LARS Information Note 040174

SIMULATION TECHNIQUES FOR ESTIMATING ERROR IN THE CLASSIFICATION OF NORMAL PATTERNS

STEPHEN J. WHITSITT
DAVID A. LANDGREBE

The Laboratory for Applications of Remote Sensing

Purdue University, West Lafayette, Indiana

1974

LARS Information Note 040174

SIMULATION TECHNIQUES FOR ESTIMATING ERROR IN THE CLASSIFICATION OF NORMAL PATTERNS*

Stephen J. Whitsitt
David A. Landgrebe

Abstract

Methods of efficiently generating and classifying samples with specified multivariate normal distributions are discussed. Conservative confidence tables for sample sizes are given for selective sampling. Simulation results are compared with classified training data. Techniques for comparing error and separability measures for two normal patterns are investigated and used to display the relationship between error and the Chernoff bound.

*The work described in this paper was supported by National Aeronautics and Space Administration Grant No. NGL 15-005-112.

INTRODUCTION

There has been a significant amount of effort devoted to the design and evaluation of functions which "measure" the relative effectiveness of statistical pattern recognition schemes in classifying data. Two of the more notable ones are the Bhattacharyya distance [5] (a special case of distribution pairs of the subsequent Chernoff bound [6, 16, pp. 116-126]), and the divergence [7,8]. The motivation for these "distance measures" is that in some cases, theoretical recognition error cannot be obtained easily. In the case of the normal assumption, the error expression is generally difficult if not impossible to evaluate analytically. A technique [9,10] has been developed for obtaining theoretical error in a two-class problem using a Bayes decision rule and gaussian assumption. But error in recognition problems with an arbitrary number of normal classes has not in general been expressed in a manner which can be analyzed easily.

Because of this problem, "distance" measures and bounds have great appeal. In multiple-class problems, some sort of average of the distance between pairs of classes often is used as a performance measure of various classification schemes (such as selecting feature sets).

The ability of a separability measure to predict performance in statistical pattern recognition ultimately depends on its relationship with theoretical error. Some relationships between error and the Bhattacharyya distance and divergence are known [11,12,13,14]. These relationships are in the form of bounds on error. For the two cited separability measures, the most important relationship is that two-class error is bounded by one-half of the Bhattacharyya coefficient [12], and accuracy (one minus error) for two normal classes appears to be bounded above and below by an empirical relationship with the divergence described in [15]. From this empirical relationship, it appears that probability of correct recognition is less than or equal to the value of the normal distribution function at one-half of the square root of the divergence. That is

$$P_c \leq \mathrm{erf}_*(\sqrt{D}/2) \ , \qquad (1)$$

although this has not been proven yet.

It is interesting to note that in [15], the paper to which much of the motivation for use of divergence has been attributed, part of the relationship between divergence and accuracy was obtained using a Monte-Carlo type of simulation. It seems apparent, in looking over some of the literature dealing with these and other error bounds, that a simulation type of analysis would have something to offer in understanding the relationship between error and these bounds.

Many advances in the use of error bounds have improved (at a cost of high mathematical complexity in some cases) error prediction in very specific areas (for instance, the use of the Chernoff bound in information theory and likelihood decoding error analysis [17, pp. 131-135; 18, pp. 394-398; 19]). In the case of two gaussian distributions, one of the tightest known bounds on error which can be easily evaluated, the Chernoff bound, is "close" in predicting error for only special cases (such as in [16, pp. 126-133]). For more general two-class problems such as the one used in [15], an example in [10, p. 73] shows a case where this bound does insignificantly better than the Bhattacharyya coefficient (tightest known bound for normal data which can be expressed explicitly), which, in this case, isn't very close to actual error. Experience has shown that this is often the case in data from natural patterns such as multi-spectral data [1] modeled by the normal distribution.

In many problems, however, it is not so important that a distance measure bound error, as it is that it should tend to indicate which classification scheme is best (not necessarily the same thing). This is especially important in the case of multiple-hypothesis pattern recognition, because even the tightest bounds lose most of their "potency" when they are averaged over all pairs of classes [20]. Also, measures which aren't averages over class pairs have yet to yield any analytic simplicity [2]. However, if one

separability measure has a weaker relationship with theoretical error than another, it must be considered as a less reliable source of separability information.

Simulation can provide a useful relationship between specific classification problems and the numbers produced by separability measures. For instance, the average divergence might be used to narrow a large number of feature sets down to several which have the highest value. Then, rather than classify the training samples using these feature sets and compare (especially if this is physically cumbersome), one might generate and classify samples with the same distribution as the training classes. Or, it may be the case that a researcher requires easy access to a large number of samples with a specific distribution in order to make a carefully controlled comparison of classification error and separability measures.

The major disadvantage, when compared to most separability measures, is the amount of machine time used to classify the samples. Also, the method is Monte-Carlo and not exact. Hence the degree of confidence varies with the number of samples used. These two drawbacks will be examined in this note.* Also, certain properties of pairs of normal

---

*In a forthcoming paper a new statistic for error will be introduced for cases where distributions are specified.

patterns are used to reduce the size of the sample space of mean vectors in case the relationship between recognition rate and other pair-wise separability measures is to be studied. Examples of all of the techniques are presented. Much of the material is tutorial in nature, but provides a necessary background for the methods described.

A THEORETICAL BASIS FOR SIMULATION AND CONFIDENCE BOUNDS FOR THE RESULTS

If one has available samples from the mixture density, a method of estimating error in using a decision rule which partitions the sample space is well known [4]. This method, random sampling or error counting, does not give estimates of conditional class error. However, precise confidence tables are available [22,4,10, p. 147] for computing sample size. Another method known as selective [4] or stratified [3, p. 255] sampling does yield these estimates and has an estimate for error with smaller variance than random sampling. Some conservative confidence tables are now developed for selective sampling. No assumption of class distributions is made.

