an interval of time Ω may be written as an orthonormal

series with uncorrelated coefficients. The expansion is in terms of the eigenfunctions of the autocovariance function of the process. When the statistics of the process are unknown and a sample autocovariance is taken as an estimate of the autocovariance function of the process, an expansion in terms of the eigenvectors of the sample autocovariance matrix is identical to Principal Component Analysis in multivariate statistics.

Consider a sample of size n from a k-dimensional distribution, $k \leq n$. This sample may be represented geometrically as a sample cluster of n points in k-dimensional Euclidean space R_k. Suppose we wish to project this cluster orthogonally onto an s-dimensional Euclidean space R_s, $s \leq k$, so as to obtain the greatest possible s-dimensional scatter of the projected points (the term scatter will become clear in the solution below). The problem is to determine the direction of projection with respect to the coordinate system of R_k.

The solution of this statistical problem can be stated in the following result due to Hotelling (1933):

Suppose $(x_{1\xi}, \ldots, x_{k\xi}, \ \xi = 1, \ldots, n)$ is a sample of size $n > k$ from a k-dimensional distribution whose covariance matrix is positive definite. Let $\|u_{ij}\|$ be the internal scatter matrix of this sample, i.e.,

$$u_{ij} = \sum_{\xi=1}^{n} x_{i\xi}, \ i = 1, 2, \ldots, k \quad \text{where } \bar{x}_i = \frac{1}{n} \sum_{\xi=1}^{n} x_{i\xi}, \ i = 1, 2, \ldots, k$$

and let it be positive definite with probability 1. Let

$$(c_{1_p}, \ldots, c_{k_p}), \ p = 1, \ldots, s \tag{28}$$

be s k-dimensional unit vectors, that is, such that $\sum_{i=1}^{k} c_{i_p}^2 = 1$, $p = 1, \ldots, s$, and set

$$z_{p\xi} = \sum_{i=1}^{k} c_{ip} x_{i\xi} \quad p = 1, \ldots, s. \tag{29}$$

Let $\|\tilde{u}_{pq}\|$ be the internal scatter matrix of the sample $(z_{1\xi}, \ldots, z_{s\xi}$ $\xi = 1, \ldots, n)$.

The values of the vectors which maximize the scatter $|\tilde{u}_{pq}|$ are the solutions of the s sets of equations

$$\sum_{j=1}^{k} (u_{ij} - \lambda_p \delta_{ij})\, c_{jp} = 0, \quad i = 1, \ldots, k \tag{30}$$

$p = 1, \ldots, s$ where $\lambda_1, \ldots, \lambda_s$, are the s largest roots of the characteristic equation

$$|u_{ij} - \lambda|\delta_{ij} = 0, \tag{31}$$

δ_{ij} being the Kronecker δ, and where $\lambda_1 > \ldots > \lambda_s$ with probability 1. Furthermore, these vectors are orthogonal, and the maximum value of $|\tilde{u}_{pq}|$ is the product $\lambda_1 \lambda_2 \ldots \lambda_s$. The proof of the above statement can be found in Ref. 7.

Both the above analyses are essentially a least mean squares linear fit of a set of S orthonormal waveforms to a set of n observed waveforms. We also note that Principal Component Analysis is sometimes referred to as Principal Factor Analysis.[23]

5. The Choice of Threshold in Linear Classification Problems

In linear classification procedures an observation (signal) $\underline{x}$ is classified as coming from the first population if $\underline{b}' \cdot \underline{x} \leq c$ and as from the second if $\underline{b}' \cdot \underline{x} > c$. The vector $\underline{b}$ and the constant c are chosen to provide maximum discrimination, in some specified sense, between the two classes. In this section the choice of c will be discussed.

