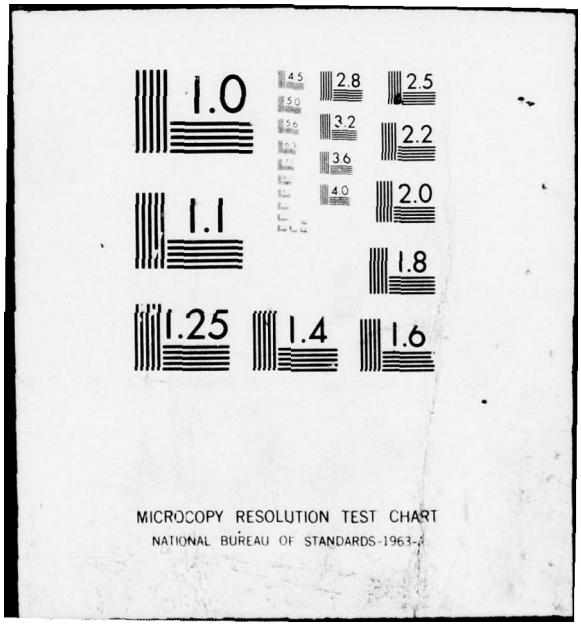
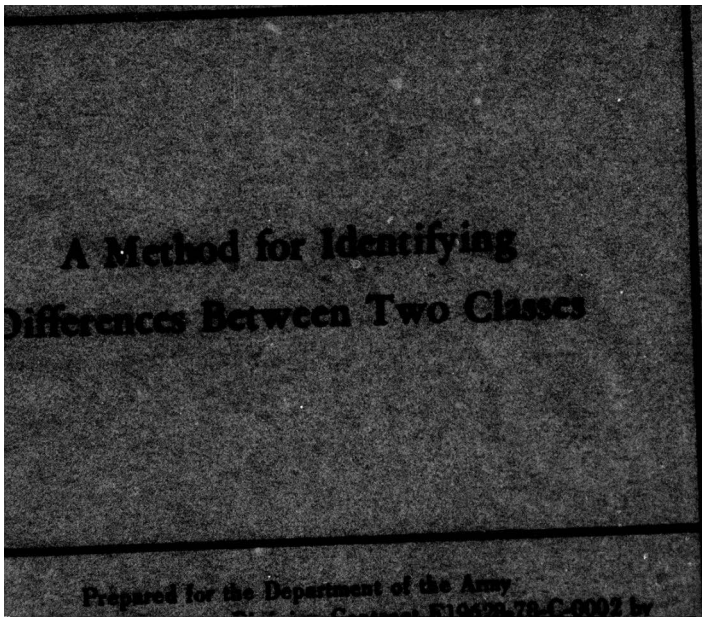
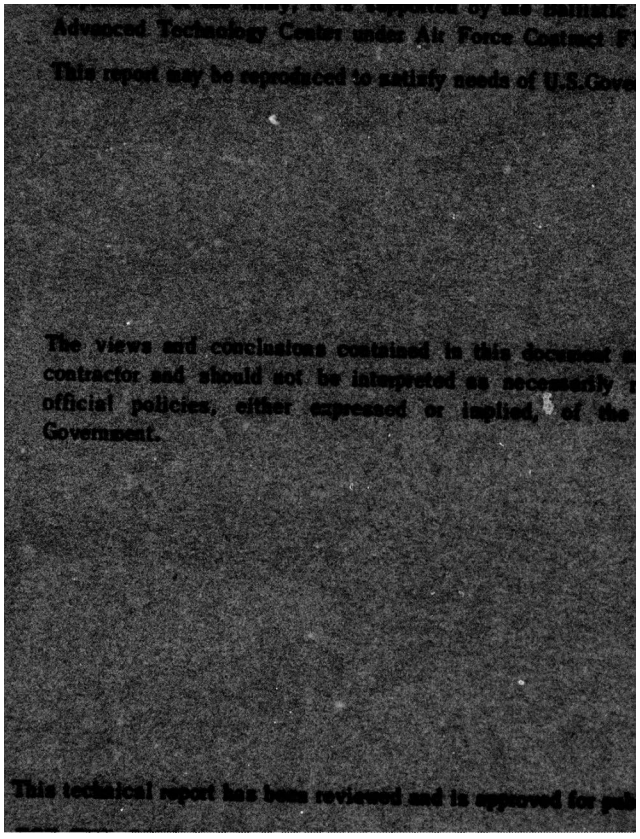
AD-A064 724 MASSACHUSETTS INST OF TECH LEXINGTON LINCOLN LAB F/G 12/1
A METHOD FOR IDENTIFYING DIFFERENCES BETWEEN TWO CLASSES.(U)
NOV 78 N E LINDGREN F19628-78-C-0002
UNCLASSIFIED TN-1978-38 ESD-TR-78-296 NL

1 OF 1
AD
AD A064724

END
DATE FILMED
4-79
DDC


1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0
4.5
5.0
5.6
MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A


A Method for Identifying
Differences Between Two Classes
Prepared for the Department of the Army


Advanced Technology Center under Air Force Contract F
This report may be reproduced to satisfy needs of U.S.Gove
The views and conclusions contained in this document a
contractor and should not be interpreted as necessarily
official policies, either expressed or implied, of the
Government.
This technical report has been reviewed and is approved for pub

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

LINCOLN LABORATORY

A METHOD FOR IDENTIFYING DIFFERENCES BETWEEN TWO CLASSES

N. E. LINDGREN

Group 32

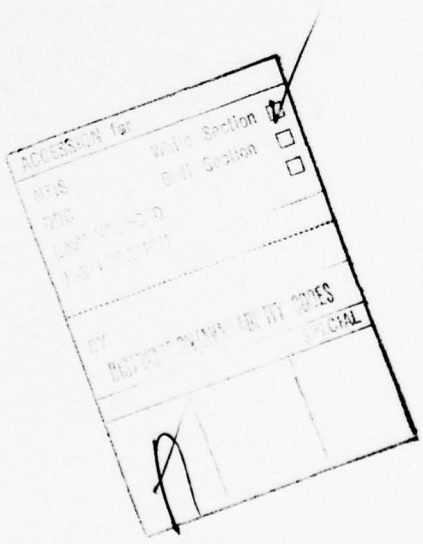
TECHNICAL NOTE 1978-38

29 NOVEMBER 1978

LEXINGTON MASSACHUSETTS

ABSTRACT

A non-parametric measure of the difference between sample distributions of a random variable for two classes of data is presented. The method involves counting the number of class reversals among the ordered set of two class data and provides a normalized measure of class intermingling. Applications of the method to the discrimination-feature selection problem are described.