Suppose that one has N_i samples from class i, and that the classification scheme under consideration classifies L_i of these samples correctly. Let P_{ci} be the conditional probability of correct classification for class i using this scheme. Since L_i is binomial with parameters N_i and P_{ci}, it

is well known that the maximum likelihood estimate $\hat{P}_{ci}$ for P_{ci} is

$$\hat{P}_{ci} = \frac{L_i}{N_i} \quad , \tag{2}$$

unbiased. Further, suppose that there are M classes in this particular example, and that $N_i = P_i N$, where P_i is the a priori probability of class i, so that a total of N samples are used. Then the maximum likelihood estimate (see [10, pp. 145-148] and [27, pp. 47-48]) for overall theoretical error $P_c = \Sigma P_i P_{ci}$ is

$$\hat{P}_c = \sum_{i=1}^{M} P_i \hat{P}_{ci} = \frac{1}{N} \sum_{i=1}^{M} L_i \quad , \tag{3}$$

unbiased. The absolute error in $\hat{P}_c$ is $|\hat{P}_c - P_c|$, and its variance is $\Sigma P_i P_{ci}(1 - P_{ci})/N$ [10]. Using a basic inequality of probability theory [21, p. 157], it can be shown that for any $\delta > 0$,

$$P\{|\hat{P}_c - P_c| \geq \delta\} \leq \frac{\Sigma P_i P_{ci}(1 - P_{ci})}{N \delta^2} = B_1 \tag{4}$$

(all summations from 1 to M). That is, the probability that the estimated overall error $\hat{P}_c$ differs from the actual overall error P_c by more than δ is bounded by B_1. But note that B_1 depends on the individual P_{ci}. If these were known, P_c could be computed exactly.

A confidence bound with no dependence on the individual P_{ci} may be easily obtained by noting that

$$\max_{P_{ci}} [\Sigma P_i P_{ci}(1 - P_{ci})] = \frac{1}{4} \quad . \tag{5}$$

Hence

$$P\{|\hat{P}_c - P_c| \geq \delta\} \leq B_1 \leq \frac{1}{4N\delta^2} = B_2 \quad . \tag{6}$$

So we use $N = 1/(4B_2\delta^2)$.

As an example, suppose that for a ten class problem, it is desired that the error in the estimate be greater than 0.01 not more than 5% of the time. This corresponds to a 95% confidence that $\hat{P}_c$ is within 0.01 of P_c. For $B_2 = 0.05$, $M = 10$, equal priors P_i, and $\delta = 0.01$, we find $N = 50{,}000$ (5000 per class) samples required. For comparison, the assumption that $P_{ci} = .8$ and use of B_1 would result in the requirement that 32,000 samples be used.

It might be noted that some similarity exists between this confidence expression and the classic confidence tables of [22] for random sampling. In the case of the latter, however, it is known that the distribution of the error in the estimate is binomial. This allows one to construct a much tighter confidence interval (or looking at it another way, use fewer samples). $|\hat{P}_c - P_c|$ is in general binomial only for $M = 2$ in this paper. Further, the confidence of $1 - B_2$ (which is $\geq P\{|P_c - \hat{P}_c| \leq \delta\}$) corresponds to the interval

$$P_c - \frac{1}{2\sqrt{NB_2}} \leq \hat{P}_c \leq P_c + \frac{1}{2\sqrt{NB_2}} \qquad (7)$$

In the classic confidence tables, these intervals are not symmetric unless $P_c = 0.5$. As an example, let $P_c = 0.5$. For M = 2 and 95% confidence that the error does not exceed .05 ($\hat{P}_c$ within .05 of P_c at least 95% of the time), we require 2000 samples using B_2 (and B_1), while only about 400 samples are required using the knowledge that N_i is binomial.

A graph of error δ versus the total number of samples N is presented in Figure 1 for confidence levels of 75, 90, 95, and 99%. A log-log scale is used in order to present a useful range of values. Because of the conservative nature of the bound, modest choices of δ and confidence level may lead to large sample sizes. In fact the 95% confidence line for random sampling with $P_c = 0.5$ would lie just above the 75% line in Figure 1, even though the variance of the selective sampling statistic is, in general, smaller. However, if one needs the estimates $\hat{P}_{ci}$, the latter statistic is more convenient (one may always use the tables of [22] to compute confidence in the individual P_{ci}), and does not require randomization on the class numbers.

When sample sizes are large, an approximation may be used. For fixed M and increasing N, the distribution function of $\hat{P}_c$ tends to become normal regardless of the values of the P_{ci} [21,29, pp. 256-257]. The N+1 discontinuities in

the distribution become small "jumps." The confidence approximation using (5) becomes

$$P[|\hat{P}_c - P_c| \geq \frac{Z_{\alpha/2}}{2\sqrt{N}}] \approx 1 - \alpha = B_3 \qquad (8)$$

where 100α is the percent confidence and $Z_{\alpha/2}$ is value at which the normal distribution function is $1 - \alpha/2$. Now we use $N = Z^2_{\alpha/2}/(4\delta^2)$. Figure 2 gives the resulting relationship for 75, 90, 95 and 99% confidence. In the example above for M = 10, we find that B_3 yields 9600 samples required (B_1 32,000; B_2 50,000). In the other example for M = 2, we get 385 (B_1, B_2 2000; binomial 400). The latter example points out the need for large sample sizes in using B_3. If M is increased, even larger sizes are probably needed.

EFFICIENT GENERATION AND CLASSIFICATION OF NORMAL SAMPLES

Let us assume that a source of independent, normally distributed samples is available. Such a source can be approximated by using a power-residue technique to generate pseudo-random samples with approximately uniform distribution. Sets of these samples may then be normalized in accordance with the central-limit theorem to produce approximately normal samples. One of the most commonly used techniques employing this procedure is described in [23, pp. 94-96] (this reference describes the theoretical basis for the algorithm used on IBM/360 computers in the SSP subroutine

RANDU). Samples generated using this method have little sample correlation [24]. Another well known method is the inverse method. It is faster than the above (using typical set sizes), requires only one uniformly distributed sample, and, for all practical purposes, is not truncated. Let the random variable X be uniformly distributed on the interval from zero to one. Let $F(\cdot)$ represent the desired distribution with inverse $F^{-1}(\cdot)$. Then $Y = F^{-1}(X)$ has distribution $F(\cdot)$. For F normal, good approximations are available [26, pp. 191-192; see SSP subroutine NDTRI]. This reference [26] is the reason the method for normal F is sometimes called Hastings method. Other fast procedures are given in [27, pp. 90-95].

