FEATURE SELECTION FOR DIAGNOSIS OF VECTORCARDIOGRAMS*

D. E. Gustafson
Charles Stark Draper Laboratory
Cambridge, Mass. 02142

A. Akant and S. K. Mitter
Massachusetts Institute of Technology
Cambridge, Mass. 02139

ABSTRACT

The automatic classification of vectorcardiograms and electrocardiograms into disease classes using computerized pattern recognition techniques has been a much studied problem. To date, however, no system exists which meets desired accuracy and noise immunity requirements and development of new techniques continues. An important aspect of the problem is that of feature selection, in which the functions of data reduction and information preservation are performed. In this paper, the problem of linear feature extraction is studied and a modified form of the Karhunen-Loeve expansion is developed which appears to have some advantages for the present application. Comparison with other feature selection methods is made using a two-dimensional example. Finally, some areas for future research are pointed out.

I. INTRODUCTION

An important application of pattern recognition theory in biomedical engineering applications is in automated diagnosis of electrocardiograms (ECG) or vectorcardiograms (VCG). This is demonstrated both in the wealth of literature on the subject and the development of several readily available programs [1 - 5]. Nevertheless, even after two decades or so of work, problems still remain in achieving desired classification accuracy.

Bailey, et al [6], have tested three of the most popular U.S. programs using two digital representations of the same analog ECG tracing, separated by one millisecond in time. Although the computer diagnoses should be identical (i.e., cardiologist diagnosis is unchanged), diagnostic statements were identically reproduced less than 80% of the time. Reproducibility was most affected by the algorithms used for feature extraction, pattern recognition, consistency checks and noise handling. Reproducibility was less affected by analog high frequency noise attenuation. As Bailey notes, poor reproducibility performance obviates the need for more time-consuming clinical evaluation. Since such performance does not eliminate the need for human intervention and review, much work remains to be done to develop a truly robust diagnostic program.

The problem of ECG or VCG diagnosis is generally divided into two essentially independent parts; (1) rhythm analysis, (2) waveform morphology, or contour, analysis. Rhythm analysis is principally concerned with the determination of the sites and rates of cardiac pacemakers and impulse propagation through the cardiac conduction system. Waveform morphology analysis is principally concerned with persistent patterns of wave shapes in an attempt to describe the state of the working muscle masses. In this paper, we will be concerned only with the problem of morphology analysis.

The pattern recognition problem is generally divided into two sequential steps; (1) feature extraction, (2) classification. This dichotomy is usually made for simplicity, since selection of features in most practical problems depends on the structure of the classifier and on the training data. Here we consider only the problem of feature extraction. In particular, we consider only linear feature extraction and comment on several aspects of this very important problem, motivated by our experience while working with cardiologists in designing a feature extractor for VCGs.

This paper is organized as follows. The problem of representing the signal and noise processes and filtering of the noise processes is considered in Section II. The feature extraction problem is discussed in Section III, with emphasis on linear feature extraction. A modified Karhunen-Loève expansion technique is proposed in Section IV which may have some potential advantages in automated medical diagnosis. Comparisons of several proposed linear feature extraction methods are given in Section V, using a simple two-dimensional example.

*This work was supported by the U.S. Air Force School of Aerospace Medicine under Contract F41609-75-C-0019

Finally, a discussion of the overall pattern recognition problem within the context of diagnosis of VCGs is given in Section VI. Several areas of research, which are suggested by the practical problem at hand, are outlined.

II. REPRESENTATION OF SIGNAL AND NOISE PROCESSES

The measured analog waveforms (ECG or VCG) at time t are denoted by the vector y(t) and assumed to consist of the underlying cardiographic signal x(t) plus an unwanted noise term n(t); $y(t) = x(t) + n(t)$.

2.1 Signal Process

There are several ways of modeling the signal process x(t) over a finite interval $[0, T]$. The signal may be represented as the output of a lumped-parameter, time-varying linear or nonlinear dynamical system. There has been very little success in this area, since the signal is stochastic and time-varying and not accurately represented using low order linear models. One is faced, inevitably, with a difficult identification problem.

An alternate approach is taken here and consists of representing the signal process as a sample from an ensemble of statistically nonstationary waveforms. Since computation will be done in discrete time we consider in the sequel only the discrete time representation of the signal. Let the j^{th} component of the signal at time t_i be denoted by $x_j(i)$, with the interval $[0, T]$ containing n sampling times. For a signal of dimension m we define the mn vector:

$$x = [x_1(1), \cdots, x_1(n), x_2(1), \cdots, x_m(n)]^T \quad (1)$$

Then the signal vector x is represented in terms of a set of scalar parameters $\alpha = \{\alpha_j; j = 1, \cdots, N\}$ as:

$$x = \bar{x} + \sum_{j=1}^{N} \alpha_j \phi_j \quad (2)$$

where the set of N vectors

$$\phi = \{\phi_j ; j = 1, \cdots, N\} \text{ span } R^N,$$

$N = mn$, and $\bar{x}$ is the ensemble mean of x. The coefficient set α, computed on-line from the data x and the basis vector set ϕ, are the features to be used in the classification algorithm.

In any practical scheme, N will be of the order of several hundred. Thus the desire is to pick out only the most informative features. The representation (2) then becomes, for m features:

$$x = \bar{x} + \sum_{j=1}^{m} \alpha_j \phi_j + e_m \quad (3)$$

where e_m is the representation error. The desire is to minimize some measure of e_m over the ensemble by selecting an appropriate rule for choosing the features, and by properly selecting the basis vectors. Note that, ideally, the basis vector selection should depend on the discrimination rule and the data used for training. Thus, feature selection should be done by employing learning rather than on an apriori basis.

Selection of the basis vectors and the rule for selecting the coefficients is given in Section IV. We now turn to a discussion of the noise processes and methods used to remove them from the signal process.

2.2 Noise Processes

In the algorithms under development, the low-frequency baseline shifts are first eliminated. Then an averaged heartbeat is computed by averaging together several successive heartbeats. The individual beats are lined up at their fiducial points defined as the point of maximum downward slope of the QRS complex.

Inspection of analog VCG data reveals four major sources of noise; (1) powerline ripple (60 Hz), (2) baseline drift, (3) high frequency noise, (4) artifacts. For the techniques proposed here, it appears that baseline drifts are the most significant noise source. Any 60 Hz components are essentially orthogonal to the feature space, as is high-frequency noise. That is, the elements of α are all essentially zero. Artifacts (e.g., muscle noise) appear to either be easily detectable or average out with time.

An example of a particularly severe baseline encountered in real data is given in Figure 1a. In Figure 1b, an estimate of the underlying signal (original minus estimated baseline), using a first-order Kalman filter with optimized filter time constant and gain, is shown. Note the induced overshoot indicative of an S wave and the ST segment slope error in the estimated waveform.