Let us first assume that $\underline{b}' \cdot \underline{x}$ has one of two possible distributions $N(\mu_1, \sigma_1^2)$ or $N(\mu_2, \sigma_2^2)$ i.e., normal with mean μ_1 and variance σ_1^2 or normal with mean μ_2 and variance σ_2^2. We lose no generality if we let $\mu_2 > \mu_1$ and designate the population I. Let us denote by L_1 the loss associated with the misclassification of an individual from population I and by L_2 the loss associated with misclassification of an individual from population II, $L_1, L_2 > 0$. Let us further denote the a priori probability of population I by p and of population II by $q = 1-p$. Let P_I be the probability that a random observation (signal) from population I is classified as having arisen from II, and P_{II} the probability that a random individual of II is classified as having arisen from I. We wish to obtain the constant c which minimizes the expected loss, i.e., min $[L_1 p P_I + L_2 q P_{II}]$.

$$P_I = \frac{1}{\sqrt{2\pi}} \int_{\frac{\mu_1 - c}{\sigma_1}}^{} e^{-x^2/2} \, dx \tag{32}$$

$$P_{II} = \frac{1}{\sqrt{2\pi}} \int_{\frac{c - \mu_2}{\sigma_2}}^{} e^{-x^2/2} \, dx \tag{33}$$

Taking the partial derivative with respect to c we obtain

$$\frac{\partial(L_1 p P_I + L_2 q P_{II})}{\partial c} = -\frac{L_1 p}{\sigma_1 \sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{\mu_1 - c}{\sigma_1}\right)^2\right]$$

$$+ \frac{L_2 q}{\sigma_2 \sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{c - \mu_2}{\sigma_2}\right)^2\right] \tag{34}$$

Equating the derivative to zero and rearranging, we obtain

$$2 \ln \frac{L_2 q \sigma_1}{L_1 p \sigma_2} + \left(\frac{\mu_1 - c}{\sigma_1}\right)^2 - \left(\frac{c - \mu_2}{\sigma_2}\right)^2 = 0 \tag{35}$$

which is a quadratic in c with the following roots

$$c = \frac{1}{\sigma_2^2 - \sigma_1^2} \left\{ \sigma_2^2 \mu_1 - \sigma_1^2 \mu_2 \pm \sigma_1 \sigma_2 \left[(\mu_2 - \mu_1)^2 - 2(\sigma_2^2 - \sigma_1^2) \ln \frac{L_1 p \sigma_1}{L_2 q \sigma_2} \right]^{1/2} \right\} \tag{36}$$

There are three possibilities for the roots of equation 35. There are no real roots (Fig. 16) no roots fall in (μ_1, μ_2) interval (Fig. 17) one and only one root falls in (μ_1, μ_2) (Fig. 18). If a root should fall at one of μ_1, μ_2, this may be considered as a limiting case of the situation when no roots fall in (μ_1, μ_2). When there are no real roots the situation is trivial, and all individuals are classified into one population. When no roots fall in (μ_1, μ_2) linear discrimination is not very helpful, and quadratic discrimination is indicated. Let us consider the situation when one and only one root falls in (μ_1, μ_2). When