Designate a normal density for class i with n by 1 mean vector M_i and n by n covariance matrix K_i as $N(M_i,K_i)$. Let Q_i be an orthogonal transformation which diagonalizes K_i as

$$Q_i^t K_i Q_i = \Lambda_i = \begin{pmatrix} \lambda_1 & & & 0 \\ & \lambda_2 & & \\ & & \ddots & \\ 0 & & & \lambda_n \end{pmatrix} \tag{9}$$

where Λ_i is the n by n diagonal matrix of eigenvalues λ for K_i (so that Q_i is a matrix with eigenvectors of K_i for its columns [25, pp. 80-99]). Form n by 1 random vectors X with density $N(\underline{0},I)$ by taking n normal samples with zero mean and unit variance and use them for the components of X. If

we define

$$\Lambda_i^{1/2} = \begin{pmatrix} \sqrt{\lambda_1} & & & 0 \\ & \sqrt{\lambda_2} & & \\ & & \ddots & \\ 0 & & & \sqrt{\lambda_n} \end{pmatrix}, \qquad (10)$$

then the random vector $Y_i = Q_i \Lambda_i^{1/2} X + M_i$ is $N(M_i, K_i)$.

These steps may be graphically depicted as the illustration in Figure 3 shows. Cross-sections (level surfaces or surfaces of constant probability) of densities for n=2 are shown which are in this case, ellipses. Figure 3a represents the desired distribution. $\Lambda_i^{1/2}$ scales the samples to obtain variances which correspond to those of the principal components of the original covariance matrix. Q_i rotates the samples (or coordinate system) until the principal components of the density are parallel to those of the desired density (same correlation between features). Adding M_i locates the mean at the desired value.

This method is quite straightforward, but very time consuming. Each vector X must be multiplied by the matrix $Q_i \Lambda_i^{1/2}$ for a total of n^2 multiplications and n^2 additions <u>for each X</u>. Adding M_i to each X could be eliminated by shifting all of the distributions by that amount. This also raises the possibility of classifying the X samples directly in a

transformed feature space rather than transforming them first and classifying in the original space. Since classification error is invariant under linear transformations and shifts, and $\hat{P}_{ci}$, $\hat{P}_c$ are the only desired results, we note that

$$X = \Lambda_i^{-1/2} Q_i^t (Y_i - M_i) \tag{11}$$

is $N(\underline{0},I)$ and transform all of the other class parameters as (see Appendix A)

$$K_j^* = \Lambda_i^{-1/2} Q_i^t \, K_j Q_i \Lambda_i^{-1/2} \tag{12}$$

$$M_j^* = \Lambda_i^{-1/2} Q_i^t (M_j - M_i) \quad . \tag{13}$$

Thus we can use $N(\underline{0},I)$ samples directly to represent class i and obtain $\hat{P}_{ci}$ by classifying these samples using the above expressions for the other covariance matrices and mean vectors. This process can be characterized as a transformation of the feature space to fit the samples, rather than a transformation of the samples to fit the feature (although the two are equivalent). In other words, the feature space is transformed in a manner analogous to going backwards in Figure 3 from 3e to 3b.

It might also be noted that the normalizing process used to obtain $N(\underline{0},I)$ samples from uniformly distributed samples could be incorporated into this procedure to eliminate more unnecessary computations (e.g., don't normalize).

Also, $\log|K_j^*|$ (which is used in the decision rule) is just $\log|K_j| - \log|K_i|$, so that these need be computed only once for the entire simulation.

Applications for Normal Data

The normal assumption appears to work reasonably well in classified designs when applied to agricultural categories of multispectral data [1]. Recently, a powerful test of normality was developed and used on this data, the results of which lead one to believe that in some cases, the assumption is not unreasonable [28]. Using the same data with classes defined in [1], an experiment was conducted to compare the results of estimating P_c by simulation with the value obtained by classifying training samples, using statistics* obtained from those samples. Eight classes (corn, soybeans, wheat, alfalfa, bare soil, oats, clover, rye) were used with 12 features (wavelength bands). One thousand samples per class were generated by the methods described above for each of the feature sets {1}, {1,2},...{1,2,...,12}. The results for estimating overall error, $P_e = 1 - P_c$, and conditional error, $P_{ei} = 1 - P_{ci}$ for the class wheat, are given in Figure 4, a and b. Agreement seems to be fairly good, with simulation results appearing more optimistic in terms of accuracy, as might be expected (the gencrated data should fit the normal assumption better).

***maximum likelihood for mean and covariance with bias correction applied to the latter.**

TWO CLASS PROBLEMS

Simulation studies of two class separability measures for certain types of distributions may yield useful information for classifier design. An example is given in [15], where recognition rate is compared to divergence values for two normal patterns. Knowledge of the behavior of such measures may allow the researcher to define new measures for M class problems which improve performance in feature selection.*

For normal patterns, it is well known that both covariance matrices may be simultaneously diagonalized, one into the identity matrix. Then the transformed means of these classes may be shifted so that the class with identity covariance has its mean at the origin. Thus, all cases of pairs of normal patterns may be simulated by considering only classes with diagonal covariance matrices, one equal to the identity with zero mean vector. In this case computation of separability measures such as the Chernoff and Bhattacharyya bound, divergence, and even true error are relatively straightforward ([10, pp. 72, 284,62**-64] respectively). One need generate values for the parameters

---

*A forthcoming paper will explore this topic.

**Changing the sign in Equation 3-51 and 3-52 from + to -.

of the class with arbitary mean vector only.

The major problem is the amount of samples from the parameter space (of mean and variance components) needed to obtain representative results. Obvious symmetry (in the sense of error) allows the use of only non-negative mean components. Yet another type of symmetry exists. We see from Figure 5a that there is reflective symmetry about lines of equal mean components. Here a two-feature example is sketched to show that for every set of mean and variance components chosen in the subset of non-negative mean components, a simple permutation of these component values yields a different distribution with the same error, still contained in this subset. Proceeding to the general case of n features, it is apparent that this property yields the requirement that only mean vectors with monotone components are required. Since there are 2^n combinations of signs for the components of an arbitrarily chosen mean vector, and because the restriction to positive signs leaves n! choices of inequalities between components (fix m_1 on the real line, leaving two places for m_2, three for m_3, etc.), the restriction of, say, $m_1 \geq m_2 \geq \ldots > m_n$ reduces the size of the set of possible mean vectors with components restricted in magnitude by a factor of $1/(n!2^n)$. Figure 5b depicts this process for n=3 and $a \geq m_1 \geq m_2 \geq m_3 \geq 0$.