The overshoot was eliminated by using an adaptive filter in which the gain was decreased with increasing measurement residual amplitude. However, this resulted in sluggish behavior, as shown in Figure 1c. In an effort to improve response a second-order non-adaptive filter was tried. This did improve performance, as seen in Figure 1d. However, the causal filter resulted in significant phase shifts for a waveform with a respiratory component (Figures 2a and 2b). Since slope monitoring is desirable to eliminate regions of high baseline slope, the results of Figure 2b were judged inadequate.

A non-causal, symmetric, moving-window filter with zero phase shift characteristic was next tried and gave the best overall performance (Figure 3). Window length was 800 msec, using equal weighting for both baseline and slope

estimates. Computation time was decreased by using only every 20th sample (at 4 msec/sample). No aliasing problems were encountered since higher harmonics were outside the bandwidth of the moving-average filter.

An interesting problem was encountered using the non-adaptive moving-window filter. As shown in Figure 4, for certain heart rates a fictitious T wave component was introduced due to the energy in the QRS complex. This problem was eliminated by making the filter adaptive. Incoming data for which a three-point slope estimate exceeded 10 mv/sec were eliminated and an extrapolated zero-slope estimate was substituted. The next data point was also neglected to allow for broad QRS complexes (e.g., due to left bundle branch block). This adaptive, moving-window filter has performed adequately in all tests made to date.

In the sequel we assume that the noise has been eliminated from the measured signal or will not affect the feature extraction process. As discussed at the beginning of this subsection, this appears to be an achievable condition.

III. FEATURE EXTRACTION

Perhaps the most important problem in VCG or ECG computer diagnosis is feature extraction, in which a small set of numbers is selected to represent the complex waveforms of each patient for later pattern recognition. These numbers (features) should be selected to provide: (1) optimal discrimination, (2) minimum sensitivity to noise, artifact, heart size and orientation, body shape, etc.

The essential function of the feature extractor is data reduction, in which only diagnostically relevant measurements are preserved. In most practical problems, the feature set is to be selected to give the lowest probability of misclassification. However, since the data are not independent this is no simple matter. For example, suppose our data consisted of N samples and we wish to use only M(< N) of these as features. We cannot, in general, pick these as the M singly most informative features. Cover [7] has given an example where the two best independent measurements are not the two best. The generalization of this result is that an exhaustive search may be required over all M-element subsets of the N data points.

Essentially all programs in present-day use select features based on the cardiologist's knowledge and experience. Among typical features used are location of onset and end of the QRS complex, P wave, T wave, and ST segment depression. What may be neglected, however, are potentially important features relating to subtle changes in waveform shapes and correlations between waveform segments, locations and durations.

Diagnostic statements are generally made on the basis of threshold logic, using a hierarchical decision tree structure. In so doing, the statistical nature of the problem may be overlooked and the decision structure, determined apriori, restricts the flexibility of the system. While the cardiologist is an expert pattern recognizer, the crucial question here is "Does the cardiologist know what features he uses?" in the sense of being able to quantify them. The answer to this question based on present operational systems, appears to be in the negative.

With this in mind, it would appear desirable to develop an approach in which features are selected on the basis of efficient classification into disease categories and insensitivity to noise. This is the essence of the approach given in this paper.

A statistical feature selection procedure is presented here for use in ECG/VCG diagnosis. Risk of misclassification and time weighting are included to enhance performance. The technique is, in principle, easily extendable to more general medical diagnosis problems.

Although most previous work in this area is based on deterministic ideas, recently there has been more effort put into statistical approaches. Cornfield, et al [2], have used a Bayesian approach to classification, using a feature set composed principally of waveform onset and end times. Balm [8] tried a correlation technique, using as features 36 evenly-spaced samples over each QT interval. Muciardi and Gose [9] compared several statistical techniques for selecting optimal features from a set of 157 features chosen by cardiologists. Okajima, et al [10], considered an adaptive, matched filter approach using, as features, 30 evenly-spaced samples over each QRS complex.

Recently, some work has been reported using Karhunen-Loève expansions over portions of the PT segment. Karlsson [11] used 20 QRS samples and 27 ST interval samples and concluded that nine features were sufficient for diagnosis. Hambley, et al [12] used 120 evenly-spaced samples over the QRST interval and concluded that eight features were sufficient. Kittler and Young [13] used 20 evenly-spaced QRS samples from each axis of three-lead vectorcardiograms. Their data indicate that at least 30 features were required to achieve minimum classification error probability.

In work [14, 15] performed at the Charles Stark Draper Laboratory, the use of a K-L expansion over the entire PT interval has been studied, using an 8 msec sampling rate. A total of 300 samples (100/axis) were used for each record (N= 300). The dominant eigenvalues were found using an efficient method for symmetric matrices given in Wilkinson and Reinsch [16], which find all eigenvalues which lie within a given range. The data consisted of 198 VCG records, supplied by cardiologists at the U.S. Air Force School of Aerospace Medicine (USAF/SAM), with approximately

equal numbers of normal and abnormal records. The mean and several of the eigenvectors are shown in Figures 5 - 8, with the i^{th} eigenvector associated with the i^{th} largest eigenvalue of the ensemble covariance matrix. The data are presented in an orthogonal (u, v, w) frame determined for each record so that the v axis contains the least amount of the time-averaged energy. The u axis is aligned with the maximum T wave amplitude in the u-w plane and the w axis completes an orthogonal triad. The u-w plane contains about 90% of the total energy, indicating that the cardiac dipole vector is essentially confined to the u-w plane. Preliminary tests [15] indicate that improved classification accuracy can be obtained using the (u, v, w) axes rather than the (x, y, z) axes of the Frank lead system, since sensitivity to heart orientation is reduced.

Several points should be noted from the data of Figures 5-8. The eigenvectors clearly represent processes which are quite non-stationary and time-varying. For wide-sense stationary processes, the eigenvectors are sinusoidal. The higher order eigenvectors contain higher frequency components of the QRS complex. It was found, however, that no significant P wave components appeared in the first twelve eigenvectors. This has led to problems in reconstructing the low energy P waves, which have high diagnostic content. It thus appears desirable to modify the K-L expansion, perhaps using time weighting of the data, to enhance feature extraction. The percentage of the total ensemble energy contained in the first few eigenvectors is shown in Figure 9. Six eigenvectors contain 90% and 12 eigenvectors contain 98%. However, it appears that part of the remaining 2% may be of significance.

The objective of the remainder of this paper is to suggest some generalizations of the K-L expansion approach to feature extraction, motivated by problems encountered during research on diagnosis of VCGs.