CONTENTS

I. INTRODUCTION

Whenever data is collected by machine in some experiment as for example in radar observation of some object, or the taking of an electrocardiogram, or seismic exploration, etc., there is often a super abundance of data collected on each radar pulse, single heartbeat, or single detonation, etc., of which only a small fraction is useful for determining the distinguishing aspects of the particular experiment. In the case of a single radar pulse illuminating an object there might be over 2700 items of data recorded, such as amplitude and phase for two polarizations for perhaps hundreds of incremented ranges that encompass the object. If it is desired to know what category of object is reflecting the radar pulse, it can turn out that only a few items out of the 2700 data are important, and if it can be determined in advance through preliminary studies which those important data are, then an enormous savings in computation can be achieved by editing out the unimportant data. This is the classical problem of feature selection.

The purpose of this note is to describe a technique called the method of reversals, which can be used to select what items of data or what types of measurements are important, in first order, for distinguishing between two categories or two classes of subjects. It is a method for sifting through the two class data base and quantifying the importance of each measurement

type. Just a few examples of two class problems are distinguishing between threatening and non-threatening vehicles in the radar defense case, healthy and abnormal heart patients in the electrocardiogram case, and the presence or absence of subterranean oil in the seismic exploration case. Essentially the method is one of determining the degree of intermingling of the two classes for each kind of measurement or for a given combination of measurements. If, for some particular measurement type or combination of types, the data for the two classes is thoroughly intermingled, then that measurement or combination by itself is worthless for distinguishing between the two classes. Conversely, if some measurement or combination easily distinguishes between the two classes then the degree of class intermingling for it will be small.

A number of feature extraction techniques rely on rotating an N dimensional data space (each of the dimensions corresponds to one of the N observables) in a manner so as to preserve a maximum amount of discrimination information when some of the dimensions, after rotation, are eliminated from consideration. The choice of rotation and selection of saved dimensions often requires ordering the eigenvalues of some combination matrix of the two class correlation (or covariance) matrices If, however, the data population of one of the classes is fewer than the number of dimensions, then the correlation or

covariance matrices cannot be determined and the rotation techniques are inapplicable. A not infrequent example in the radar case is to have fewer pulses of data than the number of (interesting) samples in range taken for each pulse, in which case the needed matrices cannot be formed. If the data population is small, but large enough to estimate the required matrices, those estimates can be very poor due to the low density of data in the full N dimensional space. In such cases an examination of class separability for each observable alone or some combination of a few observables taken together can be very useful in identifying important features for discrimination.

It is important to note here that the reversals method is not an optimal discrimination-feature extraction technique.* The method is useful for looking at class differences of a single random variable, which can be a single measurement type or a combination of measurement types. By quantifying the class differences of the single random variables, one can rank them in their ability to differentiate between the two classes. If the user wishes to examine combinations of observables, to make use of potential class separation that may reside in the correlations

---

*Optimality is defined in terms of some particular set of rules, so that even optimal discrimination-feature extraction techniques need not be good feature extraction techniques when the rules are not well chosen for a given problem.

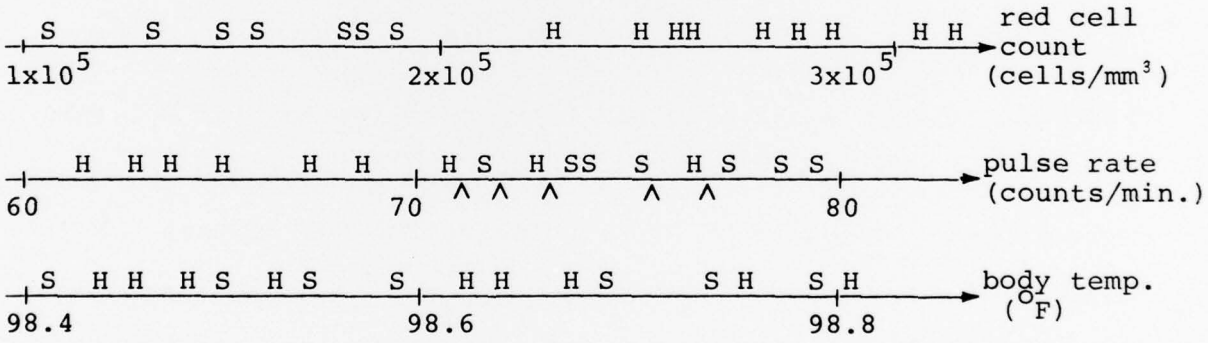
between observables, then the user must choose what combinations of observables to examine with the technique. There are intelligent methods for determining appropriate combinations of observables (such as determining the Fisher linear classifier projection direction for correlated pairs of observables), but that aspect of the problem is not the focus of this paper. The situation may be summed up as follows. Whatever separability is found and ranked by the reversals technique is there, but there is no guarantee that all the available separation has been uncovered.

The method of reversals is one technique among many for quantifying the differences of single variables between two classes. Other techniques, described and compared in the last section, are the K factor method (finding the separation in the class mean values normalized by the average intra-class standard deviation), the equal classification-error value for a single threshold test, the distance measures (Bhattacharyya, Kullback divergence, Matusita, etc.), and some non-parametric tests for determining at what level of significance the two class populations are considered different.

In the next section the reversals method is introduced by way of a simple example.

II. AN ILLUSTRATIVE EXAMPLE

Suppose we are trying to determine the important measurements for early diagnosis of disease X and we have data for many kinds of measurements such as body temperature, pulse rate, red cell count, etc. We have a data bank of these measurements for a class of N_1 healthy people as well as N_2 people suffering from disease X. The first step in the technique is to order all the values for each measurement regardless of class. Let an H denote each of the ($N_1 = 9$) healthy persons and an S denote each of the ($N_2 = 7$) sick persons in the illustration below.



Next, scan along each measurement axis and count the number of class reversals encountered. For example on the red cell count axis there is only one reversal, where the string of S's meets the string of H's. Let the reversal count be denoted by R, i.e., R=1 for the red cell measurements. For the pulse rate measure-

ment, we get R = 5, as shown by the caret symbols beneath the axis at the reversal points. For the body temperature measurement the reversal count is R = 9. Obviously the red cell count measurement distinguishes best between the two classes for these three axes.