The method is readily applied to experiments where an attempt to pre-determine covariance and mean values is desired. These values may be incremented by a fixed amount

over a range of numbers, so as to insure that representative combinations are covered (one objection to a random approach). Generating random components is a bit more difficult if one desires a uniform distribution on the set of possible mean components. This would involve more complicated software to compute the assignment of probability mass for successive mean components conditioned on the value of a previous one. Experience has shown that order statistics or random walks (m_i uniform on m_{i+1} to a) give satisfactory results.

As an example, 40,000 sets of parameters were generated, 1,000 each for sets of 2, 3, and 4 components, and 37,000 for one component (due to time considerations in computing error for n>1), and both P_c and the Chernoff bound were computed. The result is given in Figure 6. Order statistics for uniformly distributed random numbers on the interval from 0.0 to 6.0 were used to obtain mean components. Variance values were obtained from numbers uniform on .01 to 25.0. P_c was computed using the method of [10].

One interesting possibility raised by the above example is that a relationship between the Chernoff distance C (minus the log of the coefficient) and P_c, similar to that of the divergence, may exist. For equal covariance matrices,

$$P_c = \text{erf}_* (\sqrt{2C}) \qquad (14)$$

Plotting the right hand side of (14) with P_c yields Figure 7, suggesting that

$$P_c \leq \text{erf}_* \ (\sqrt{2C}) \quad , \tag{15}$$

but which has not been proven. A check of the numbers generated has thus far established empirical agreement with (15).

Summary and Conclusion

Motivation for the use of Monte-Carlo type simulation in the study of classifier design includes avoiding the difficulty in obtaining error exactly, and the desire to obtain relationships between error and separability measures for various classes of density functions. Selective sampling was reviewed and conservative confidence bounds for sample sizes developed. The confidence relationships are weaker than those for random sampling. However, random sampling does not provide controlled size estimates of conditional class errors. Methods of generating and classifying normal data were discussed and an example representing classification of multispectral agricultural data was given. For studies of pair-wise separability measures involving normal patterns, methods of selecting statistical parameters efficiently were given. An example depicting the relationship between the Chernoff bound and correct recognition was presented. The results suggest the possibility of the existence of a tight lower bound on error in terms of the Chernoff distance for normal patterns.

BIBLIOGRAPHY

1. K. Fu, D. Landgrebe, and S. Phillips, "Information Processing of Remotely Sensed Data," Proceedings of the IEEE, April, 1969, p. 639.

2. P. Devijver, "On a New Class of Bayes Risk in Muli-hypothesis Pattern Recognition," IEEE Transactions on Computers, C-23, January, 1974, pp. 70-80.

3. H. Cramer, The Elements of Probability Theory and Some of Its Applications, Wiley, N.Y., 1955.

4. W. Highleyman, "The Design and Analysis of Pattern Recognition Experiments," Bell System Technical Journal, V. 41, pp. 723-744.

5. A. Bhattacharyya, "On a Measure of Divergence between Two Statistical Populations Defined by their Probability Distributions," Bulletin of the Calcutta Mathematical Society, V. 35, No. 3, pp. 99-110, September, 1943.

6. H. Chernoff, "A Measure of Asymptotic Efficiency for Tests Based on a Sum of Observations," Annals of Mathematical Statistics, 1952, V. 23, pp. 493-507.

7. A. Kullback, Information Theory and Statistics, Dover, N.Y., 1960.

8. H. Jeffreys, Theory of Probability, Oxford University Press, 1948, p. 158.

9. K. Fukanaga and T. Krile, "Calculation of Recognition Error for Two Multivariate-Gaussian Distributions," IEEE Transactions on Computers, March, 1969, pp. 220-229.

10. K. Fukanaga, Introduction to Statistical Pattern Recognition, Academic Press, N.Y., 1972.

11. D. Lainiotis, "A Class of Upper Bounds on Probability of Error for Multihypothesis Pattern Recognition," IEEE Transactions on Information Theory, November, 1969, pp. 730-731.

12. T. Kailath, "The Divergence and Bhattacharyya Distance Measures in Signal Selection," IEEE Transactions on Comm. Tech., February, 1967, pp. 52-60.

13. G. Toussaint, "Comments on 'The Divergence and Bhattacharyya Distance in Signal Selection,'" IEEE Transactions on Comm., June 1972, p. 485.

14. D. Lainiotis, "On a General Relationship between Estimation, Detection, and the Bhattacharyya Coefficient," IEEE Transactions on Information Theory, July, 1969, pp. 504-505.

15. J. Marill and D. Green, "On the Effectiveness of Receptors in Recognition Systems," IEEE Transactions on Information Theory, January, 1963, pp. 11-17.

16. H. Van Trees, Detection, Estimation, and Modulation Theory, Part I, Wiley, N.Y., 1968.

17. R. Gallager, Information Theory and Reliable Communication, Wiley, N.Y., 1968.

18. J. Wozencraft and J. Jacobs, Principles of Communications Engineering, Wiley, N.Y., 1965.

19. C. Shannon, R. Gallager, E. Berlekamp, "Lower Bounds to Error Probability for Coding on Discrete Memoryless Channels: I and II," Information Theory and Control, V. 10, 1967, pp. 65-103 and 522-552.

20. S. Whitsitt, "Predicting Performance in Multiclass Statistical Pattern Recognition," Ph.D. Thesis Proposal, Department of Electrical Engineering, Purdue University, 1972.

21. M. Loeve, Probability Theory, Van Nostrand, N.J., 1955.

22. C. Clopper and E. Pearson, "The Use of a Confidence, or Fiducial Limits Illustrated in the Case of the Binomial," Biometrika, V. 26, 1934, pp. 404-413.

23. G. Gordon, System Simulation, Prentice-Hall, Englewood Cliffs, N.J., 1969.

24. S. Whitsitt, "Random Noise in Multispectral Classification," LARS Information Note No. 102670, Purdue University, 1970.

25. J. Indritz, Methods in Analysis, Macmillan, N.Y., 1963.

26. C. Hastings, Approximations for Digital Computers, Princeton University Press, Princeton, N.J., 1955.

27. T. Naylor, J. Balintfy, D. Burdick, K. Chu, Computer Simulation Techniques, Wiley, N.Y., 1966.

28. D. Kessell, "Error Evaluation and Model Validation," Purdue University, Department of Electrical Technical Report, TREE 72-23, August, 1972.