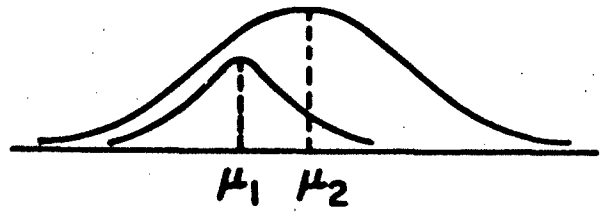


Fig. 16 No Real Roots

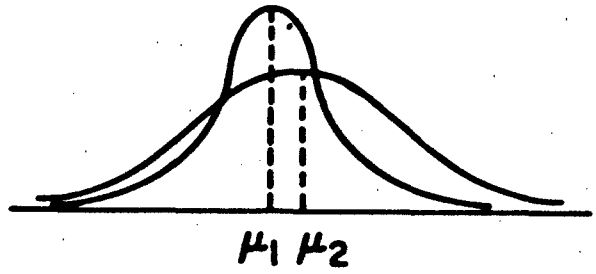


Fig. 17 No Roots in (μ_1, μ_2) Interval

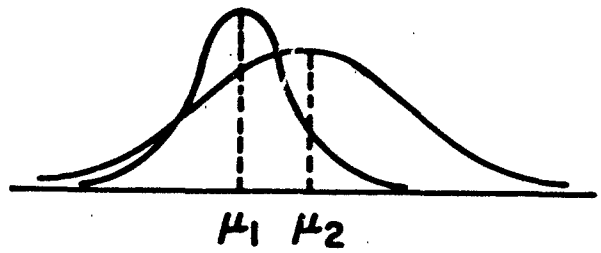


Fig. 18 One Root in (μ_1, μ_2) Interval

a root falls in (μ_1, μ_2), this is the root which minimizes $(L_1 p P_I + L_2 q P_{II})$ and is therefore the root desired. The other root maximizes $(L_1 p P_I + L_2 q P_{II})$ and therefore will not be used. We also notice that when $\sigma_1 < \sigma_2$ the root which falls in (μ_1, μ_2) is the larger of the two and when $\sigma_2 < \sigma_1$ it is the smaller, thus in both cases the positive square root is required.

Let us now consider the case when neither a priori probabilities nor loss functions are known. There can be two criteria for choosing c. One would be the c which minimizes the sum of the two types of errors, and the other which minimize the larger of the error quantities.

Consider now the minimizing of max (P_I, P_{II}). Since P_I and P_{II} are monotonic, decreasing and increasing respectively, in c, the desired c is located such that $P_I = P_{II}$. From Eq. 32 and Eq. 33 we obtain

$$\phi\left(\frac{c - \mu_1}{\sigma_1}\right) = \phi\left(\frac{\mu_2 - c}{\sigma_2}\right) \tag{37}$$

thus,

$$\frac{c - \mu_1}{\sigma_1} = \frac{\mu_2 - c}{\sigma_2} \tag{38}$$

Solving, we obtain

$$c_1 = \frac{\mu_1 \sigma_2 + \mu_2 \sigma_1}{\sigma_2 + \sigma_1} . \tag{39}$$

Setting $L_1 = L_2 = p = q = 1$ in Eq. 36 we obtain (taking the positive square root) the c which minimizes $[P_I + P_{II}]$

$$c_2 = \frac{1}{\sigma_2^2 - \sigma_1^2} \left\{ \sigma_2^2 \mu_1 - \sigma_2^2 \mu_2 + \sigma_1 \sigma_2 \left[(\mu_2 - \mu_1)^2 - 2(\sigma_2^2 - \sigma_1^2) \ln \frac{\sigma_1}{\sigma_2} \right]^{\frac{1}{2}} \right\} \tag{40}$$

It should be noted that if $\sigma_1 = \sigma_2$, the c's under both criteria reduce to a c dependent upon only the centroids

$$c_3 = \frac{\mu_1 + \mu_2}{2} . \tag{41}$$

Under some conditions $\mu_1 + \mu_2/2$ is a good compromise between c_1 and c_3. In Ref (21) Riffenburgh and Clunies-Ross find the conditions under which c_1 is better than c_3 and conditions under which c_2 is better than c_3. In the B, A plane where $B = \sigma_2/\sigma_1$ and $A = \mu_2 - \mu_1/\sigma_2 + \sigma_1$ they find four regions, in region (1) no linear discriminator is reasonable, in region (2) c_3 is a compromise between c_1 and c_2, in region (3) c_2 is better than c_1, and in region (4) both c_2 and c_1 are better than c_3.

The appropriate threshold c is chosen according to the above considerations.

6. Processing of Stochastic Signals

As described in Section VI, a commonly used procedure for analyzing stationary stochastic processes is to first compute the power spectrum or autocorrelation function of each signal in question, and then to use either of these functions as the signal to be analyzed, rather than the original time signal.

A method will not be presented which yields the same results as the above mentioned standard procedure, but which eliminates the need for computing the power spectrum or autocorrelation function of each signal.

Consider the standard procedure, given a stochastic process x(t). First the autocorrelation function $\phi_{xx}(\tau)$ is computed. Assume the maximum correlation span of the process to be T/2. Then an auxiliary function f(t), $0 < t < T$, can be defined such that $f(t) = \phi_{xx}(t - T/2)$. These auxiliary functions f(t) then are used to define a basis of eigenfunctions which are solutions of the equation

$$\int_0^T \phi_{ff}(t - \tau)\,\psi_i(\tau)\,d\tau = \lambda_i \psi_i(t) \qquad (42)$$

The coordinates of any sample in the subspace thus defined are given by

$$c_i = \int_0^T f(t)\,\psi_i(t)\,dt. \qquad (43)$$

The following procedure has been developed so that the coefficients can be determined without computing the autocorrelation $f(t) = \phi_{xx}(t-T/2)$ for each sample function.