IV. A MODIFIED K-L EXPANSION

4.1 Formulation

For the linear signal model of (3), a popular approach to feature extraction is to utilize the K-L expansion. To facilitate notation, let superscript k denote association with the k^{th} member of the ensemble. Then (3) becomes

$$x^k = \bar{x} + \sum_{j=1}^{m} \alpha_j^k \phi_j + e_m^k \qquad (4)$$

with the first two terms on the right-hand side representing the estimate of x^k. The K-L expansion is defined by first constraining ϕ to be an orthonormal set. Then, the coefficients are selected by:

$$\alpha_j^k = \phi_j^T x^k \qquad (5)$$

and the basis vectors $\{\phi_j\}$ are eigenvectors of the ensemble covariance matrix

$$R = \frac{1}{k-1} \sum_{k=1}^{P} (x^k - \bar{x})(x^k - \bar{x})^T \qquad (6)$$

with P the number of samples in the population. The eigenvectors are ordered according to the magnitude of their corresponding eigenvalues, with ϕ_j associated with the j^{th} largest eigenvalue. The K-L expansion has been extensively studied [17, 18] and its properties are well-known. The intent here is to point out some deficiencies in the K-L expansion, as applied to feature extraction for ECG/VCG diagnosis in particular and to medical diagnosis in general.

The K-L expansion minimizes the cost function

$$J_m = \sum_{k=1}^{P} (e_m^k)^T (e_m^k) \qquad (7)$$

Note that each member of the ensemble receives equal weighting and that all time points are also given equal weight. Intuitively, one would expect that ensemble members with high misclassification risk should be weighted more heavily and that portions of the waveform with higher information content should receive higher weighting so that ϕ is representative of the diagnostically important data. Consider a modification of (7) to the form

$$J_m = \sum_{k=1}^{P} r^k (e^k)^T Q (e^k) \qquad (8)$$

with $r^k > 0$, $Q > 0$. Then r^k is the cost, or risk, associated with misclassifying x^k and Q allows the errors to be time weighted.

Chien and Fu [19] have considered the case of ℓ independent classes. Their result is that the weighting coefficients take the form $r^k = p_i/n_i$, where member x^k is in the i^{th} class Γ_i, which contain n_i members and occurs with apriori probability p_i. Note, however, that if p_i is very small, then ϕ will not contain vectors characteristic of the i^{th} class and discrimination of the i^{th} class will be difficult. Thus, r^k should reflect the risk of misclassification of x^k, as well as the probability of occurrence of x^k.

4.2 Solution

The solution of the modified problem defined by (4) and (8) is given as follows. Without loss of generality, we take the set $\phi = \{\phi_i\}$ to be Q-conjugate;

$$\phi_j^T Q \phi_\ell = \delta_{j\ell}; \; j = 1, \cdots, N; \; \ell = 1, \cdots, N \qquad (9)$$

Then minimizing (8) with respect to α_j^k yields

$$\alpha_j^k = \phi_j^T Q x^k \tag{10}$$

If minimization is now done over the set $\{\phi_i\}$, the result is that ϕ_j is the solution of the eigenequation

$$\tilde{R} Q \phi_j = \lambda_j \phi_j \tag{11}$$

where

$$\tilde{R} = \sum_{k=1}^{P} r^k (x^k - \bar{x})(x^k - \bar{x})^T \tag{12}$$

and the eigenvalues are ordered as $\lambda_1 > \lambda_2 > \lambda_3 > \cdots > \lambda_N$.

The effect of Q is a non-singular transformation of coordinates. If Q is factored as $Q = WW^T$ and we define $z^k = W^T x^k$, $\psi_j = W^T \phi_j$, then (10) and (11) become:

$$\alpha_j^k = \psi_j^T z^k, \quad \tilde{R}_z \psi_j = \lambda_j \psi_j \tag{13}$$

where $\tilde{R}_z = W^T \tilde{R} W$ is the covariance matrix associated with z.

The question now to be asked is how to choose the set $\{r^k\}$ and Q. Ideally, these should be selected to minimize classification error. One method for doing this would be an iterative scheme in which the global minimum of, say, the Bayes risk would be sought by varying $\{r^k\}$ and Q. This is a complex nonlinear programming problem which is further complicated by the fact that analytic gradients are not generally available. As an alternative, we will consider in the sequel a two-class problem in which the discriminatory information is easily defined. Diagnostically, this may be thought of as a screening problem with classes "Normal" and "Questionable".

Some insight into the two-class problem may be gained by considering the relationship of the modified K-L expansion and classical linear discriminant analysis for the two-class problem.

4.3 Relation to Discriminant Analysis

In statistical discriminant analysis [17], within-class and between-class scatter matrices are used to develop criteria for class discrimination. The within-class scatter matrix is given by:

$$S_w = c_N R_N + c_Q R_Q; \; c_N + c_Q = 1 \tag{14}$$

and the between-class scatter matrix is:

$$S_b = S_w + c_N c_Q (m_N - m_Q)(m_N - m_Q)^T \triangleq S_w + M$$

where m_N and R_N are the mean and covariance matrix for the Normal class and m_Q and R_Q are the mean and covariance matrix for the Questionable class. Note that S_b is the ensemble scatter matrix.

The usual interpretation given to c_N and c_Q is that they are the respective apriori class probabilities; $c_N = p_N$, $c_Q = p_Q$. However, since usually $p_N >> p_Q$, S_w carries very little information relative to the questionable class. By including class risk factors r_N and r_Q and and defining $c_N = r_N p_N$, $c_Q = r_Q p_Q$, S_w will contain more information relative to the questionable data since the risk r_Q, associated with misclassifying questionable data, is much higher than the risk r_N, associated with misclassifying normal data.

Efficient discrimination is enhanced by simultaneously maximizing S_b and minimizing S_w. A convenient scalar discrimination measure is [17]:

$$J = \mathrm{tr}(S_w^{-1} S_b) \tag{16}$$

which is invariant under any nonsingular transformation of coordinates, which is a desirable feature.

We now consider the problem of finding an optimal set of m features. Define the m x n matrix A as:

$$A = \begin{bmatrix} \psi_1^T \\ \psi_2^T \\ \vdots \\ \psi_m^T \end{bmatrix} \tag{17}$$

with $\{\psi_i\}$ an orthonormal set. Then the feature vector for x^k is;

$$\alpha^k = [\alpha_1^k, \ldots \alpha_m^k]^T = A x^k \tag{18}$$

and the measure to be maximized is:

$$J_m = \mathrm{tr}[(AS_w A^T)^{-1} (AS_b A^T)] \tag{19}$$

It can be shown [17] that J_m is maximized if ϕ_i is the eigenvector of $S_w^{-1} S_b$ corresponding to the eigenvalue μ_i, with the eigenvalues ordered as $\mu_1 > \mu_2 > \ldots > \mu_m$. Further,

$$J_m = \sum_{i=1}^{m} \mu_i \tag{20}$$

The solution, then, requires solving an eigenvalue problem of the form;

$$S_w^{-1} S_b \phi_i = \mu_i \phi_i \tag{21}$$

and the eigenvalues μ_i are positive real since S_w and S_b are symmetric and positive-definite. We remark that the method proposed by Kittler and Young [13] is identical to solving (21) although derived in a different manner.

