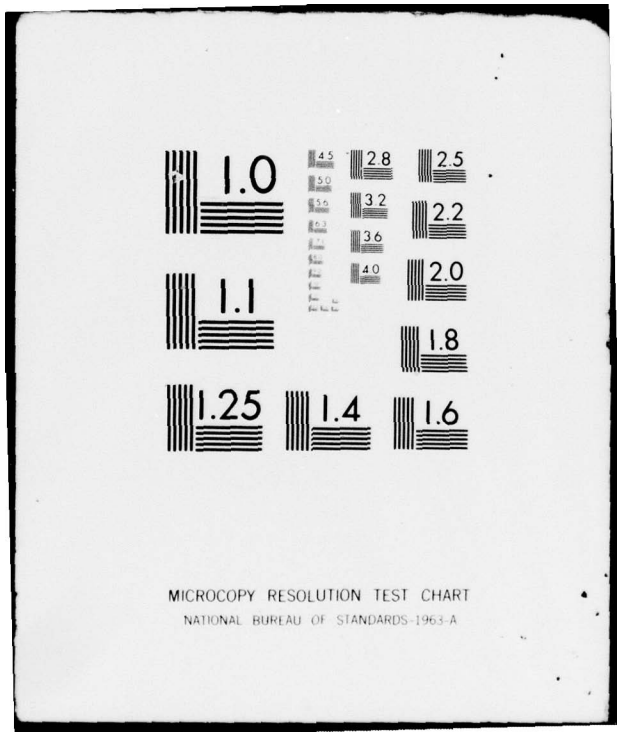
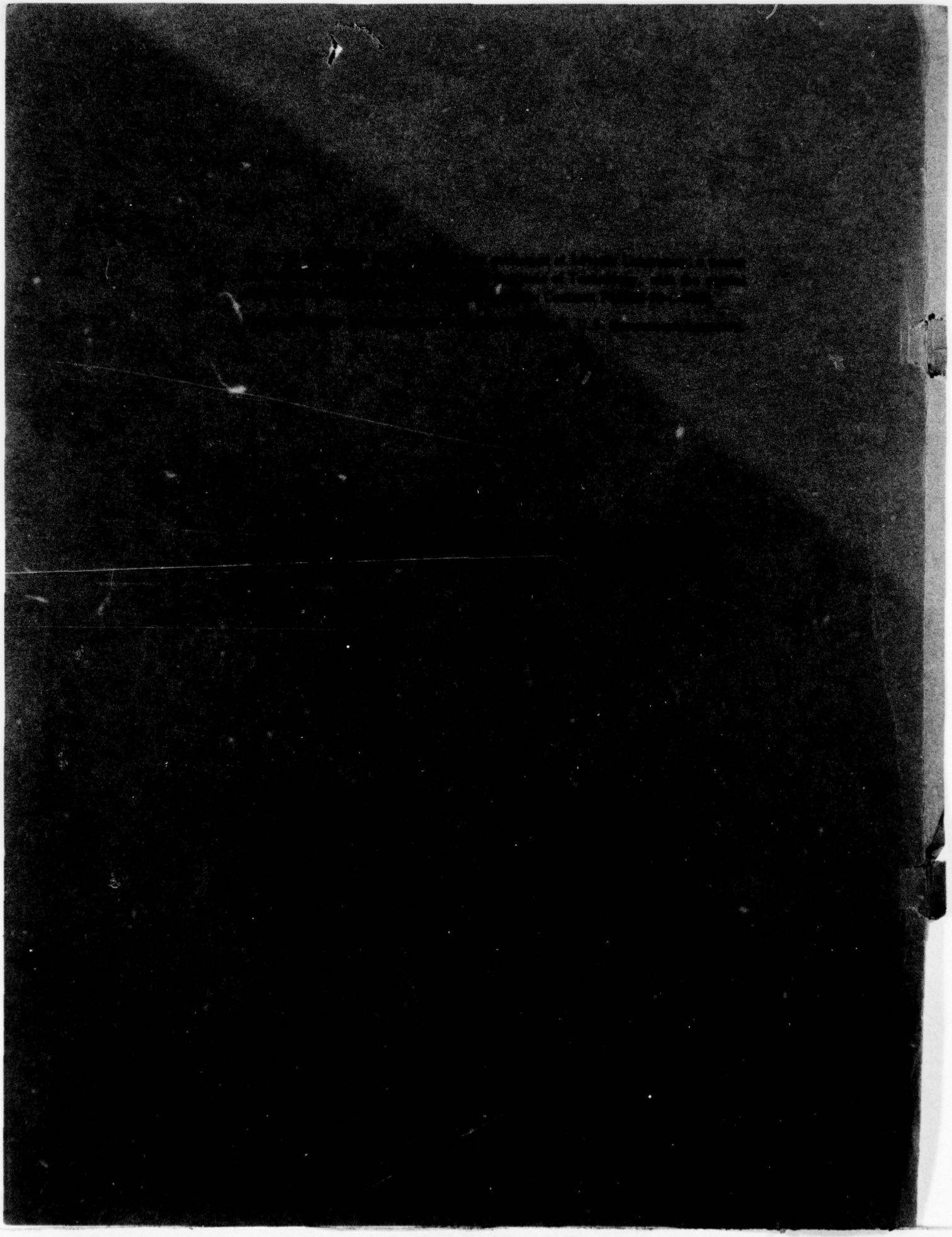
AD-A056 352 MASSACHUSETTS INST OF TECH LEXINGTON LINCOLN LAB F/G 12/1
FURTHER CONSIDERATION OF SAMPLE AND FEATURE SIZE.(U)
APR 78 I T YOUNG F19628-78-C-0002
UNCLASSIFIED TN-1978-13 ESD-TR-78-80 NL