29. W. Feller, An Introduction to Probability Theory and Its Applications, VII, Wiley, N.Y., 1966.

APPENDIX A

Let Z be $N(M_j, K_j)$. Then $X = \Lambda_i^{-1/2} Q_i^t (Z - M_i)$ has mean vector

$$M_j^* = E[\Lambda_i^{-1/2} Q_i^t (Z - M_i)] = \Lambda_i^{-1/2} Q_i^t (EZ - M_i)$$

$$= \Lambda_i^{-1/2} Q_i^t (M_j - M_i) \qquad \text{(A1)}$$

and covariance matrix

$$K_j^* = E[\Lambda_i^{-1/2} Q_i^t (Z - M_i) - \Lambda_i^{-1/2} Q_i^t (M_j - M_i)][\text{same}]^t$$

$$= \Lambda_i^{-1/2} Q_i^t [E(Z - M_j)(Z - M_j)^t] Q_i \Lambda_i^{-1/2}$$

$$= \Lambda_i^{-1/2} Q_i^t K_j Q_i \Lambda_i^{-1/2} \qquad \text{(A2)}$$

Thus classifying $Z \sim N(M_j, K_j)$ is equivalent to classifying $X \sim N[\Lambda_i^{-1/2} Q_i^t (M_j - M_i), \Lambda_i^{-1/2} Q_i^t K_j Q_i \Lambda_i^{-1/2}] = N(M_j^*, K_j^*)$

which for class i is $N(\underline{O}, I)$. In fact if we define the discriminant for class j at X as

$$g(X) = C_j + \log|K_j^*| + (X - M_j^*)^t K_j^{*-1} (X - M_j^*) \qquad \text{(A3)}$$

where C_j is the cost and a priori probability constant, we find that substitution of (A1) and (A2) yields

$$g(X) = C_j + \log|K_j| - \log|K_i| + (Z - M_j)^t K_j^{-1}(Z - M_j)$$

$$= g(Z) - \log|K_i| \quad . \tag{A4}$$

Thus the discriminant values differ by a constant.

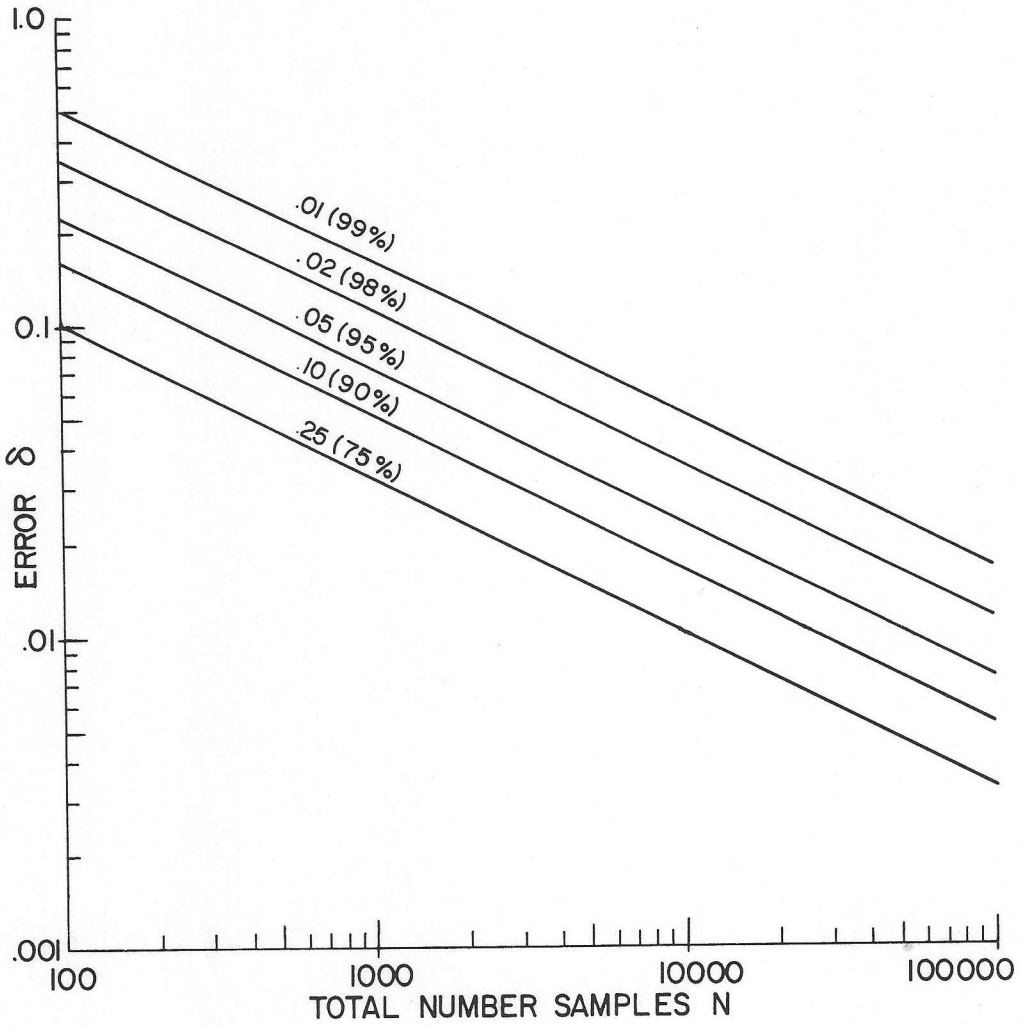


Figure 1: Conservative Confidence Values for Selective Sampling.