First consider that the cross-correlation between input, $x(t)$, and output, $y(t)$, of any linear time invariant system with weighting function $h(t)$ is given by [19]

$$\phi_{xy}(\tau) = \int_{-\infty}^{\infty} \phi_{xx}(\tau - u)\, h(u)\, du$$

$$= \int_{o}^{T} \phi_{xx}(u - \tau)\, h(u)\, du\,, \qquad (44)$$

since ϕ_{xx} is an even function, and it is assumed that

$$h(u) = 0 \begin{cases} u < 0 \\ u > T \end{cases}$$

But, from above, $\phi_{xx}(u - \tau) = f(u - \tau + T/2)$
so that

$$\phi_{xy}(\tau) = \int_{o}^{T} f(u - \tau + T/2)\, h(u)\, du\,,$$

and therefore

$$\phi_{xy}(T/2) = \int_{o}^{T} f(u)\, h(u)\, du\,. \qquad (45)$$

Then, if the weighting function $h(u)$ is taken to be the i^{th} eigenfunction $\psi_i(u)$, the right-hand side of 45 becomes, by 43

$$c_i = \int_{o}^{T} f(u)\, \psi_i(u)\, du$$

The left-hand side of 45 by the definition of the cross-correlation between stationary random functions is

$$\phi_{xy_i}(T/2) = \lim_{\mathcal{T} \to \infty} \frac{1}{2\mathcal{T}} \int_{-\mathcal{T}}^{\mathcal{T}} x(t-T/2)\, y_i(t)\, dt = \overline{x(t-T/2)\, y_i(t)} \qquad (46)$$

Finally, therefore, because of the quality between 43, 45, and 46, the coefficient c_i can be evaluated as shown in Fig. 19 by averaging (for a "suitable length" of time, $\mathcal{T}$) the product of the output and delayed input of a filter whose weighting function is the i^{th} eigenfunction.

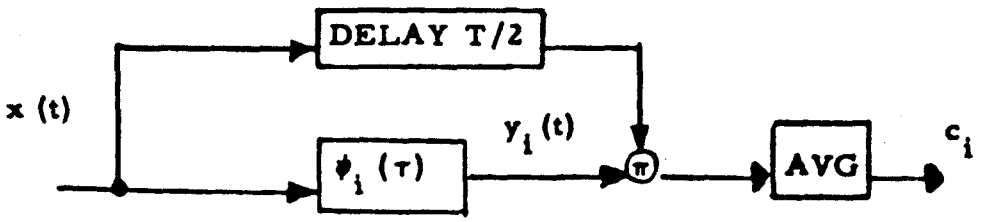


Figure 19 - Time Domain Processing

An experiment was conducted to provide a direct comparison of the standard procedure and the proposed time domain method. Computer-generated random signals were used as the data set, with 20 samples of each class used for training and 20 for test. The results on test data are shown in Fig. 20. The results show the time domain method to be significantly more effective in separating the test samples (figures of merit = 6.25 vs. 4.5 for the standard procedure.)

7. Another Proof for Theorem 4.1, Ref. 1

Theorem 4.1

Two probability distributions $\{\rho_i\}$ and $\{\lambda_i\}$ satisfying, $\rho_i \geq 0$, $\lambda_j \geq 0$, $\sum_{i=1}^{\infty} \rho_i = 1$, $\sum_{j=1}^{\infty} \lambda_j = 1$ are connected by a double stochastic matrix A_{ij} such that $\rho_i = \sum_{j=1}^{\infty} A_{ij} \lambda_j$. If the labeling of the λ_j is done in a descending order $\lambda_1 \geq \lambda_2 \geq \ldots$ then, for an arbitrary n, the sum of the first n elements of $\{\lambda_j\}$ is not less than the sum of the first n elements of $\{\rho_j\}$

Proof:*

$$\sum_{i=1}^{\infty} \rho_i = \sum_{i=1}^{n} \sum_{j=1}^{\infty} A_{ij} \lambda_j = \sum_{i=1}^{n} \sum_{j=1}^{n} A_{ij} \lambda_j + \sum_{i=1}^{n} \sum_{j=n+1}^{\infty} A_{ij} \lambda_j \qquad (47)$$

* This proof was suggested by S. Winograd.