END
DATE FILMED
8 -78
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



MASSACHUSETTS INSTITUTE OF TECHNOLOGY

LINCOLN LABORATORY

(6)

FURTHER CONSIDERATION OF SAMPLE AND FEATURE SIZE.

(10) Ian

I. T. YOUNG, Consultant

Group 69

(14) TN-1978-13

(9) TECHNICAL NOTE 1978-13

(11) 28 APRIL 1978

(12) 19 p.

(18) ESD

(19) TR-78-84

(15) F19628-78-C-0002

Approved for public release; distribution unlimited.

(16) 1227

LEXINGTON MASSACHUSETTS

ABSTRACT

In this report it is shown that in the context of a specific pattern classification decision metric the number of samples M needed to characterize a cluster described by N features is:

$$M \geq (1 + \beta^{-1})(N + 2) \qquad \text{(i)}$$

where β represents an interval width. The distance metric

$$d^2(X) = (X - \hat{\mu}_X)^t \, S_X^{-1} \, (X - \hat{\mu}_X) \qquad \text{(ii)}$$

is shown to have an F distribution which leads to result (i). An additional application of the distribution of (ii) is discussed in terms of a specific type of pattern classifier.

TABLE OF CONTENTS

I. INTRODUCTION

In choosing the number of samples that are required to estimate a correlation matrix of a multivariate Gaussian-process, a number of guidelines have been suggested. With N equal to the dimensionality of the matrix, and M equal to the number of samples the following results have been obtained:

i) $M \geq N$

ii) $M > 2N$ (Reed et al.)[1]

iii) $M > 2N$ (Cover)[2]

iv) $M > 5N$ (Foley)[3].

The first two have been studied in the context of estimation. The first result states that in order for the estimate of the matrix given in Eq. (1) to be nonsingular, the number of samples must be greater than or equal to the dimensionality of the matrix. The estimation equation for the correlation matrix formed from vectors $X_i = (X_{1i}, X_{2i}, \ldots, X_{Ni})$ is given by

$$R = \frac{1}{M} \sum_{i=1}^{M} X_i X_i^t \tag{1}$$

The second result was derived by considering the expected signal-to-noise ratio (S/N) that could be obtained in an adaptive antenna system with an infinite number of samples $[(S/N)_\infty]$ and then finding the number of samples M necessary to estimate the antenna weights such that the expected $(S/N)_M = \frac{1}{2}(S/N)_\infty$.

The third and fourth results were derived in the context of detection problems - specifically pattern recognition ones - and presented the point of view that a sufficient number of samples should be used so that samples

of a single multivariate random process would not appear to be samples from two processes. Alternatively, where two processes were present and the Bayes' error for classification could be calculated, this error could be considered as the $M = \infty$ case. We then ask, with N dimensions, how many samples are necessary such that the probability of error is close to the Bayes' error, i.e.,

$$P_M(E) \approx P_\infty(E)?$$

The results of these studies led to (iii) and (iv).