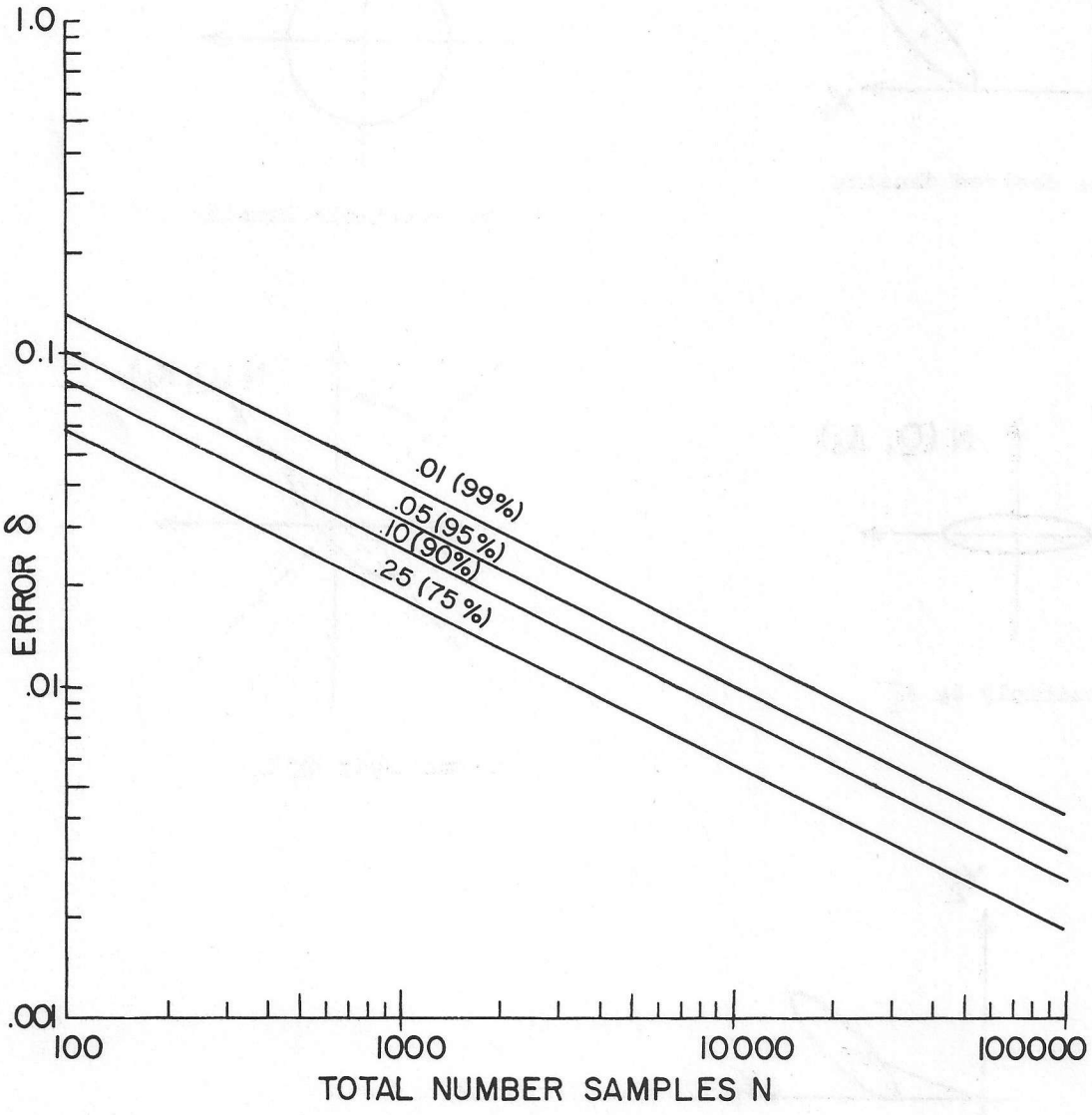
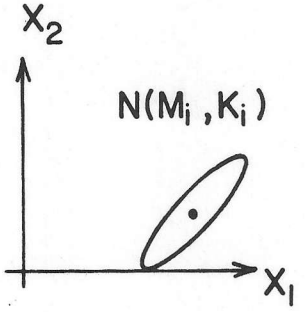
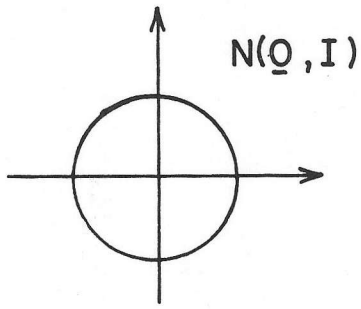


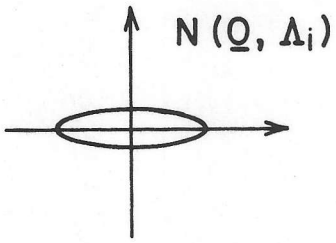
Figure 2: Confidence Values for Selective Sampling Using the Normal Assumption.



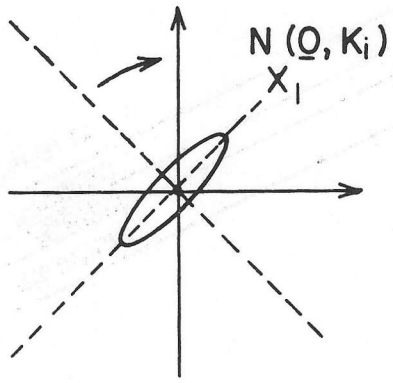
a) desired density



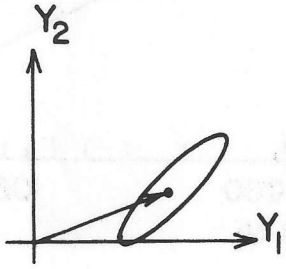
b) available density



c) multiply by $\Lambda_i^{\frac{1}{2}}$



d) multiply by Q_i



e) add M_i

Figure 3: Feature Transformation.