Since the number of reversals on any axis will usually increase if the number of members in the two classes is increased (unless, for example, there is perfect separation between the classes), it is desired to have a normalized measure of importance which is independent of the number of data points N_1 and N_2. The measure W defined as

$$W \triangleq \frac{\hat{R} - R}{\hat{R} - 1}$$

has that property and also has the properties that $W \rightarrow 1$ for perfect separation of the two classses and $W \rightarrow \pm\epsilon \approx 0$ for those measurements incapable of distinguishing the two classes. $\hat{R}$ is defined as the number of reversals expected if class 1 and class 2 were statistically identical for the particular measurement. It is calculated from the formula

$$\hat{R} = \frac{2 N_1 N_2}{N_1 + N_2}$$

In the present example, $N_1 = 9$ and $N_2 = 7$ for all the axes so that $\hat{R} = \frac{2 \times 9 \times 7}{9 + 7} = 7.875$. The reversal measure W is then

calculated for each axis.

$$W_{\text{red cells}} = \frac{7.875 - 1}{7.875 - 1} = 1$$

$$W_{\text{pulse rate}} = \frac{7.875 - 5}{7.875 - 1} = 0.418$$

$$W_{\text{temperature}} = \frac{7.875 - 9}{7.875 - 1} = -0.164$$

Looking at the resultant W values we would then conclude that the red cell count measurement is an outstanding measurement for diagnosing disease X, the pulse rate measurement is of moderate value, and the body temperature measurement used alone is worthless. Negative values of W arise when more reversals are counted than the expected number of reversals for statistically identical classes.

The results of a different example, a defense radar discrimination problem, are presented in Fig. 1. The radar returns from 80 different ranges were sampled, encompassing the body which was from 1 of the 2 classes, re-entry vehicles and decoys. The reversal measure W is plotted for each of the 80 range gates (solid line) and for the difference combination $P_i - P_{i-5}$ (dashed line), where P_i is the radar return in range gate i. The difference combination is proportional to the slope estimate of P as a function of range at the mid-range point.

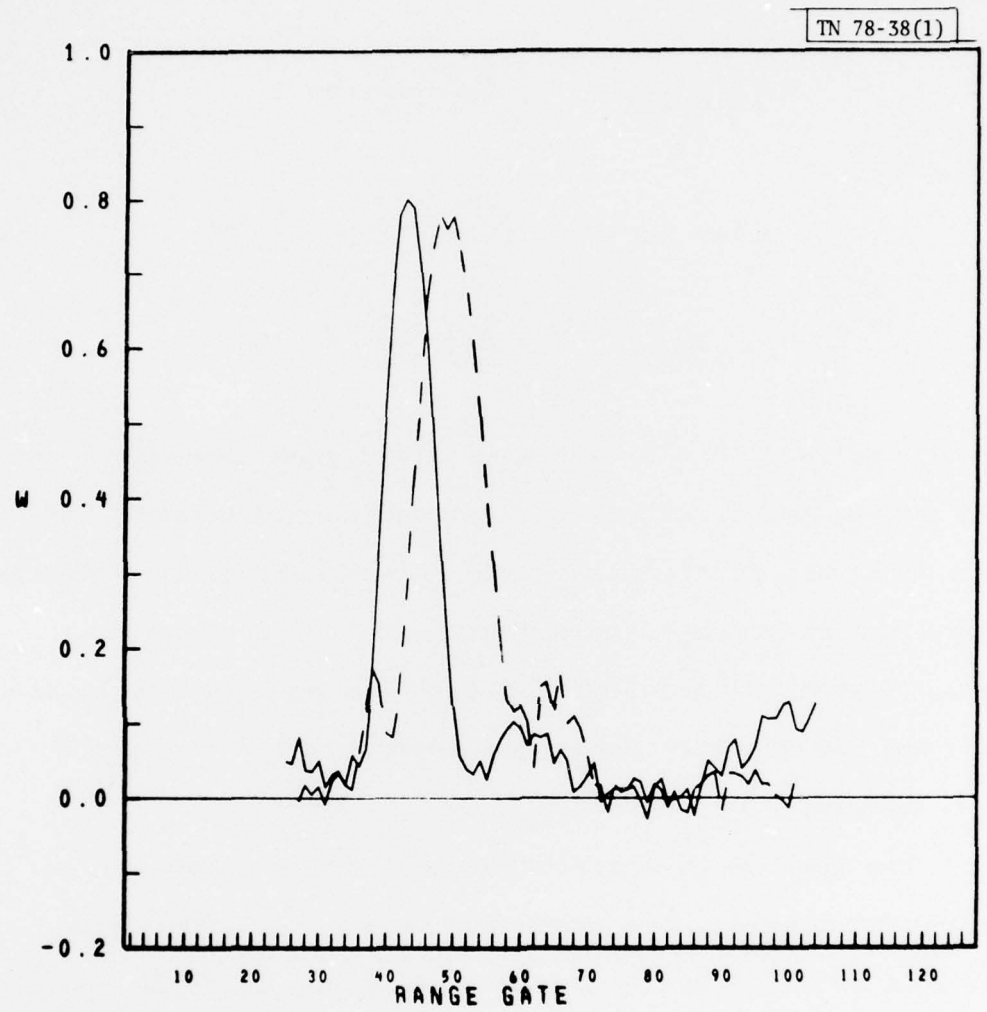


Fig.1. Comparison of reversal measure W for a radar discrimination problem utilizing the range profiles of two classes, re-entry vehicles and decoys. The profiles are most different in their absolute values (solid line) at range gate 43. The profiles are statistically most different in their slopes (dashed line) in the vicinity of range gate 49.

The highest values of W indicate which of these individual or combination measurements are most capable of separating the two classes. From the figure it can be seen that the range gates at and near gate 43 are important on an individual basis and the slopes of the radar return profiles are very different for the two classes in the vicinity of range gate 49.

In order to see how W scales with various kinds of class differences, the results of a number of calibration exercises are presented in the following section for statistically controlled examples. The examples demonstrate how W increases as the class distributions become more different.

III. CALIBRATION EXAMPLES

For unlike distributions of two classes of data, the mathematics for expected reversal count and standard deviation of reversal count is exceedingly difficult and depends on the particular distributions considered. Consequently Monte Carlo experiments have been employed for studying unlike distributions. The one case where mathematical analysis is tractable is when the two classes are identical. Those results are included briefly here.

It is well known from the statistical theory of runs[1-3] that when the two classes are identical the probability density of reversal count is

$$p(R) = \begin{cases} \dfrac{2\binom{N_1 - 1}{\frac{R-1}{2}}\binom{N_2 - 1}{\frac{R-1}{2}}}{\binom{N_1 + N_2}{N_1}} & (R \text{ odd}) \\ \dfrac{\binom{N_1 - 1}{\frac{R}{2}}\binom{N_2 - 1}{\frac{R}{2} - 1} + \binom{N_1 - 1}{\frac{R}{2} - 1}\binom{N_2 - 1}{\frac{R}{2}}}{\binom{N_1 + N_2}{N_1}} & (R \text{ even}) \end{cases}$$

where $\binom{a}{b} = \frac{a!}{b!(a-b)!}$

The expected number of reversals is $\hat{R} = 2N_1N_2/(N_1 + N_2)$ and the variance is

$$\sigma_R^2 = \frac{2N_1N_2(2N_1N_2 - N_1 - N_2)}{(N_1 + N_2)^2(N_1 + N_2 - 1)}$$

Consequently, $E\{W\} = 0$ and $\sigma_W = \frac{\sigma_R}{\hat{R} - 1}$.