II. A NEW APPROACH BASED ON CONFIDENCE INTERVALS

Let us assume that our sample vectors $\{X_i \mid i=1,\ldots,M\}$ come from a multivariate Gaussian random process with mean, μ_X, and covariance matrix, Σ_X. Consider the quadratic form given by

$$d^2(X) = (X-\mu_X)^t \Sigma_X^{-1} (X-\mu_X) \tag{2}$$

The term d may be considered as the distance from the sample X to the mean μ_X measured in standard deviation units. This distance may then be interpreted either in the context of a signal-to-noise ratio calculation or in the context of a minimum distance classification rule. We also observe that if μ_X and Σ_X are being estimated (by $\hat{\mu}_X$ and S_X) from a sample population by the consistent* equations:

---

* A consistent estimator $\hat{\theta}$ of a parameter θ is one for which

$$\lim_{M\to\infty} \Pr[|\hat{\theta}-\theta|>\varepsilon] = 0$$

$$\hat{\mu}_X = \frac{1}{M} \sum_{m=1}^{M} X_m \tag{3a}$$

$$S_X = \frac{1}{M-1} \sum_{m=1}^{M} (X_m - \hat{\mu})(X_m - \hat{\mu})^t \tag{3b}$$

then d^2 corresponds to the infinite sample case. Since X is a random vector $d^2(X)$ is a random variable and we would like to determine its distribution.

The Distribution of $d^2(X)$

Since X is a Gaussian random vector, distributed as $N(X|\mu_X,\Sigma_X)$, any linear function of X is a Gaussian random vector. Let

$$y = W(X-\mu_X) \tag{4}$$

then:

$$\mu_y = E[W(X-\mu_X)] = W \cdot E[(X-\mu_X)] = 0 \tag{5a}$$

$$\Sigma_y = E[yy^t] = W \Sigma_X W^t \tag{5b}$$

$$d^2(y) = (W^{-1}y)^t \Sigma_X^{-1} (W^{-1}y)$$

$$d^2(y) = y^t[(W^{-1})^t \Sigma_X^{-1}(W^{-1})]y \tag{5c}$$

where W is an (N x N) invertible matrix. Specifically we will define W as

follows: Let ϕ_i be an eigenvector of Σ_X such that $\Sigma_X\phi_i = \lambda_i\phi_i$. We define Φ to be the matrix whose rows are $\{\phi_i | i=1,\ldots,N\}$ and Λ to be the diagonal matrix $\Lambda = \text{diag}(\lambda_1, \lambda_2 \ldots \lambda_N)$. Then

$$\Sigma_X \Phi^t = \Lambda\Phi^t \tag{6}$$

and since the eigenvectors are orthonormal:

$$\Phi\Phi^t = I \tag{7a}$$

this implies

$$\Phi^t = \Phi^{-1}. \tag{7b}$$

We are now prepared to define W as:

$$W = \Lambda^{-1/2}\Phi \tag{8}$$

where $\Lambda^{1/2} = \text{diag}\ (\sqrt{\lambda_1}, \sqrt{\lambda_2}, \ldots \sqrt{\lambda_N})$. Substituting Eq. (8) into Eqs. (5b, 5c) we arrive at:

$$\Sigma_y = \Lambda^{-1/2}\ \Phi\ \Sigma_X\ \Phi^t\ \Lambda^{-1/2} = I \tag{9a}$$

$$d^2(y) = y^t y = \sum_{n=1}^{N} y_i^2 \tag{9b}$$

where y_i is the i^{th} component of the y vector and is distributed according to $N(y_i|\mu_{y_i} = 0, \sigma_{y_i} = 1)$. Then $d^2(y)$ has a χ^2 distribution with N degrees-of-freedom.[4] But $d^2(y)$ (from Eq. 5c) equals $d^2(X)$ so that $d^2(X)$ has a χ^2 distribution with N degrees-of-freedom. From the properties of the χ^2 distribution we know that

$$E[d^2(X)] = N \tag{10a}$$

$$Var[d^2(X)] = 2N \tag{10b}$$

The Distribution of $d_M^2(X)$

When only a finite number of data samples are available and the parameters μ and Σ must be estimated, Eq. (2) takes the form:

$$d_M^2(X) = (X - \hat{\mu}_X)^t \, S_X^{-1} \, (X-\hat{\mu}_X) \tag{11}$$

To develop the distribution of $d_M^2(X)$ we will need the T^2 statistic. The definition and distribution of the T^2 variable are rephrased from Anderson[5] in the following:

Theorem: Define $T^2 = p^t S^{-1} p$ where p is distributed according to $N(p|0,\Sigma)$ and $(M-1)S$ is independently distributed as

$$(M-1)S = \sum_{m=1}^{M} Z_m Z_m^t$$

where the Z_m are independently distributed as $N(Z_m|0,\Sigma)$. Then

$$T^2 \left[\frac{M-N}{(M-1)N} \right] = F$$

has a central F distribution with (N,M-N) degrees-of-freedom.

We treat the random vector X as a sample independent of the set used to estimate μ_X and Σ_X.* Then, under the null hypothesis that X is distributed according to $N(X|\mu_X, \Sigma_X)$ and $\hat{\mu}_X$ is distributed according to $N(\hat{\mu}_X|\mu_X, \Sigma_X/M)$ and S_X is as given in Eq. (3b), we set

$$T^2 = \frac{M}{M+1} (X - \hat{\mu}_X)^t S_X^{-1} (X - \hat{\mu}_X) \tag{12a}$$

$$T^2 = \frac{M}{M+1} d_M^2(X) \tag{12b}$$

With $p = \sqrt{\frac{M}{M+1}} (X - \hat{\mu}_X)$ we have that the random variable T^2 has an F-distribution with (N,M-N) degrees-of-freedom and a critical region specified by

$$T^2 \geq \left[\frac{(M-1)N}{M-N} \right] F_{N,M-N}(\alpha) \tag{13}$$

---

* Fukunaga and Kessel[6] have considered the case where X is part of the same sample set used to estimate μ_X and Σ_X. Using the scalar statistic

$$u = \frac{1}{M-1} (X - \hat{\mu}_X)^t S_X^{-1} (X - \hat{\mu}_X)$$

they have derived a test for the multivariate normality of the data without prior knowledge of the mean and covariance parameters.

with α significance level. We may now determine $E[d_M^2(X)]$ and $Var[d_M^2(X)]$ as follows:

$$E[d_M^2(X)] = \frac{M+1}{M} E[T^2] = \left[\left(\frac{M+1}{M}\right)\left(\frac{M-1}{M-N}\right)N\right] E[F_{N,M-N}(\alpha)] \quad (14a)$$

$$Var[d_M^2(X)] = \left(\frac{M+1}{M}\right)^2 Var[T^2] = \left[\left(\frac{M+1}{M}\right)\left(\frac{M-1}{M-N}\right)N\right]^2 Var[F_{N,M-N}(\alpha)] \quad (14b)$$

It can be shown that the expected value of the F-distribution with (k_1, k_2) degrees-of-freedom is:[4]

$$\mu_F = \frac{k_2}{k_2-2} \qquad k_2 > 2 \quad (15a)$$

The variance is:[4]

$$\sigma_F^2 = \frac{2k_2^2(k_1 + k_2 - 2)}{k_1(k_2 - 2)^2(k_2 - 4)} \qquad k_2 > 4 \quad (15b)$$

Assuming that the stronger of these conditions is true, i.e., $k_2 = M-N > 4$ we have

$$E[d_M^2(X)] = \left[\left(\frac{M+1}{M}\right)\left(\frac{M-1}{M-N}\right)N\right]\left[\frac{M-N}{M-N-2}\right]$$

$$\boxed{E[d_M^2(X)] = \left(\frac{M+1}{M}\right)\frac{N(M-1)}{M-N-2}} \quad (16)$$

and

with α significance level. We may now determine $E[d_M^2(X)]$ and $Var[d_M^2(X)]$ as follows:

$$E[d_M^2(X)] = \frac{M+1}{M} E[T^2] = \left[\left(\frac{M+1}{M}\right)\left(\frac{M-1}{M-N}\right)N\right] E[F_{N,M-N}(\alpha)] \tag{14a}$$

$$Var[d_M^2(X)] = \left(\frac{M+1}{M}\right)^2 Var[T^2] = \left[\left(\frac{M+1}{M}\right)\left(\frac{M-1}{M-N}\right)N\right]^2 Var[F_{N,M-N}(\alpha)] \tag{14b}$$