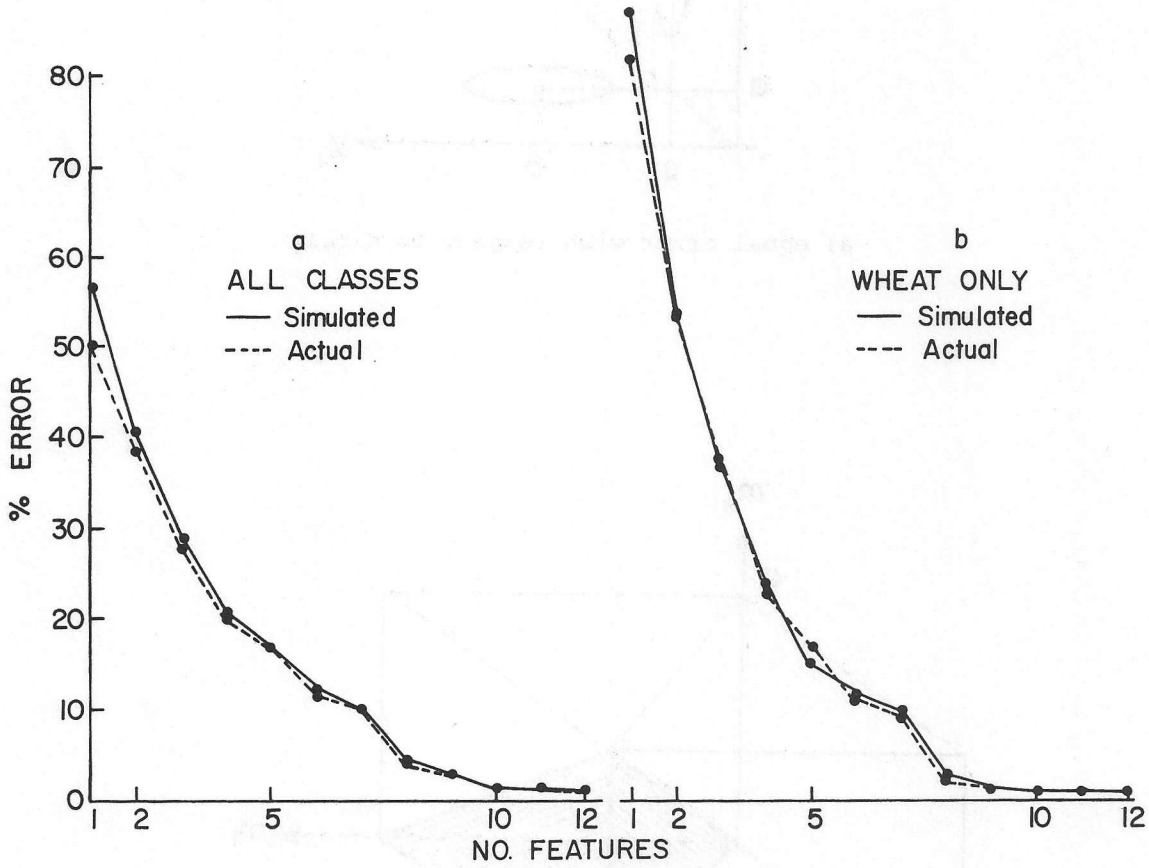
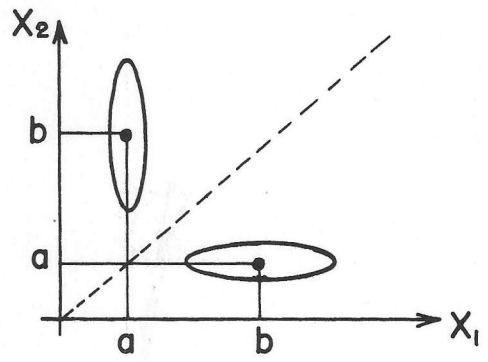
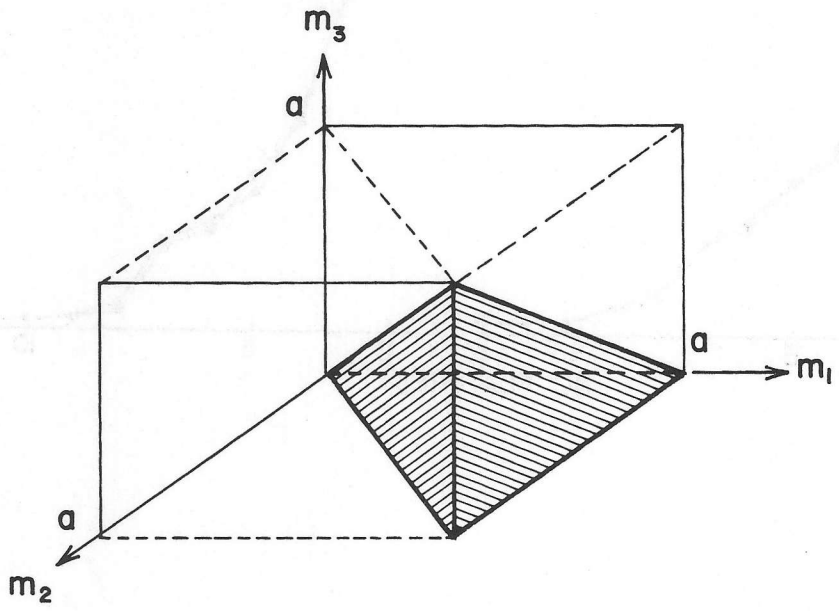


Figure 4: Simulation Results for Normal Data.



a) equal error with respect to $N(\underline{0},I)$



b) $a \geq m_1 \geq m_2 \geq m_3 \geq 0$

Figure 5: Mean Vector Space Reduction.