Since $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n \geq \lambda_{n+1} \geq \ldots$

$$\sum_{p=1}^{n} a_p \lambda_p \geq \lambda_{n+1} \quad \text{for any } a_p \geq 0 \text{ and } \sum_{p=1}^{\infty} a_p = 1 \tag{48}$$

$$\text{Let } a_p = \frac{\sum_{i=n+1}^{\infty} A_{ip}}{\sum_{p=1}^{n} \sum_{i=n+1}^{\infty} A_{ip}} \tag{49}$$

note that since A is doubly stochastic

$$\sum_{p=1}^{\infty} \sum_{i=n+1}^{\infty} A_{ip} = \sum_{p=1}^{n} \sum_{i=n+1}^{\infty} A_{pi} \tag{50}$$

$$\therefore \quad a_p = \frac{\sum_{i=n+1}^{\infty} A_{ip}}{\sum_{p=1}^{n} \sum_{i=n+1}^{\infty} A_{pi}} \tag{51}$$

by (47) and (48).

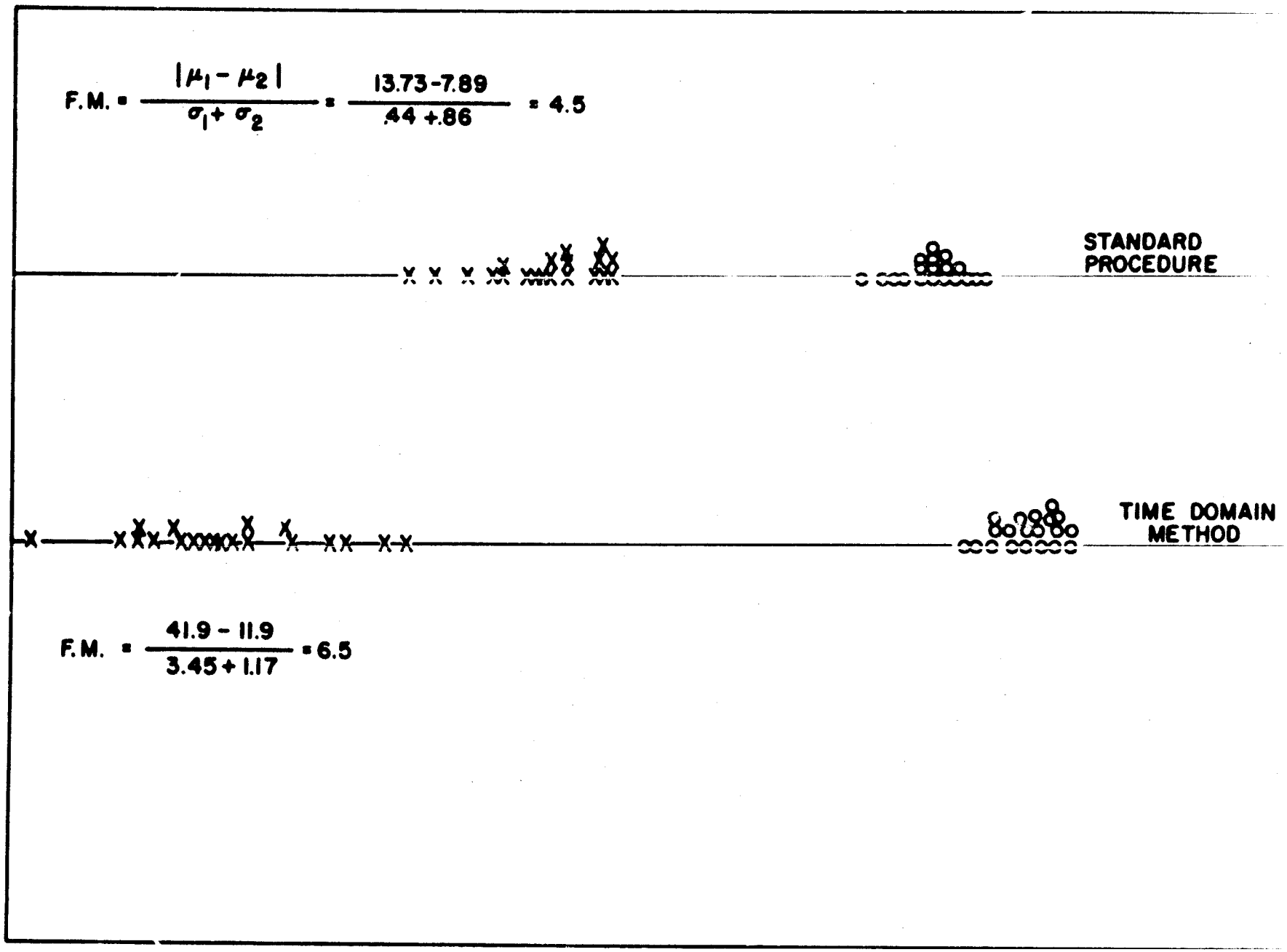
$$\sum_{i=1}^{\infty} \rho_i \leq \sum_{j=1}^{n} \sum_{i=1}^{n} A_{ij} \lambda_j + \sum_{i=1}^{n} \sum_{j=n+1}^{\infty} A_{ij} \sum_{p=1}^{n} a_p \lambda_p = \tag{52}$$