It can be shown that the expected value of the F-distribution with (k_1, k_2) degrees-of-freedom is:[4]

$$\mu_F = \frac{k_2}{k_2-2} \qquad k_2 > 2 \tag{15a}$$

The variance is:[4]

$$\sigma_F^2 = \frac{2k_2^2(k_1 + k_2 - 2)}{k_1(k_2 - 2)^2(k_2 - 4)} \qquad k_2 > 4 \tag{15b}$$

Assuming that the stronger of these conditions is true, i.e., $k_2 = M-N > 4$ we have

$$E[d_M^2(X)] = \left[\left(\frac{M+1}{M}\right)\left(\frac{M-1}{M-N}\right)N\right]\left[\frac{M-N}{M-N-2}\right]$$

$$\boxed{E[d_M^2(X)] = \left(\frac{M+1}{M}\right)\frac{N(M-1)}{M-N-2}} \tag{16}$$

and

$$V[d_M^2(X)] = 2\left(\frac{(M-2)N}{M-N-4}\right)\left(\frac{M+1}{M}\right)^2\left(\frac{M-1}{M-N-2}\right)^2 \tag{17}$$

Effects of Finite Sample Size

We now have two expressions for distance; one based on finite sample size $d_M^2(X)$ and one on infinite sample size $d^2(X)$. As M increases we have from Eqs. (16) and (17)

$$\lim_{M\to\infty} E[d_M^2(X)] = N = E[d^2(X)] \tag{18}$$

$$\lim_{M\to\infty} \mathrm{Var}[d_M^2(X)] = 2N = \mathrm{Var}[d^2(X)] \tag{19}$$

Let us now determine (in the Reed[1] or Foley[3] sense) the value of M required so that the $E[d_M^2(X)]$ is with 100 β% of its true value. Then

$$E[d_M^2(X)] = N\left(\frac{M+1}{M}\right)\left(\frac{M-1}{M-N-2}\right) = N(1 \pm \beta) \tag{20}$$

Since the left side of Eq. (20) approaches the limiting value from above this may be written as

$$\left(\frac{M+1}{M}\right)\left(\frac{M-1}{M-N-2}\right) = 1 + \beta \tag{21}$$

Solving for M in terms of N and β gives:

$$M = \frac{(1+\beta)(N+2) \mp \sqrt{(1+\beta)^2(N+2)^2 - 4\beta}}{2\beta} \tag{22}$$

For small values of β an excellent approximation is:

$$\boxed{M = (1+\beta^{-1})(N+2)} \tag{23}$$

In Fig. 1 we plot M versus N for various values of β. For a given value of N this represents the minimum number of samples required to match the design goal, i.e., that $d_M^2(X)$ be within a certain percentage of $d^2(X)$. Note that for $\beta=1$ we return to the Reed result (ii) that $M \geq 2N$.

III. APPLICATION TO A CERTAIN TYPE OF PATTERN CLASSIFIER

In a number of pattern classification schemes we might wish to refrain from assigning a class label when the distance to each of the prototype clusters is too large. If the distance squared from a point to be classified X to a cluster ω_i is given by

$$\begin{aligned} d^2(X, \omega_i) &= (X-\hat{\mu}_i)^t \, S_i^{-1} \, (X-\hat{\mu}_i) \\ &= d_M^2(X,\hat{\mu}_i) \ , \end{aligned} \tag{24}$$