By equating (21) and (11), it can be seen that the time weighting matrix Q is explicitly given by

$$Q = \tilde{R}^{-1} S_w^{-1} S_b \tag{22}$$

Since only the symmetric part of Q affects J_m of (8), (22) can be reduced to

$$Q = \tilde{R}^{-1} + \frac{1}{2}(\tilde{R}^{-1} S_w^{-1} M + M S_w^{-1} \tilde{R}^{-1}) \tag{23}$$

The relation (23) provides a way of determining the equivalent weighting in the modified K-L expansion to give the discriminant analysis solution. The value of this is in increasing understanding of the linear discriminant method in terms of the intuitive idea of time weighting of the data. Note that (23) implies that, in general, cross-weighting of the original data is required. However, if Q of (23) is diagonalized by an orthogonal transformation matrix C so that CQC^T is diagonal, then no cross-weighting is required in the transformed coordinates. This is equivalent to a transformation of the data from $\{x^k\}$ to $\{Cx^k\}$.

Note from (15) and (21) that if the class means, m_N and m_Q, coincide, the linear discriminants all contain equal discrimination information and linear discriminant analysis offers no solution to the feature extraction problem.

4.4 Eigenvector Computation

The solution of (21) involves a matrix inversion which may not be feasible for covariance matrices of large dimension, due to ill-conditioning. In our studies on vectorcardiograms, 100 samples/axis at a sampling rate of 125 samples/sec were used, so that the covariance matrices were of dimension 300, and attempts at inversion failed. An iterative technique for obtaining dominant eigenvalues of non-symmetric matrices [20] was tried but was numerically unstable.

Another method of solving (21) is via simultaneous diagnonalization [17]. However, since all eigenvalues of S_w must be obtained, the ill-conditioning of S_w still prevented obtaining a solution. Kittler [21] who considered the same problem, has proposed a Fourier representation to obtain the eigenvalues of S_w. The K-L expansion reduces to the discrete Fourier transform if the process is wide sense stationary; in this case, the covariance matrix elements along diagonals parallel to the main diagonal are equal. While this may hold approximately in some problems, it did not for the covariance matrices encountered in this study(see Figs.5-8).

It appears that a worthwhile approach is the reduction of dimensionality, i.e., representation of the individual waveforms by fewer time samples. Since all waveforms are to be sampled at the same time points, the sample points may be selected on the basis of the ensemble statistics.

Although, ideally, selection should be based on minimizing the probability of misclassification, a more tractable method would be to minimize some measure of the expected representation errors for given interpolation polynomials. This is a problem in approximation theory which could be attacked using, as approximating classes, finite-order algebraic and trigonometric polynomials or splines [22].

V. COMPARISON OF LINEAR DISCRIMINANT METHODS

In this section we will consider a specific idealized two-class discrimination problem which exhibits a particularly interesting structure. The classes are shown in Figure 10. The feature space is the two-dimensional plane and in general, the classes will not be linearly separable. The problem considered is optimal linear feature extraction by reduction to one discriminant feature. The results of seven methods of feature extraction will be presented. These are the following: The K-L expansion, the Fukunaga-Koontz method [23], the Fisher method [24], discriminant analysis [17], the Chernoff method [25], the method of Chien and Fu [19], and the modified K-L expansion. The probability of classification error can be computed explicitly for each of these feature extractors. The deficiencies of these feature extractors are clearly pointed out by this example which represents a topologically complex two class configuration while maintaining analytical tractability.

(a) K-L Expansion In this case, the maximum eigenvalue of R of (6) is sought. Note that $R = S_b$ with c_N and c_Q the proportion of normal

and questionable records in the data. The decision rule for the discriminant direction is

$$\frac{a^2/2 + (c_Q/c_N)\,\overline{r^2}(1+S(2\theta)) + 2\,c_Q m_{Qx}^2}{b^2/2 + (c_Q/c_N)\,\overline{r^2}\,(1-S(2\theta))} \underset{y}{\overset{x}{\gtrless}} 1 \quad (24)$$

where $\overline{r^2} = (r_1^2 + r_2^2)/2$, $S(\theta) = (\sin\theta)/\theta$ and m_{Qx} is the x component of the mean of class Q:

$$m_{Qx} = \frac{2}{3}\,\frac{r_2^3 - r_1^3}{r_2^2 - r_1^2} S(\theta) \quad (25)$$

Note that increasing $\overline{r^2}$ and m_{Qx} favors selection of the x direction as desired. The y direction is always chosen if b is sufficiently larger than a, but may also be chosen if $a > b$, $\theta > \Pi/2$ and m_{Qx} is sufficiently small.

(b) Fukunaga-Koontz Method In this method a normalization technique is used for feature extraction. A mixture covariance matrix $R = R_N + R_Q$ is formed and transformed by a matrix U such that $URU^T = I$. Then two matrices $\bar{R}_N = UR_NU^T$ and $\bar{R}_Q = UR_QU^T$ are formed. The eigenvectors of $\bar{R}_N$ and $\bar{R}_Q$ are identical and the eigenvalues, denoted respectively by λ_i^N and λ_i^Q are related by : $\lambda_i^N = 1 - \lambda_i^Q$.

For classification purposes we wish to choose the features whose eigenvalues are closest to 0 or 1 in general. That is, eigenvalues which are far from 0.5 are more appreciated since they lead to a good dichotomy of the classes. A cost criterion suggested in [23] is

$$C = \frac{1}{m}\sum_{j=1}^{m}(\lambda_j - 0.5)^2 + \frac{1}{m}\sum_{j=1}^{m}\{(1-\lambda_j) - 0.5\}^2$$

As C increases the features are more effective. However in our problem where m = 1 the choice of either eigenvalue leads to the same cost due to symmetry. With this method two eigenvectors are necessary for two-class problems with different mean vectors. Therefore this method does not apply to this two dimensional example.