When the class distribution densities are not identical there are no general formulas for $E\{W\}$ and σ_W since they depend on the specific class density functions. Monte Carlo experiments have been run for special families of the class densities, as the two classes are made more different. In the following three cases, differences in the means, spreads, and skewness are respectively illustrated.

a. Two Gaussians with Different Means

In a first series of experiments, two Gaussian densities having the same standard deviation are separated by increasing amounts between their means. At each stage of separation 100 experiments were performed, each time generating 200 random numbers from each Gaussian distribution, counting the class reversals, and calculating the value of W. In Fig. 2 the mean value of W at each stage of separation is plotted, along with

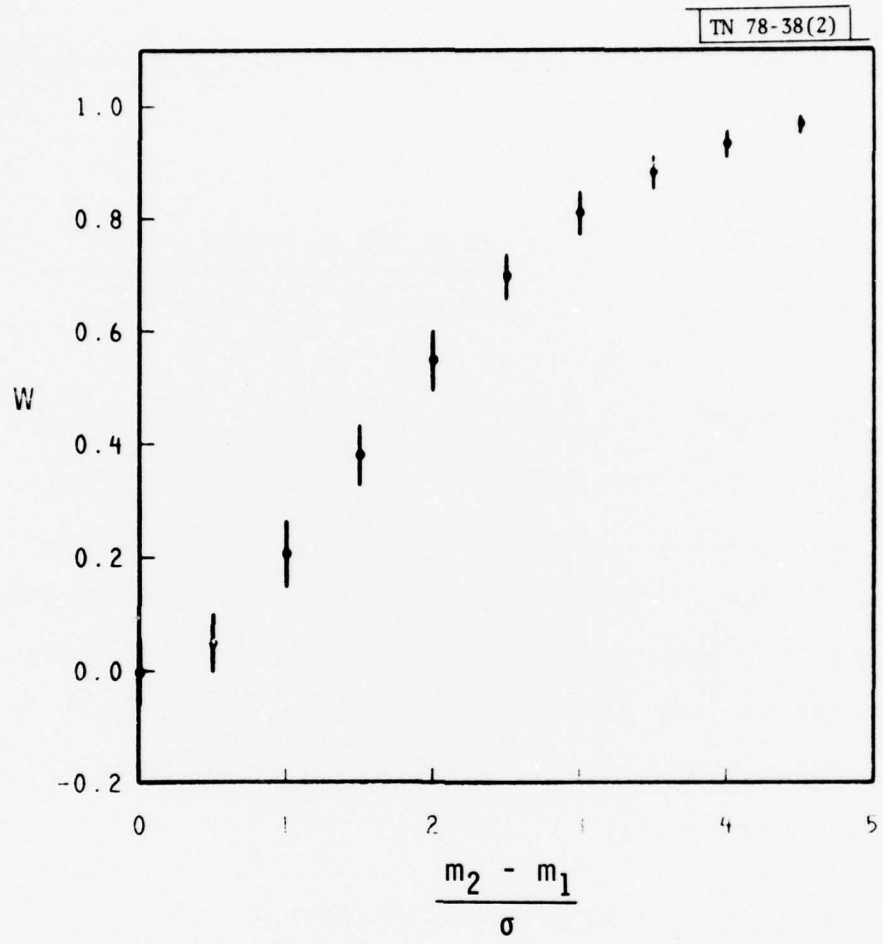
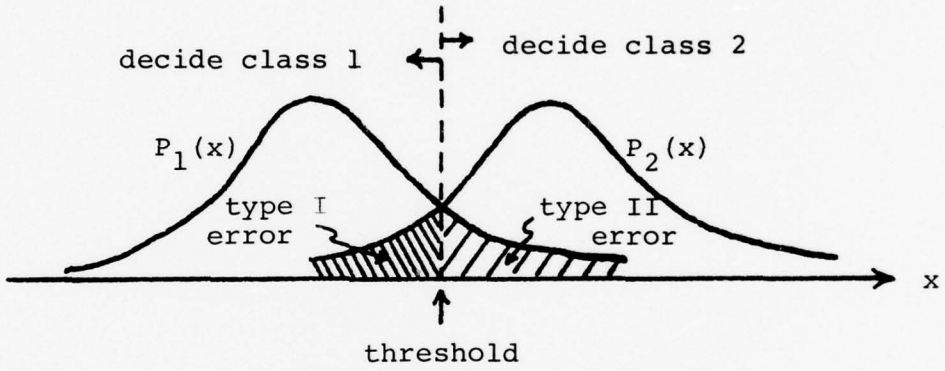


Fig.2. Reversal measure as a function of the separation of the means of two Gaussian distributions with the same spread. The abscissa is identical to the often used K value.

the associated error bars which extend one standard deviation above and below the mean value.

A different measure of class separation of two Gaussians having the same standard deviation is the equal error value on an operating characteristic curve describing the decision performance of a threshold test. The expected location of the threshold for equal decision errors is where the two Gaussian density functions cross, as seen in the sketch.



It can be easily visualized that when the two distributions are pulled further apart, the equal error value diminishes. In Fig. 3 is shown the correlation between reversal measure and equal error value in a series of 35 Monte Carlo experiments.

b. Two Gaussians with Different Spreads, but the Same Mean

Here is a case where the single threshold equal error measure is blind to differences in the distributions. The