Substituting (51) we get

$$= \sum_{j=1}^{n} \lambda_j \sum_{i=1}^{n} A_{ij} + \sum_{p=1}^{n} \sum_{i=n+1}^{\infty} A_{ip} \lambda_p = \sum_{p=1}^{n} \lambda_p \left[\sum_{i=1}^{n} A_{ip} + \sum_{i=n+1}^{\infty} A_{ip} \right]$$

$$= \sum_{p=1}^{n} \lambda_p$$

$$\therefore \quad \sum_{i=1}^{n} \rho_i \leq \sum_{p=1}^{n} \lambda_p \qquad \text{Q. E. D.}$$


F.M. = $\frac{|\mu_1 - \mu_2|}{\sigma_1 + \sigma_2} = \frac{13.73-7.89}{.44+.86} = 4.5$
STANDARD PROCEDURE
TIME DOMAIN METHOD
F.M. = $\frac{41.9 - 11.9}{3.45 + 1.17} = 6.5$

SECTION IX

CONCLUSIONS AND SUGGESTIONS FOR FUTURE WORK

The L-K and Discriminant Analysis methods studied provide a promising basic methodology for computer-assisted analysis and classification of biological data. When properly used, the computer programs which have been developed should become useful tools in the hands of research workers in the biological sciences. It would, however, be prudent to consider them still as experimental tools until they have been exercised and evaluated on much more data.

It is suggested that future research be devoted to the considerations of methods of analysis based on other than the mean-square criterion [16,17], and on extensions to multiple time signals[18].

REFERENCES

1. S. Watanabe, The Loève-Karhuen Expansion as a Means of Information Compression for Classification of Continuous Signals, AMRL-TR-65-114, Aerospace Medical Research Laboratories, Wright-Patterson Air Force Base, Ohio, October 1965.

2. R. Courant and D. Hilbert, "Methods of Mathematical Physics," I, Interscience Publishers, New York, 1953, Ch. III.

3. K. Karhunen, "Über Linear Methodden in der Wahrscheinlichkeitscrechnung," Ann. Ac. Sci. Fennicae, AI 37, Helsinki (1947).

4. M. Loeve, Probability Theory, D. Van Nostrand, Princeton, New Jersey, 1955.

5. C. W. Helstrom, Statistical Theory of Signal Detection, Permagon Press, The MacMillian Co., New York, 1960, Chs. 4, 11.

6. K. L. Jordan, Jr., Discrete Representations of Random Signals, Technical Report 378, Research Laboratory of Electronics, MIT, July 14, 1961.

7. S. S. Wilks, Mathematical Statistics, John Wiley and Sons, Inc. New York & London, 1960, Ch. 18.

8. W. J. Dixon, ed. Biomedical Computer Programs, Health, Sciences Computing Facility, Dept. of Preventative Medicine and Public Health, School of Medicine, UCLA.