(c) Fisher Method:

In this method of two class discrimination we seek to compute a direction d in the feature space such that orthogonally projected samples from the two classes onto d are maximally discriminated. The discrimination criterion suggested by Fisher is related to the ratio of the projected class differences relative to the sum of the projected within-class variability. Specifically, the Fisher discriminant is obtained by solving for the unit vector d which maximizes the ratio

$$J = \frac{d^T M d}{d^T S_w d}$$

In order to solve for the Fisher direction d we take the derivative of J with respect to d and set equal to zero. This leads to:

$$[M - \lambda S_w]\, d = 0$$

The solution is obtained by solving for the eigenvectors. The rank of M is 1 and therefore only one nonzero solution exists. The eigenvector corresponding to the non-zero eigenvalue is the Fisher direction and is given by:

$$d = \alpha S_w^{-1}[m_N - m_Q]$$

where α is a normalization constant.

It can be seen that, regardless of the shape of the classes, the resulting discrimination direction will be along the vector joining the means of the two classes whenever these classes have diagonal covariances. This is not always a desirable direction as can be seen for the extreme example given in Figure 11. In this example knowing the y-coordinate of a sample is a much better discriminant for classification than knowing the x-coordinate. The Fisher method however will never give the y-direction for discrimination for this family of two-class problems.

(d) Linear Discriminant Analysis Using (21), the eigenvector associated with the principal eigenvalue of $S_w^{-1} S_b$ can be shown to always lie along the x axis (i.e., along the line joining the class means).

(e) Chernoff Method Chernoff suggests two closely related measures for discriminating between two distributions. The method leads to the discrimination direction d given by

$$d = S^{-1}(m_N - m_Q)$$

Here S is a particular mixture of the class covariance matrices

$$S = t\,R_N + (1-t)\,R_Q; \quad t \in [0,1]$$

and t is chosen to optimize a certain criterion. The details of the choice of t and the theoretical justifications for this choice can be found in [25]. For our two-class problem we can obtain the feature extractor without explicitly computing t. Since both R_N and R_Q are diagonal, S is diagonal and d is always along x.

(f) Chien and Fu Method In this method the eigenvalues of S_w are found, where c_N, c_Q are

interpreted as apriori class probabilities. Thus, S_w is the average within-class scatter matrix. The best linear discriminant direction is the eigenvector associated with the principal eigenvalue. Note that this method does not take the class means into consideration, which can be a deficiency in two-class discrimination.

(g) Modified K-L Expansion This method is studied by considering the effects of using risk parameters and time weighting of the data separately in the modified K-L expansion, as given by (10) and (11). For equal time weighting $Q = I$ and the relevant covariance matrix is S_b of (15) with $c_N = r_N p_N$, $c_Q = r_Q p_Q$, where r_N, r_Q are the misclassification risks, and the decision rule is given by (24). Generally $r_Q > r_N$ since misclassifying a questionable VCG is more costly than misclassifying a normal one. Inspection of (24) shows that as r_Q/r_N increases the x axis is favored, and as $r_Q/r_N \to \infty$ the decision rule becomes

$$S(2\theta) \underset{y}{\overset{x}{\gtrless}} 0$$

which favors x for $0 < \theta < \pi/2$ and y for $\pi/2 < \theta < \pi$, independent of a, b and m_{Qx}. Note that increasing c_Q has the effect of increasing the sensitivity to differences in the class means.

The effect of time weighting of the data can be seen by assuming that the problem is that of classifying on the basis of scalar measurements x(t) at two times, t_1 and t_2, and associating the values at these times with the x and y components respectively, of the two classes of Figure 10. Our "feature" is now the most informative time. The weighting matrix Q corresponding to the discriminant analysis solution can then be obtained from (22) and is found to be diagonal. Assuming $c_Q = c_N = 1/2$, the weighting of the sample at time t_1 relative to the one at time t_2 is:

$$\frac{Q_{11}}{Q_{22}} = \frac{b^2 + \overline{r^2}(1-S(2\theta))}{a^2 + \overline{r^2}(1+S(2\theta))} \left[1 + \frac{4m_{Qx}^2}{a^2 + \overline{r^2}(1+S(2\theta))} \right]$$

Note that $z(t_1)$ (the x component) gets heavier weighting as b increases, as a decreases, and as m_{Qx} increases, which is intuitively the correct behavior. As $\overline{r^2}$ increases, the sensitivity to the sector angle θ increases and in the limit as $\overline{r^2} \to \infty$, $z(t_1)$ is given higher weighting if $\pi/2 < \theta < \pi$ and $z(t_2)$ is given higher weighting if $0 < \theta < \pi/2$. For θ small, this is equivalent to giving the highest weighting to the time sample with smallest variation.

VI DISCUSSION

In this paper we have proposed the use of a modified K-L expansion to find the optimum set of features for VCGs. This modification was introduced due to our desire to time-weight the data and employ individual risk factors for each VCG in the training set. The weighting in the modified K-L expansion which gives the optimum two-class discriminants was found by maximizing a measure involving the within class and between-class scatter matrices. It would be desirable to generalize this to the multi-class problem.

Since the meaning of a single linear discriminant for more than two classes is not clear, it appears that some kind of local feature extraction is required, where distant classes are ignored.

Although this paper has considered only VCGs, the methods are applicable to ECGs as well. The main problem then is a computational one due to the increase in dimensionality. With a view to overcoming the dimensionality problem, a study has been conducted at the Charles Stark Draper Laboratory [15] to determine the feasibility of estimating the VCG (3 signals) of a patient from his ECG (12 signals) for purposes of data reduction. It has been found that the average heartbeat waveforms of an individual as obtained from the ECG and the VCG, are very accurately related by a linear transformation, with the provision that phase shifts among the ECG signals are accounted for in the transformation. However attempts to find a single transformation which is valid for an entire group of individuals have not been successful.

The persisting need to reduce the set of ECG signals then led to an investigation of principal factor analysis. The results of a study performed on a set of nine ECG records supplied by USAF/SAM indicate that the use of four, five, or six standard principal factors accounts, respectively, for 97.7%, 98.8% and 99.3% of the total ECG signal energy. This, coupled with visual assessment of original and reconstructed waveforms, suggests that data reduction from 12 to approximately 6 to 8 signal components via a standard transformation may suffice for ECG analysis. Furthermore, comparison with VCG data for the same patient set shows that the intrinsic plane concept used for the VCG holds for the ECG to essentially the same degree of accuracy. It was found that the VCG and ECG intrinsic planes are aligned to within 13 degrees in all cases studied.

An intermediate stage between feature selection and pattern classification is cluster analysis. Cluster analysis may be used to aid both the feature selection and the pattern classification process. Several comments should be made at this point. Firstly, the feature selection and pattern classification processes are not decoupled, since a true measure of good feature selection is minimizing pattern classification

error. The whole process is indeed a closed-loop system not unlike that of optimum stochastic control where the estimation and control functions cannot in general be separated. What we are doing here in control parlance is to do a linear analysis in which we decouple the feature selection and pattern classification steps. Since the optimum classifier is in general non-linear this decoupling in general is not valid. Furthermore the optimum features are probably nonlinear functions of the data.