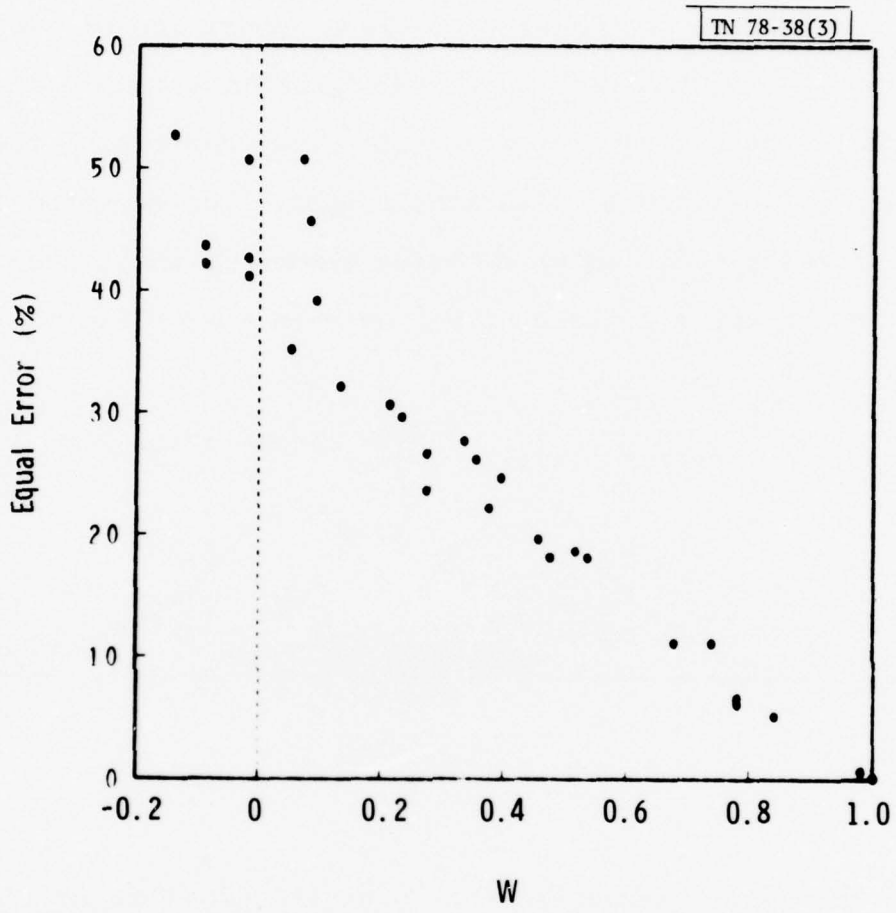


Fig.3. Correlation of reversal measure with equal error measure for Gaussian distributions with the same standard deviation, but different means. $N_1 = N_2 = 100$.

expected equal error is always 50% regardless of how unlike the distribution spreads are. The reversals technique is able to detect differences in spread. The results of a series of Monte Carlo experiments, where the standard deviation of one Gaussian distribution is varied in stages while the other Gaussian remains fixed, are shown in Fig. 4.

c. Two Skewed Distributions with the Same Mean and Spread

χ^2 distributions with ν degrees of freedom (ν = 2,3,4, 5, and 6) were used in this part. The distribution skewness, defined as μ_3/σ^3, increases as ν decreases. For a fixed value of ν, 100 experiments were performed, each as follows. 200 random numbers from the χ^2 distribution were generated as the class 1 data. The mean of the numbers was found and then each number was reflected about the mean and the resultant became an element of class 2. Thus, classes 1 and 2 had the same mean and standard deviation, and were different only in their third and higher central moments. An example histogram for the two class distributions is shown in Fig. 5 for the case ν = 3. The mean values and standard deviations of W as a function of the difference in sample skewness between the two distributions are shown in Fig. 6.

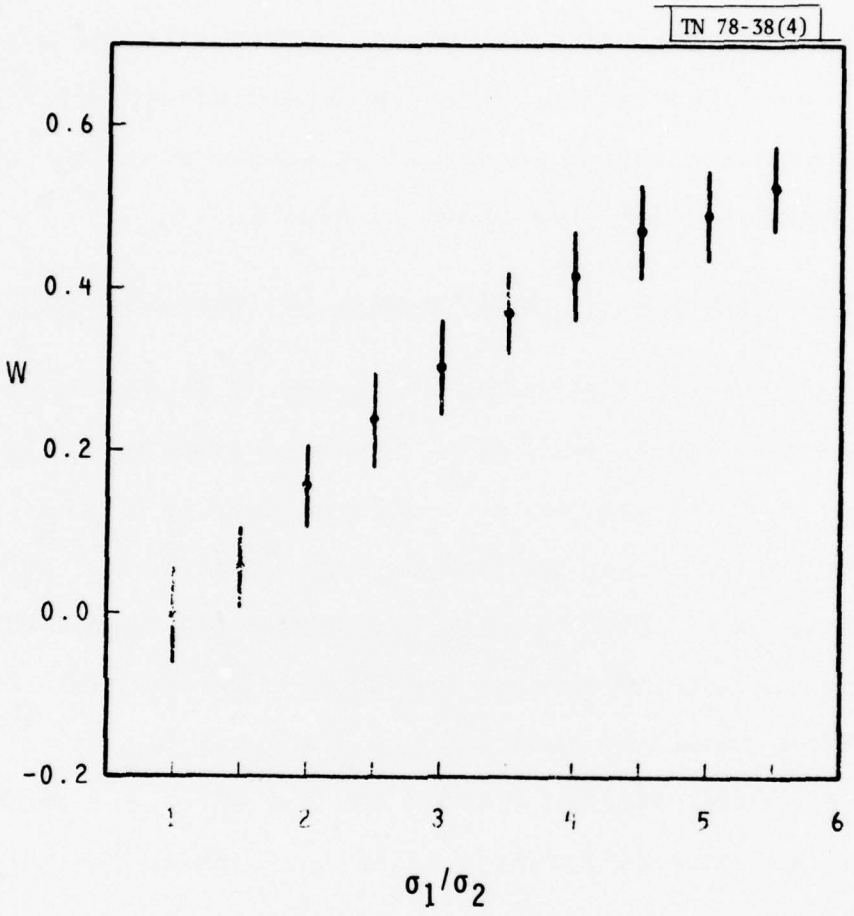


Fig.4. Reversal measure for two Gaussian distributions with different standard deviations, but with the same mean.