9. E. R. John, D. S. Ruchkin and J. Villegas, "Experimental Background: Signal Analysis and Behavioral Correlates of Evoked Potential Configur tions in Cats," Annals New York Academy of Sciences, 1963, pp. 363-420.

10. D. S. Ruchkin, J. Villegas and E. R. John, "An Analysis of Average Evoked Potentials Making Use of Least Mean Square Techniques," Annals New York Academy of Sciences, 1964,

11. D. O. Walter and W. R. Adey, "Analysis of Brain-Wave Generators as Multiple Statistical Time Series," Trans. IEEE, Prof. Group on Biomedical Electronics, Apr. 1965.

12. R. B. Blackman and J. W. Tukey, The Measurement of Power Spectra, Dover, 1958.

13. P. D. Welch, "A Direct Digital Method of Power Spectrum Estimation," IBM Journal of Research and Development, April 1961, pp. 141-156.

14. R. L. Mattson and J. E. Dammann, "A Technique for Determining and Coding Subclasses in Pattern Recognition Problems," IBM Journal of Research and Development, Vol. 9, No. 4, July 1965.

15. C. R. Rao, Advanced Statistical Methods in Biometric Research, John Wiley & Sons, New York, 1952, pp. 365-370.

16. J. R. Rice, "The Approximation of Functions", Addison-Wesley, 1964, Vol. I.

17. J. W. Tukey, "Data Analysis and the Frontiers of Geophysics," Science, June 4, 1965, pp. 1283-1289.

18. E. Wong, Vector Stochastic Processes in Problems of Communication Theory, PhD. Thesis Princeton U., 1959.

19. Y. W. Lee, Statistical Theory of Communication, John Wiley & Sons, New York, 1960, Ch. 13.

20. R. H. Riffenburgh and C. W. Clunies-Ross, "Linear Discriminant Analysis," Pacific Science, Vol. XIV, July, 1960.

21. R. A. Fisher, "The Statistical Utilization of Multiple Measurements," Ann. Eugen. 8, 376-386 (1938).

22. Semi-Automatic Imagery Screening Research Study and Experimental Investigation, Report No. 5 (AD No. 427172), Army Electronic Research and Development Laboratory, Fort Monmouth, New Jersey.

23. H. H. Harman, Modern Factor Analysis, 1960, The University of Chicago Press.

24. R. A. Penrose, "Generalized Universe for Matrices," Proc. Cambridge Phil. Soc. 51, 406-413 (1955).

25. P. Welch and R. Wimpress, "Two Multivariate Statistical Computer Programs and their Application to the Vowel Recognition Problem," J. Acoustical Soc. of America, April, 1961..

26. R. G. Casey, "Linear Reduction of Dimensionality in Pattern Recognition," IBM Report RC-1431.

27. J. Raviv and D. N. Streeter, "Linear Methods for Biological Data Processing," IBM Report RC-1513.

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)	2a. REPORT SECURITY CLASSIFICATION
Thomas J. Watson Research Center, International Business Machines Corp., PO Box 218, Yorktown Heights, New York 10598	UNCLASSIFIED
	2b. GROUP N/A

3. REPORT TITLE

RESEARCH ON ADVANCED COMPUTER METHODS FOR BIOLOGICAL DATA PROCESSING

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

Final report, 15 July 1964 - 14 July 1965

5. AUTHOR(S) (Last name, first name, initial)

Streeter, D. N., PhD
Raviv, J., PhD

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
April 1966	58	27

8a. CONTRACT OR GRANT NO. AF 33(615)-2047	9a. ORIGINATOR'S REPORT NUMBER(S)
b. PROJECT NO. 7233	IBM Research Report No. RC 1513
c. Task No. 723305	9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report) AMRL-TR-66-24
d.	

10. AVAILABILITY/LIMITATION NOTICES

Distribution of this document is unlimited.

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

13. ABSTRACT

The purpose of the research carried out under this contract has been the development of mathematical methods and computer programs for the extraction of meaningful information from biological, primarily neurophysiological, measurements. Emphasis has been placed on statistical methods suitable for separating two or more random signals and which provide insight into the underlying mechanism by which the signals are generated. Loeve-Karhunen expansion and Discriminant Analysis methods are applied to the problem of time signal classification. Experiments are performed both on computer generated time signals and on electroencephalograms. Methods of coping with the singularity problem arising from a small sample size are investigated.