In order to better approximate the overall process, one might consider cluster analysis as an intermediate stage between feature selection and pattern classification. The basic idea here could be that good feature selection leads to compact clusters with maximum separation and hence good pattern classification. Hence if the initial features were chosen in a sufficiently high dimensional space to give good approximation of the waveforms and cluster analysis performed in this high dimensional space, optimum features in a lower dimensional space could be obtained by taking into appropriate account the local geometry of the clusters as well as their global configuration.

Finally, some comments should be made about the choice of norms in the feature selection process. So far we have used the ℓ_2-norm. It appears that this particular choice may distort the true information metric, since large deviations are not always associated with high information content. The ℓ_1 norm might be a better choice for our purposes. However, the choice of optimal basis vectors with respect to an ℓ_1-norm appears to be an open problem.

ACKNOWLEDGEMENT

The authors wish to thank Col. Malcolm C. Lancaster and Col. John H. Triebwasser of the U.S. Air Force School of Aerospace Medicine for their advice and encouragement in this work.

REFERENCES

1. Bonner, R.E., Crevasse, L., Ferrer, M.I. and Greenfield, J.C., "A New Computer Program for Analysis of Scalar Electrocardiograms", Comp. Biomed. Res. Vol. 5, p. 629, 1972

2. Cornfield, J., Dunn, R.A., Batchlor, C.D. and Pipberger, H.V., "Multigroup Analysis of Electrocardiograms", Comp. Biomed. Res., Vol. 6, pp. 97-120, 1973

3. Smith, R.E. and Hyde, E.M., "Computer Analysis of the Electrocardiogram in Clinical Practice", in Electrical Activity of the Heart, edited by G.W. Manning and H.P. Ahuja, Charles C. Thomas, Springfield, Ill., p. 305, 1965

4. Wolf, H.K., Macinnis, P.J., Stock, S., Helppi, R.K. and Rautaharju, P.M., "The Dalhousie Program: A Comprehensive Analysis Program for Rest and Exercise Electrocardiograms", in Computer Application on ECG and VCG Analysis, edited by Chr. Zyweitz and B. Schneider, North-Holland, 1973

5. Specht, D.F., "Vectorcardiographic Diagnosis Using the Polynomial Discrimant Method of Pattern Recognition", IEEE Trans. Biomed. Elect. Vol. BME-14, pp 90-95, April 1967

6. Bailey, J.J., Horton, M. and Itscoitz, S.B., "A Method for Evaluating Computer Programs for Electrocardiographic Interpretation. III. Reproducibility Testing and Sources of Program Errors", Circulation, Vol. 50, pp. 88-93, July 1974

7. Cover, T.M., "The Best Two Independent Measurements are not the Two Best", IEEE Trans. on Systems, Man and Cybernetics, pp. 116-117, Jan. 1974

8. Balm, G.J., "Crosscorrelation Techniques Applied to the Electrocardiogram Recognition Problem", IEEE Trans. Bio-Med. Eng., BME-14, No. 4, October, 1967

9. Mucciardi, A.N. and Gose, E.E., "A Comparison of Seven Techniques for Choosing Subsets of Pattern Recognition Properties", IEEE Trans. Comput. Vol. C-20, No. 9, pp. 1023-1031, 1971.

10. Okajima, M., Stark, L., Whipple, G., and Yasui, S., "Computer Pattern Recognition Techniques: Some Results with Real Electrocardiographic Data", IEEE Trans. on Bio-Med. Elect., pp. 106-114, July 1963

11. Karlsson, S., "Representation of ECG Records by Karhunen-Loeve Expansions", Digest of 7th Int. Conf. on Medical and Biological Eng., Stockholm, August, 1967

12. Hambley, A.R., Moruzzi, R.L., and Feldman, C.L., "The Use of Intrinsic Components in an ECG Filter", IEEE Trans. Bio-Med. Eng. Vol. 21, Nov. 1974

13. Kittler, J. and Young, P.C., "A New Approach to Feature Selection Based on the Karhunen-Loève Expansion", Pattern Recognition, Vol. 5, pp 335-352, 1973

14. Womble, M.E., Halliday, J.S. and Mitter, S.K., "Application of Estimation Theory to the Classification of Vectorcardiograms", presented at AAS/AIAA Astrodynamics Specialist Conference, Vail, Colorado, July 16-18, 1973

15. Gustafson, D. E., Johnson, T. L. and Akant, A., "Cardiogram Analysis and Classification Using Signal Analysis Techniques", C.S. Draper Laboratory Rept. R-853, Sept. 1974

16. Wilkinson, J. H. and Reinsch, C., Handbook for Automatic Computation, Vol. II, Linear Algebra, Springer Verlag, New York, 1971

17. Fukunaga, K. Introduction to Statistical Pattern Recognition, Academic Press, New York, 1972

18. Watanabe, S., "Karhunen-Loeve Expansion and Factor Analysis, Theoretical Remarks and Applications", Proc. Fourth Prague Conf. Inf. Theory (also Yale Univ. Tech Rept.) 1965

19. Chien, Y. T. and Fu, K.S., "Selection and Ordering of Feature Observations in a Pattern Recognition System", Inf. and Cont., Vol. 12, pp. 395-414, 1968

20. Franklin, J.N., Matrix Theory, Prentice-Hall, New Jersey, 1968

21. Kittler, J., "A Method of Feature Selection for High Dimensional Pattern Recognition Problems", 2nd Joint Int. Conf. on Patt. Recog., Aug. 1974.

22. McClure, D. E., "Nonlinear Segmented Function Approximation and Analysis of Line Patterns", Quarterly of Applied Math., Vol. 33, No. 1, pp. 1-37, April 1975

23. Fukunaga, K. and Koontz, W. L. G., " "Application of the Karhunen-Loève Expansion to Feature Selection and Ordering", IEEE Trans. Comput. Vol. C-19, No. 4, pp. 311-318, 1970

24. Fisher, R. A., "The Use of Multiple Measurements in Taxonomic Problems", Annals Eugenics, Vol. 7, pp. 179-188, Sept. 1936.