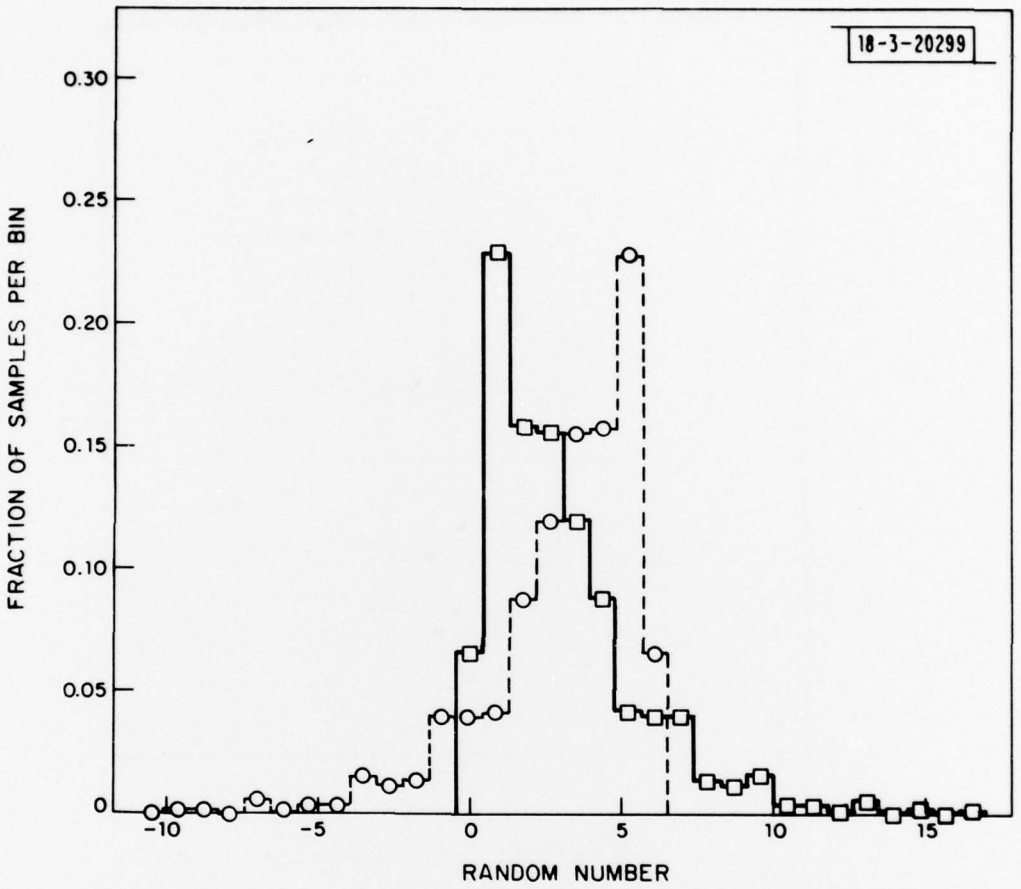


Fig.5. Example of distributions having the same mean and standard deviation. The lowest order difference is their skewness.

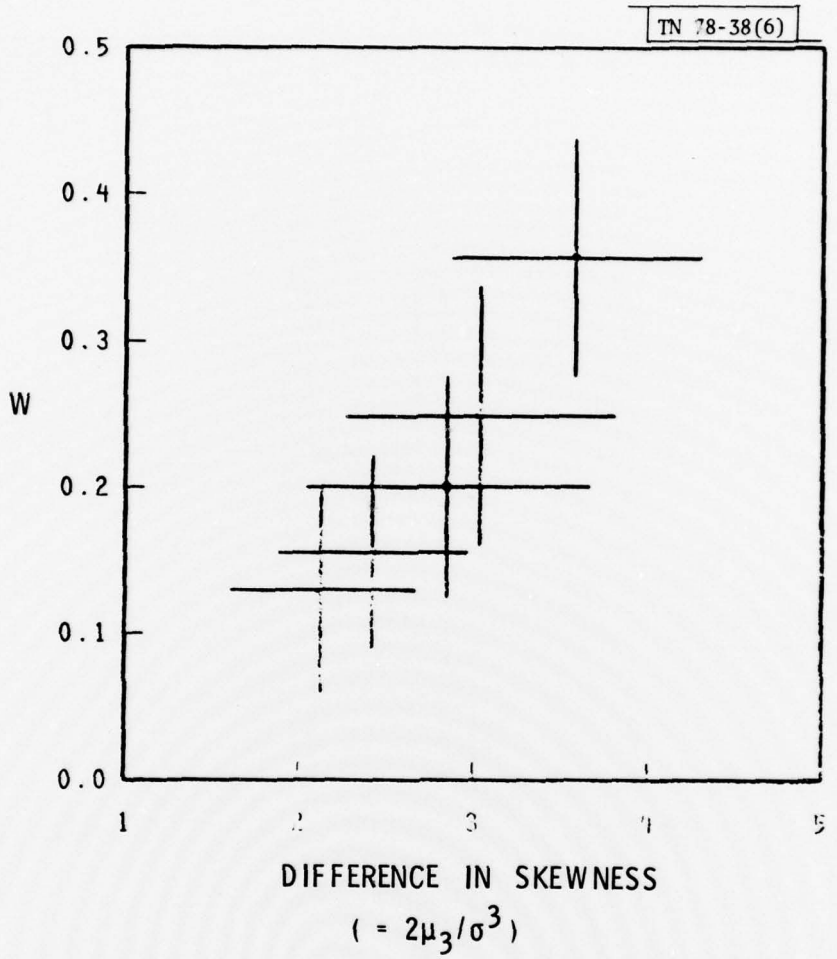


Fig.6. Reversal measure as a function of the difference in skewness of two χ^2 distributions with the same means and standard deviations.

IV. COMPARISON WITH OTHER FEATURE SELECTION TECHNIQUES

The reversals technique is a non-parametric method of measuring the difference between two distributions of a single random variable. Since the difference is quantified, in the measure W, the method is useful for comparing different random variables, which can be individual types of measurements or can be combinations of measurement types, for their ability to separate two classes. The technique has the benefit, when the nature of the underlying distributions are unknown, that no assumptions are made about the distributions and no parameters of the distributions are estimated. The type of scale on which the raw measurements for the two classes are made is unimportant. For example, the scale could be linear or logarithmic, or any other monotonically increasing scale and the results of the ordering of the data, and hence W, would be unaffected. The data need be only ordinal. For example one could measure the class difference between male and female runners in some race by completely disregarding their finishing times. All that is needed is the order in which males and females cross the finish line.

The reversals technique is easy to implement on a computer. All that is required is a sorting routine to order the data and a few lines of programming to count reversals and calculate $\hat{R}$ and W. As an example, to calculate and plot W for 155 kinds of measurements (155 sorts required) for class

populations $N_1 = 162$ and $N_2 = 150$, the total central processor time on a CDC 6600 computer was 10 seconds.

A number of other feature extraction techniques automatically incorporate correlations of the measurements, such as the methods of Fukunaga and Koontz[4] and others[5]. The methods rely on ordering the eigenvalues of matrices derived from the class correlation or covariance matrices. Another method[6] which also relies on the class covariance matrices, successively finds Fisher linear classifier projection directions, each time constraining the projection direction to be orthogonal to previously found projection directions. The methods result in a rotated multidimensional data space from which the unimportant dimensions can be neglected. The methods are powerful and practical for not too many variables. However, in the case of hundreds or even thousands of variables, finding eigenvalues and eigenvectors is computationally expensive and furthermore, the dimensionality must not exceed the number of elements in the smallest class. In such large dimension problems the reversals method can be very useful to initially reduce the number of dimensions to a manageable number.

A number of alternative measures of class separability exist such as the various "distance" measures[7]: Bhattacharyya distance, Kullback divergence distance, Kolmogorov distance, Matusita distance, etc. These distance measures require estimates of the class density distributions, which can be obtained from

histograms or Parzen estimates. The distance measures are powerful in that they can be applied to several random variables jointly. In practice the methods are limited by the problem of sorting the data into bins and smoothing the results. Different results will be obtained depending on the choice of bin size, origin for the bins, and smoothing kernel. Considering more than a few random variables jointly can become prohibitive in terms of the number of bins that must be stored. Furthermore most of the distance measures require integrals to be performed, all of which adds to the computational expense.