DD FORM 1 JAN 64 1473

Security Classification

14. KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Classification Signal analysis Biomedical signals EEG analysis Loeve-Karhunen methods Discriminant analysis Computer signal processing						

INSTRUCTIONS

1. ORIGINATING ACTIVITY: Enter the name and address of the contractor, subcontractor, grantee, Department of Defense activity or other organization *(corporate author)* issuing the report.

2a. REPORT SECURITY CLASSIFICATION: Enter the overall security classification of the report. Indicate whether "Restricted Data" is included. Marking is to be in accordance with appropriate security regulations.

2b. GROUP: Automatic downgrading is specified in DoD Directive 5200.10 and Armed Forces Industrial Manual. Enter the group number. Also, when applicable, show that optional markings have been used for Group 3 and Group 4 as authorized.

3. REPORT TITLE: Enter the complete report title in all capital letters. Titles in all cases should be unclassified. If a meaningful title cannot be selected without classification, show title classification in all capitals in parenthesis immediately following the title.

4. DESCRIPTIVE NOTES: If appropriate, enter the type of report, e.g., interim, progress, summary, annual, or final. Give the inclusive dates when a specific reporting period is covered.

5. AUTHOR(S): Enter the name(s) of author(s) as shown on or in the report. Enter last name, first name, middle initial. If military, show rank and branch of service. The name of the principal author is an absolute minimum requirement.

6. REPORT DATE: Enter the date of the report as day, month, year; or month, year. If more than one date appears on the report, use date of publication.

7a. TOTAL NUMBER OF PAGES: The total page count should follow normal pagination procedures, i.e., enter the number of pages containing information.

7b. NUMBER OF REFERENCES: Enter the total number of references cited in the report.

8a. CONTRACT OR GRANT NUMBER: If appropriate, enter the applicable number of the contract or grant under which the report was written.

8b, 8c, & 8d. PROJECT NUMBER: Enter the appropriate military department identification, such as project number, subproject number, system numbers, task number, etc.

9a. ORIGINATOR'S REPORT NUMBER(S): Enter the official report number by which the document will be identified and controlled by the originating activity. This number must be unique to this report.

9b. OTHER REPORT NUMBER(S): If the report has been assigned any other report numbers *(either by the originator or by the sponsor)*, also enter this number(s).

10. AVAILABILITY/LIMITATION NOTICES: Enter any limitations on further dissemination of the report, other than those imposed by security classification, using standard statements such as:

(1) "Qualified requesters may obtain copies of this report from DDC."

(2) "Foreign announcement and dissemination of this report by DDC is not authorized."

(3) "U. S. Government agencies may obtain copies of this report directly from DDC. Other qualified DDC users shall request through _____."

(4) "U. S. military agencies may obtain copies of this report directly from DDC. Other qualified users shall request through _____."

(5) "All distribution of this report is controlled. Qualified DDC users shall request through _____."

If the report has been furnished to the Office of Technical Services, Department of Commerce, for sale to the public, indicate this fact and enter the price, if known.

11. SUPPLEMENTARY NOTES: Use for additional explanatory notes.

12. SPONSORING MILITARY ACTIVITY: Enter the name of the departmental project office or laboratory sponsoring *(paying for)* the research and development. Include address.

13. ABSTRACT: Enter an abstract giving a brief and factual summary of the document indicative of the report, even though it may also appear elsewhere in the body of the technical report. If additional space is required, a continuation sheet shall be attached.

It is highly desirable that the abstract of classified reports be unclassified. Each paragraph of the abstract shall end with an indication of the military security classification of the information in the paragraph, represented as *(TS), (S), (C), or (U)*.

There is no limitation on the length of the abstract. However, the suggested length is from 150 to 225 words.

14. KEY WORDS: Key words are technically meaningful terms or short phrases that characterize a report and may be used as index entries for cataloging the report. Key words must be selected so that no security classification is required. Identifiers, such as equipment model designation, trade name, military project code name, geographic location, may be used as key words but will be followed by an indication of technical context. The assignment of links, rules, and weights is optional.