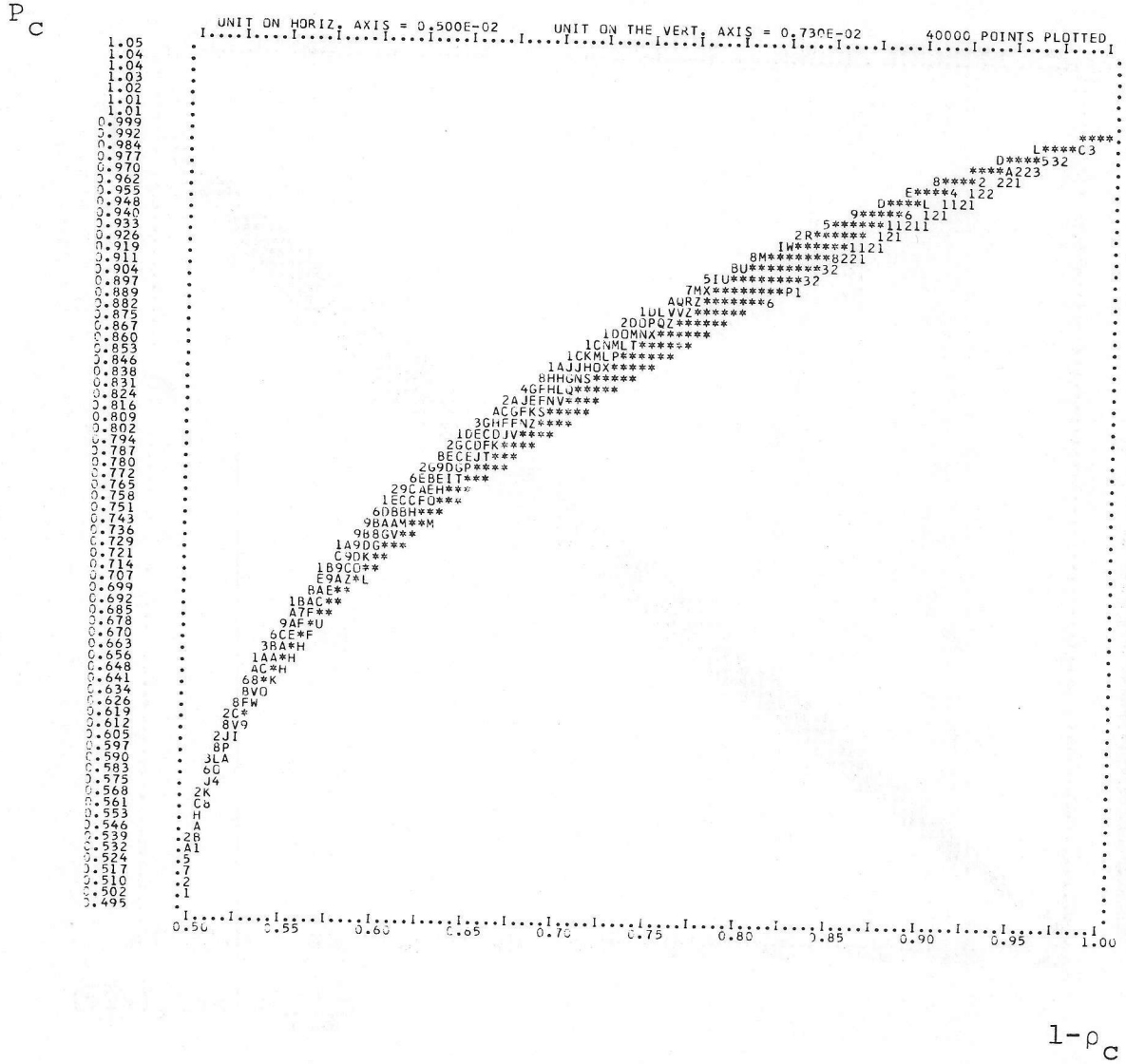


Figure 6. Probability of Correct Classification versus Chernoff Bound.

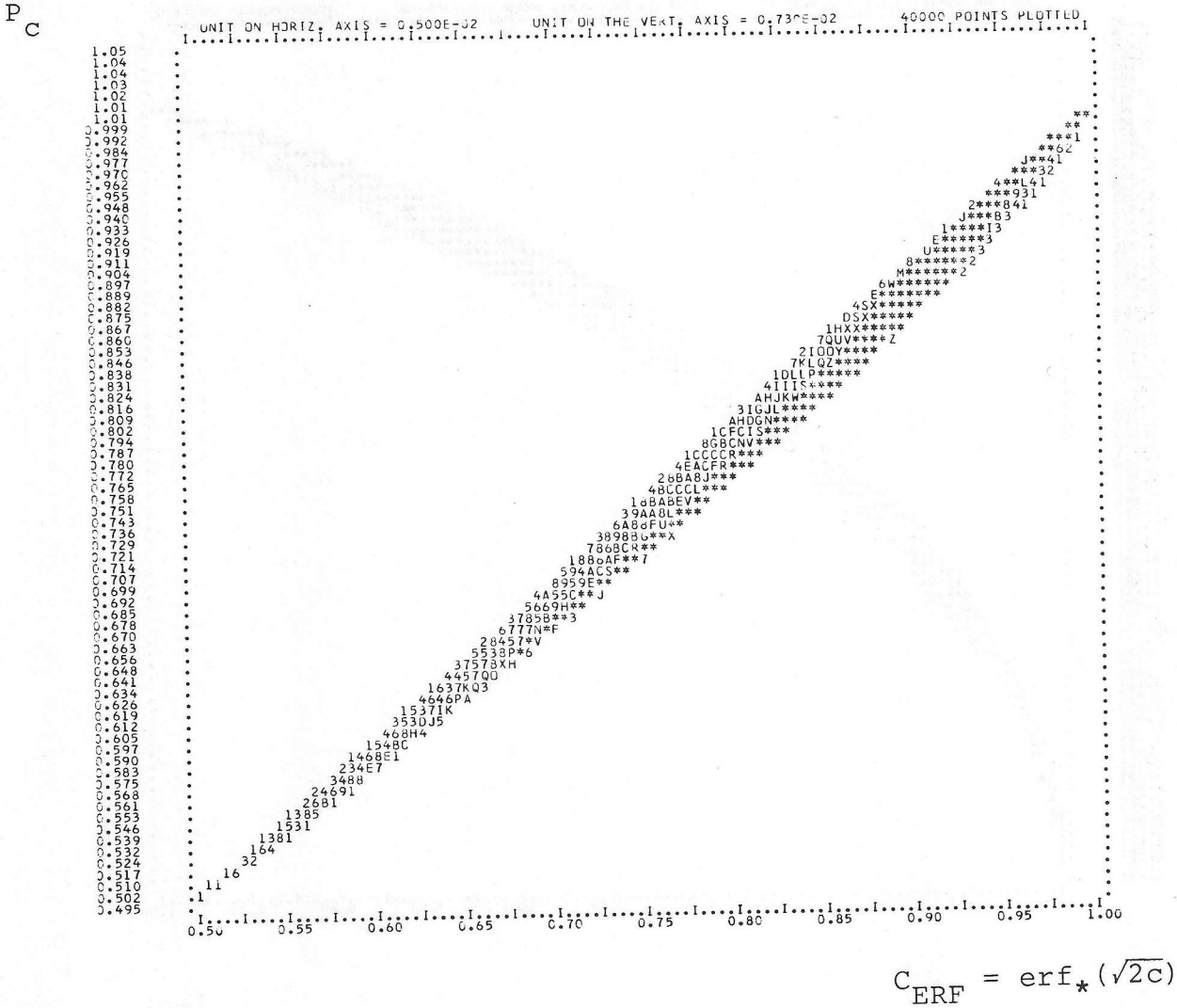


Figure 7. Probability of Correct Classification versus Error Function Chernoff Distance.