One example of comparing the Bhattacharyya distance for each of 25 kinds of measurement with the reversals measure W is shown in Fig. 7 for radar data of unknown distributions. In both cases, measurements numbered 4 - 6 are found to separate the classes best and measurements 1, 7, 13 separate the classes least. One shortcoming of the reversals method is highlighted in this example. The Bhattacharyya distance for measurement 10 is diminished significantly compared to that for measurements 4 - 6, whereas the reversal measure is not. The effect is most easily understood in the limiting case of perfect separability. If the distance between the distributions is altered, with perfect separation being maintained, the reversal count cannot change, whereas the Bhattacharyya distance measure detects the alteration.

Another distance measure that has been used is the value of K, defined as $2|m_2 - m_1|/(\sigma_1 + \sigma_2)$, where m_i and σ_i are the

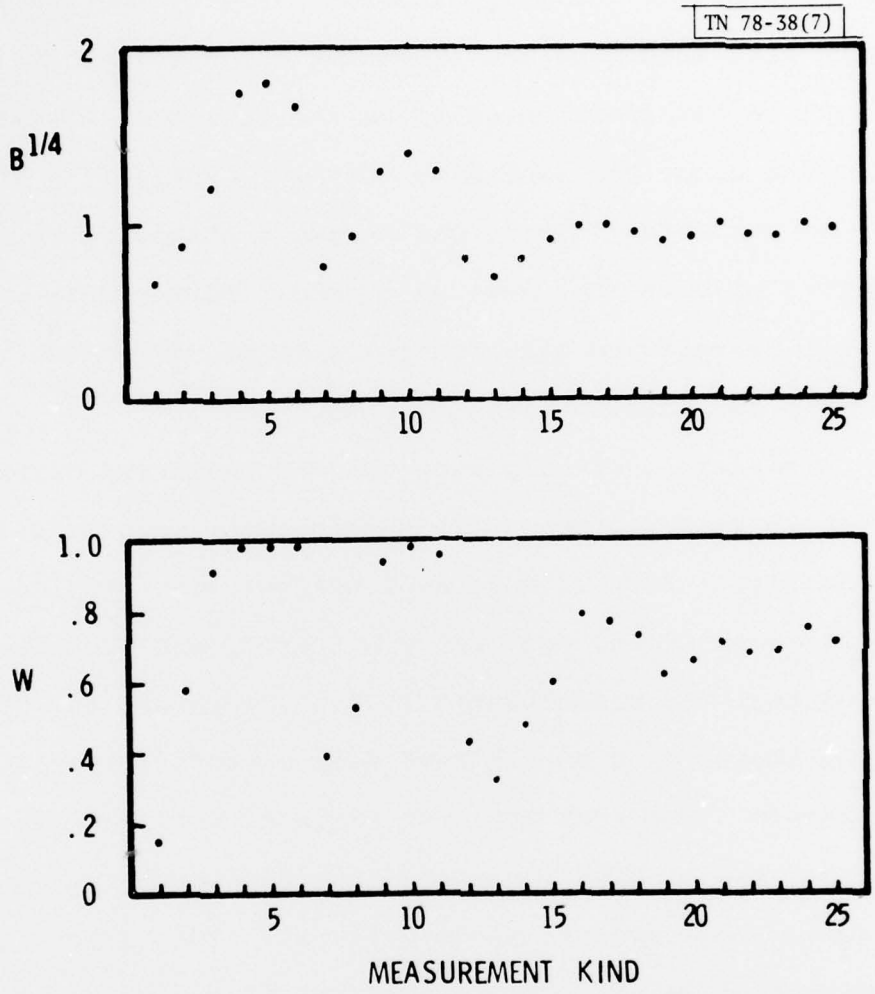


Fig.7. Comparison of the Bhattacharyya distance, B, with reversal measure, W, when applied to 25 different kinds of radar measurement, each having unknown distribution of the two classes. $N_1 = 83$, $N_2 = 102$.

sample mean and standard deviation for class i. This (parametric) measure is sensitive to differences in the mean, but shares the same disadvantages as the equal error measure, namely that it is blind to any difference in the classes when they have the same mean. Both the K measure and the equal error measure fail in the examples (B) and (C), where the standard deviation and skewness were the factors that separated the two classes. Another disadvantage of the technique is that the underlying true distribution is characterized by only the mean value and standard deviation. If there are multiple modes making up a class, so that the distribution is multi-humped and possibly asymmetric, the K measure treats the distribution in the same way as if it were single humped and symmetric. The reversal measure W requires no assumption on the number of modes and requires no estimate of the class mean and standard deviation and consequently is capable of detecting any differences in the two classes, whether due to differences in modes, symmetry, or otherwise.

Besides the distance measures, there is a whole class of non-parametric tests whose purpose is to answer the question, "Are the two classes <u>significantly</u> different?" Some of the more common non-parametric tests are the Kolmogorov-Smirnov test, the Wilcoxon rank sum test, the Mann-Whitney U test and the Wald-Wolfowitz runs test. A method of utilizing the tests as distance measures is to find for each random variable or type of

measurement the minimum level of significance at which the two classes cannot be considered identical for the given outcome of the test statistic. The measurement which then yields the smallest value of the minimum level of significance is considered best for separating the two classes. Such a method has been used by Day and Mullet [8,9], utilizing the runs test to determine which questions on an employment application form were most useful in predicting job longevity for firemen.

Unfortunately the probability density functions for the test statistics, which are for the hypothesis that the two classes are statistically identical, are generally quite broad, so that there is considerable fluctuation in the statistic outcome from one trial to another when the two classes are only slightly different. To illustrate this fluctuation, the results of 20 trials on identical classes for four non-parametric tests are compared with the reversal measure W. In each trial 500 random numbers were generated per class from the same normal distribution. The minimum level of significance, α_{min}, at which the hypothesis (for a one-sided test), that the classes are identical, must be rejected for the given outcome of the statistic was then found. This is just the area under the wing of the statistic probability density function measured outward from the given outcome of the statistic. The distance measure $D = 1-2\alpha_{min}$ which scales from D = 0 (believe that the classes are identical) to D = 1 (believe that the classes are completely separated), was

then found for each trial for each non-parametric test. The entries in the table below are the values of D for the non-parametric tests and the corresponding value of W.