25. Chernoff H., "Some Measures for Discriminating between Normal Multivariate Distributions with Unequal Covariance Matrices", Multivariate Analysis - II, Academic Press, New York, 1973

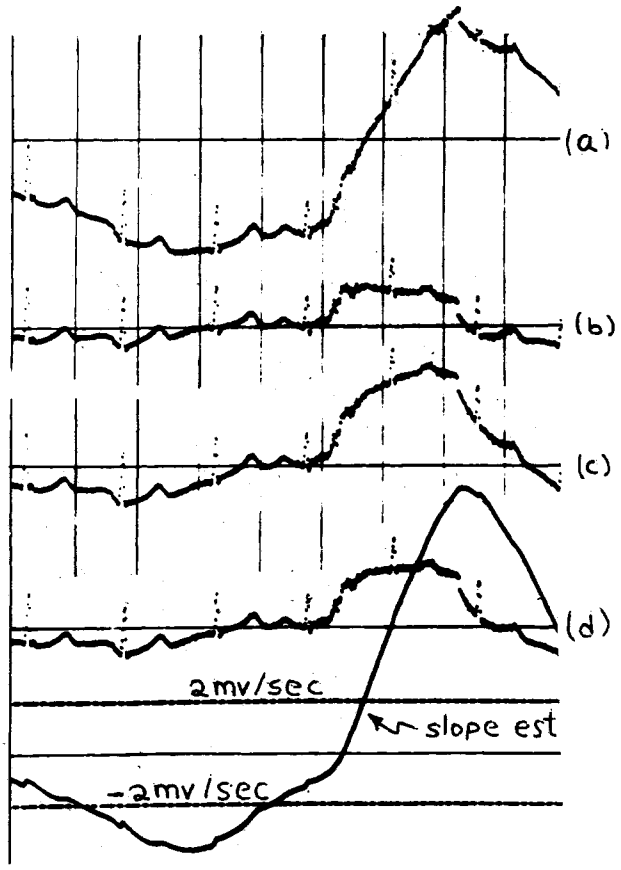


Fig. 1. Baseline Removal with Severe Jump. (a) Original record, Estimates Using: (b) 1st order Kalman filter, (c) 1st order adaptive filter, (d) 2nd order Kalman filter

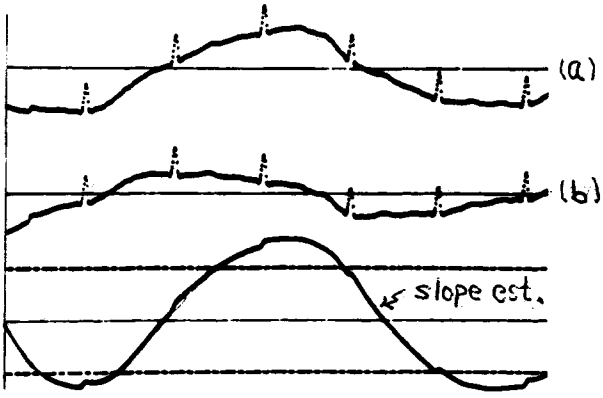


Fig. 2. Baseline Removal with Respiratory component. (a) Original record, (b) Estimates using 2nd order Kalman filter

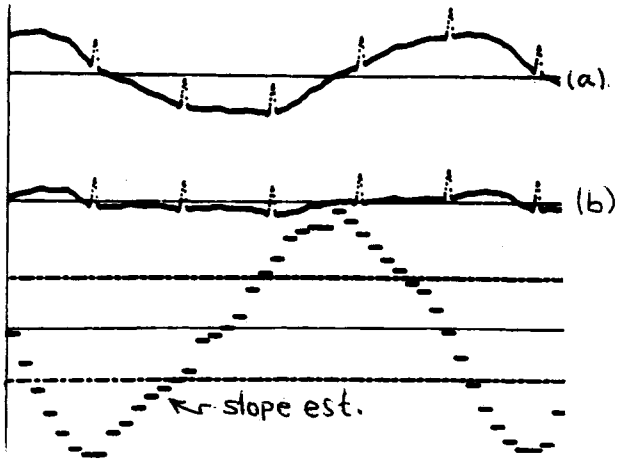


Fig. 3. Baseline Removal with Respiratory Component (a) Original record, (b) Estimates using non-adaptive moving window filter

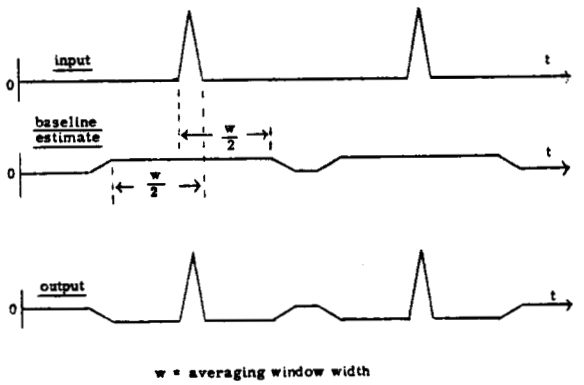


Fig. 4. Spurious T wave generation using non-adaptive moving window filter

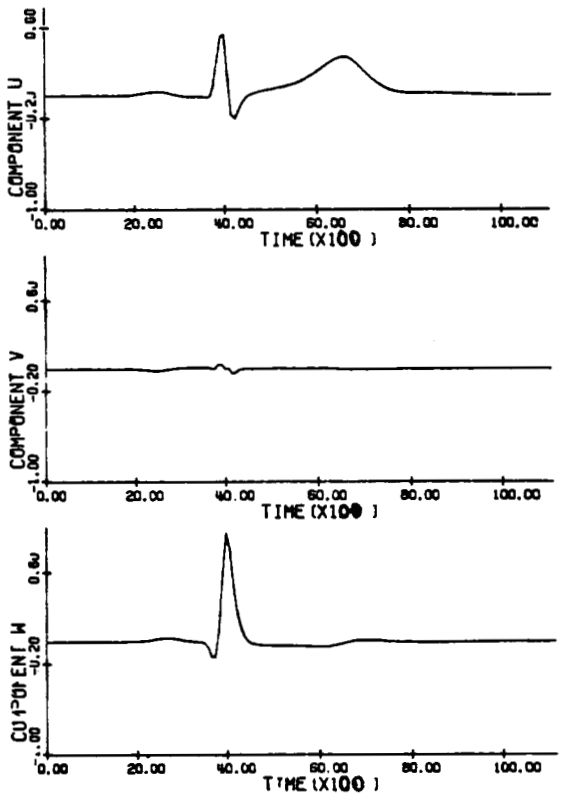


Fig. 5. Ensemble mean VCG

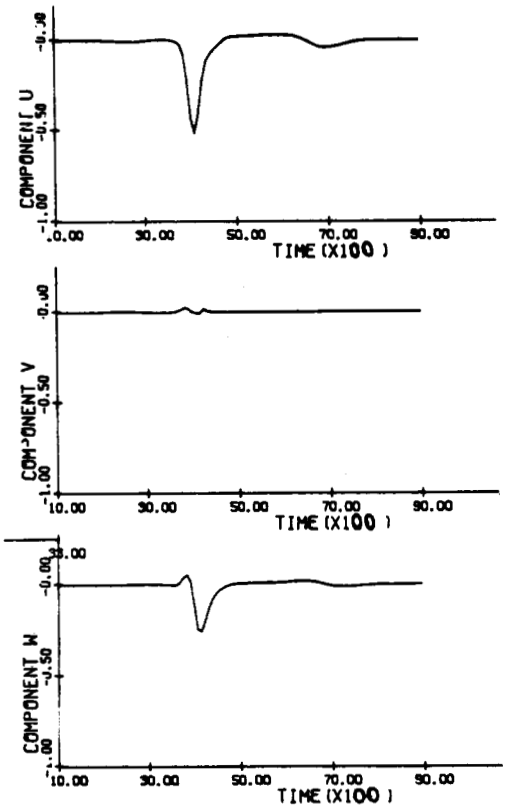


Fig. 6. 1st K-L eigenfunction

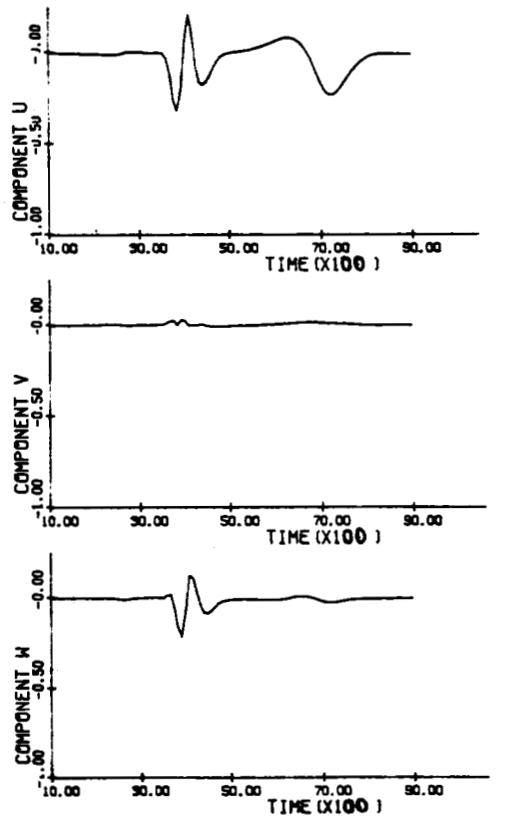


Fig. 7. 4th K-L eigenfunction

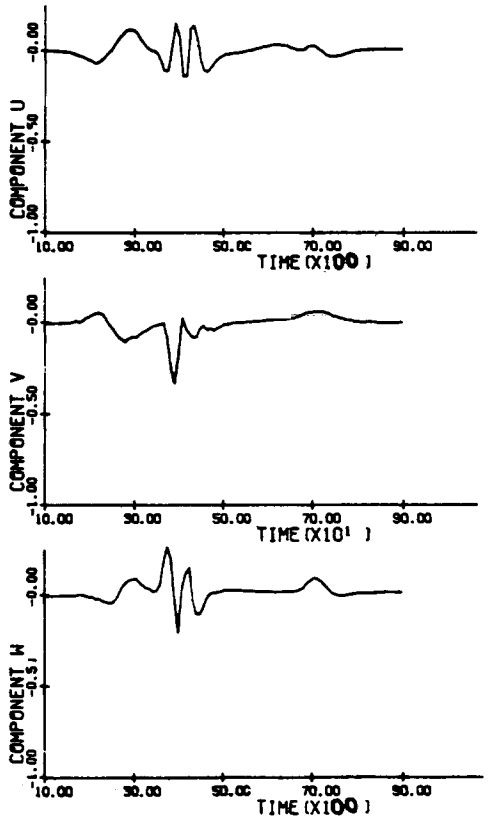
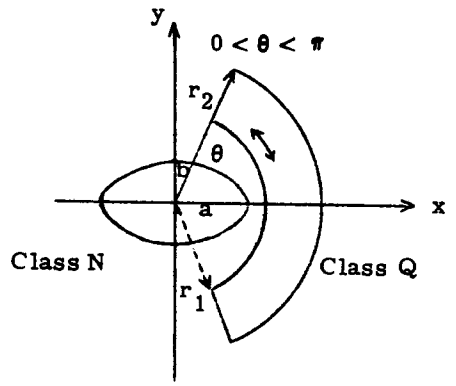


Fig. 8. 13th K-L eigenfunction



Class N:

$$p_N(\underline{x}) = \begin{cases} \frac{1}{\pi ab} & \text{if } x \in \text{ellipse} \\ 0 & \text{otherwise} \end{cases}$$

Class Q:

$$p_Q(\underline{x}) = \begin{cases} \frac{1}{\theta\,(r_2^2 - r_1^2)} & \text{if } \underline{x} \in \text{sector} \\ 0 & \text{otherwise} \end{cases}$$

Fig. 10. Two-class, two-dimensional discrimination example

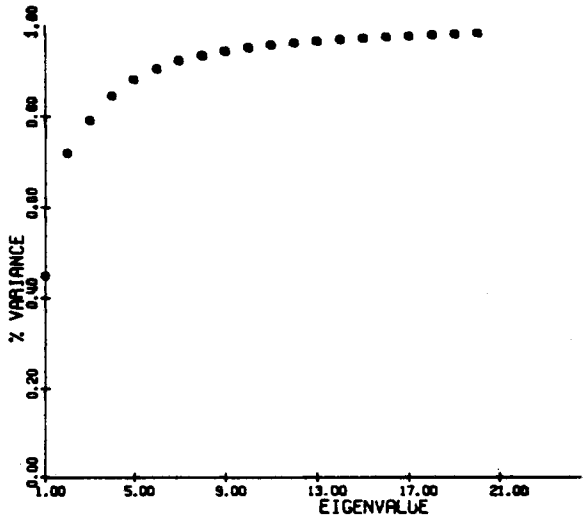


Fig. 9. Percentage of total ensemble energy in first few eigenfunctions

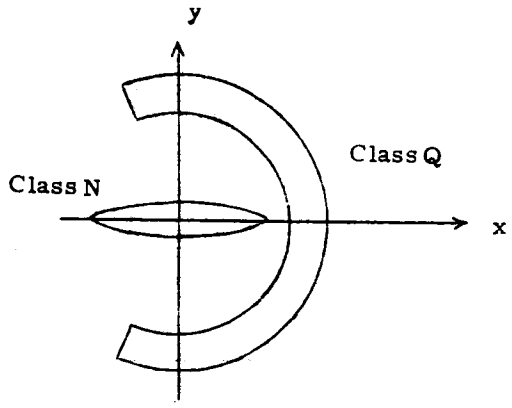


Fig. 11. A configuration which favors y-axis discrimination