Trial Number	Kolmogorov-Smirnov	Mann-Whitney	Wilcoxon rank sum	Wald-Wolfowitz	Reversals W
1	.06	.07	.46	.22	-.01
2	.77	.67	.83	.22	-.01
3	.95	.81	.91	.95	-.06
4	.02	.71	.15	.73	-.04
5	.54	.90	.05	.08	.00
6	.89	.79	.89	.18	-.01
7	.92	.83	.91	.67	.03
8	.33	.62	.19	.03	.00
9	.18	.26	.37	.27	-.01
10	.87	.56	.78	.53	-.03
11	.00	.75	.13	.67	-.03
12	.10	.85	.08	.98	-.07
13	.63	.35	.67	.27	.01
14	.49	.07	.46	.18	.01
15	.87	.28	.64	.03	.00
16	.77	.82	.91	.27	-.01
17	.02	.45	.27	.49	-.02
18	.23	.95	.02	.83	.04
19	.85	.48	.74	.53	.02
20	.94	.80	.90	.49	.02
Std. Dev. =	.37	.27	.34	.30	.03

The general result is that the standard non-parametric tests, when modified to be used as distance measures between classes, lack steadiness of results compared to the reversal measure W particularly when the classes are very similar. Also, note must be paid to the weaknesses of the particular non-parametric test.

Some tests are blind to differences in the class means and others to the class spreads.

One rather obvious point in applying any distance measure to the two class problem is that a single value of the measure does not illuminate the kind of class difference for that random variable. If we are told the value of W is 0.8, we know that the distributions are quite unlike, but we don't know if they are different in their means, spreads, skewness, number of modes, etc. A distance measure only quantifies the class difference and further study of the nature of the difference can be made from histograms of the class distributions.

In the discrimination problem once the dimensionality of the data space has been reduced to a manageable size, by whatever technique, then classifiers such as the Fisher linear classifier, the quadratic classifier, or others can be used on the remaining dimensions to classify the data. The classifiers will find the decision surface that divides the reduced dimension data space in some optimal fashion, making use of, among other things, correlations between variables that may not have been examined by the user in the initial dimension reduction. Consequently one would generally wish to save the maximum number of feasible dimensions.

Besides the aspect of reducing the dimensionality of a problem, the reversals method is useful for illuminating where

differences are between two classes. The interpretation of class differences is relatively easy in the reversals technique compared to techniques that rotate a multidimensional data space. Although the latter techniques may be optimal in some sense*, the interpretation of how the new rotated axes (features) are important in terms of the original observables can be difficult. In this sense the reversals technique is a powerful tool; the user is in full control of choosing the variables (individual and combination measurement types) to isolate the class differences, one variable at a time. The danger, of course, is that the user may neglect to examine some combination of measurement types that is, in fact, good for discrimination. The situation may be likened to probing an oracle with questions on some problem. It may happen that greater insight is achieved by asking many small questions, each answer to which we understand, than to ask one ultimate question whose answer is baffling.

---

*As one example, the Fukunaga-Koontz method is optimal in the sense that it selects features that most typify one class and simultaneously least typify the other class. That the technique is not necessarily a good technique has been shown by a counter-example in reference 6, where in a 3-dimensional problem the technique selected the 2 worst dimensions for discrimination.

ACKNOWLEDGMENTS

B. Llorens contributed the Bhattacharyya distance data used in Fig. 7 and made many helpful suggestions. Others who helped through their discussions are J. R. Delaney, K-P. Dunn, G. Q. McDowell and J. A. Tabaczynski. Computer programming assistance was given by R. M. Kelner, M. E. DeCoste and G. S. Freedman.

REFERENCES

1. W. L. Stevens, "Distribution of Groups in a Sequence of Alternatives," Annals of Eugenics 9, 10 (1939).

2. A Wald and J. Wolfowitz, "On a Test Whether Two Samples are from the Same Population", Ann. Math. Stat. 11, 147 (1940).

3. Almost any textbook with a comprehensive section on non-parametric statistics, for example, S.S. Wilks, Mathematical Statistics (Wiley, New York, 1962), pp. 145-9.

4. K. Fukunaga and W. Koontz, "Application of the Karhunen - Loeve Expansion to Feature Selection and Ordering," IEEE Trans. Computers C-19, 311 (1970).

5. Eleven feature selection techniques are reviewed in J. Kittler, "Feature Selection Methods based on the Karhunen - Loeve Expansion" in Pattern Recognition Theory and Application, eds. K. S. Fu and A. B. Whinston (Noordhoff, Leyden, Netherlands 1977), pp. 61-74.

6. D. H. Foley and J. W. Sammon, Jr., "An Optimal Set of Discriminant Vectors," IEEE Trans. Computers C-24, 281 (1975).

7. C. Chen, Statistical Pattern Recognition (Spartan Books, Rochelle Park, New Jersey 1973), pp. 57-60

8. G. J. Day and G. M. Mullet, "Discriminant Analysis Using Qualitative Predictor Variables," Report M-75-4, College of Industrial Management, Georgia Institute of Technology (1975).

9. G. M. Mullet, "Process Analysis Using the Runs Test," Jour. of Quality Technology 9, 1-5 (1977).

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER ESD-TR-78-296	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE *(and Subtitle)* A Method for Identifying Differences Between Two Classes		5. TYPE OF REPORT & PERIOD COVERED Technical Note
		6. PERFORMING ORG. REPORT NUMBER Technical Note 1978-38
7. AUTHOR(s) Nathan E. Lindgren		8. CONTRACT OR GRANT NUMBER(s) F19628-78-C-0002
9. PERFORMING ORGANIZATION NAME AND ADDRESS Lincoln Laboratory, M.I.T. P.O. Box 73 Lexington, MA 02173		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS Project No. 8X363304D215
11. CONTROLLING OFFICE NAME AND ADDRESS Ballistic Missile Defense Program Office Department of the Army 5001 Eisenhower Avenue Alexandria, VA 22333		12. REPORT DATE 29 November 1978 13. NUMBER OF PAGES 36
14. MONITORING AGENCY NAME & ADDRESS *(if different from Controlling Office)* Electronic Systems Division Hanscom AFB Bedford, MA 01731		15. SECURITY CLASS. *(of this report)* Unclassified 15a. DECLASSIFICATION DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT *(of this Report)*

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT *(of the abstract entered in Block 20, if different from Report)*

18. SUPPLEMENTARY NOTES

None

19. KEY WORDS *(Continue on reverse side if necessary and identify by block number)*

discrimination-feature selection problem
feature extraction
non-parametric measure
method of reversals
runs
pattern recognition

20. ABSTRACT *(Continue on reverse side if necessary and identify by block number)*

A non-parametric measure of the difference between sample distributions of a random variable for two classes of data is presented. The method involves counting the number of class reversals among the ordered set of two class data and provides a normalized measure of class intermingling. Applications of the method to the discrimination-feature selection problem are described.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

207 650