then we might refuse to classify the sample X if

$$d_M^2(X,\hat{\mu}_i) > \theta \ \forall \ i=1,\ldots,L$$

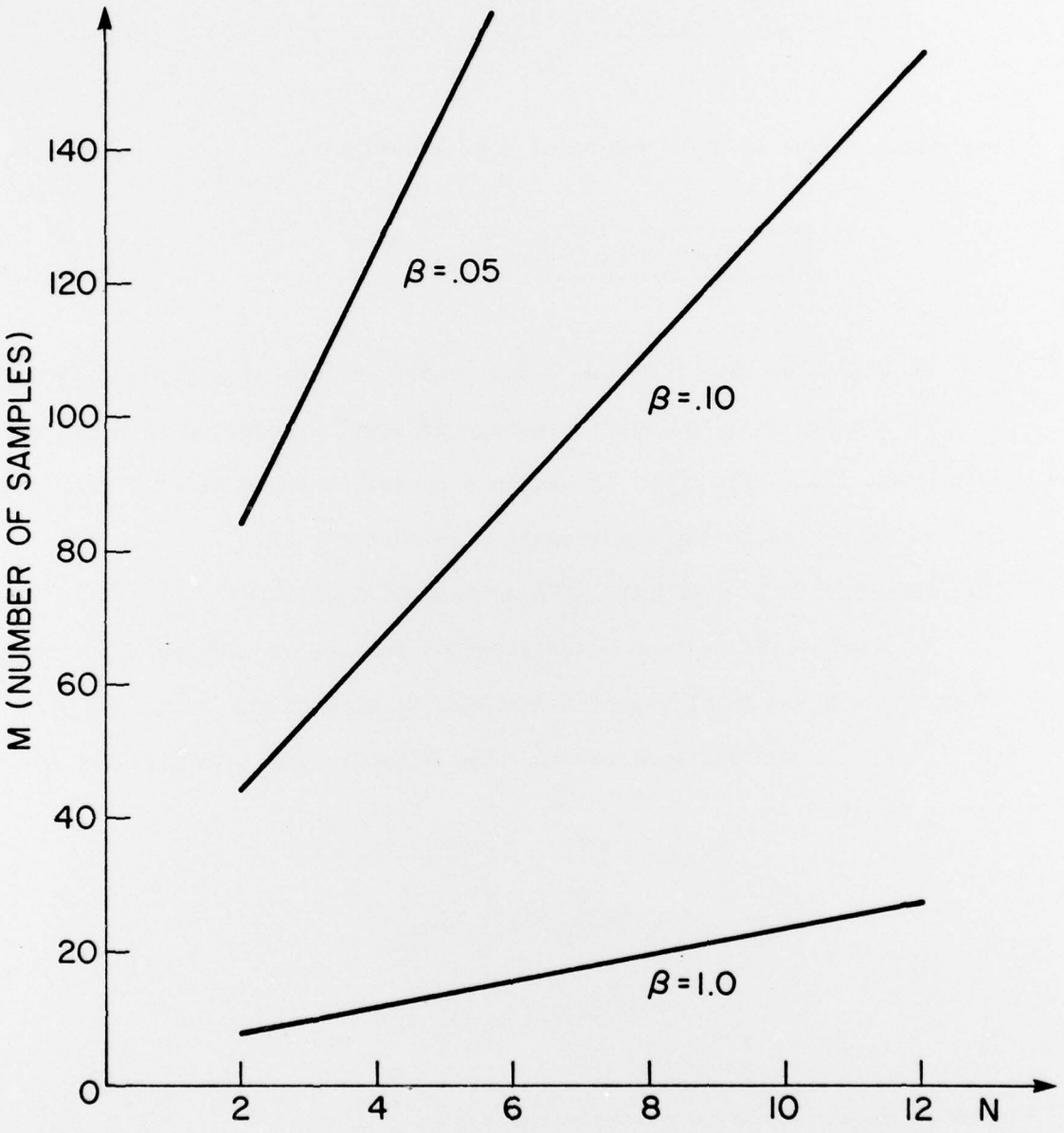


Fig.1. Number of samples for a given dimensionality and confidence range.

with L = number of different classes and θ = some arbitrary threshold. Alternatively we might consider the sample X to belong to the class labeled "other".

In this section we would like to consider how to choose the threshold θ. We do this by asking the question: If X were a member of class ω_i then how far should we expect to find it from the cluster? The answer to this is contained in Eq. (16). As an example with M=100 samples and N=6 features we have

$$E[d_M^2(X)] = \left(\frac{M+1}{M}\right) \frac{N(M-1)}{M-N-2} = 6.52 \tag{25}$$

The one standard deviation interval on the mean is given by:

$$E[d_M^2(X)] \pm \sigma[d_M^2(X)] \tag{26a}$$

$$= \left(\frac{M+1}{M}\right)\left(\frac{M-1}{M-N-2}\right)\left[N \pm \sqrt{\frac{2N(M-2)}{M-N-4}}\right] \tag{26b}$$

which for M=100, N=6 gives

$$6.52 \pm 3.93$$

Finally using F-tables[4] we may find the 99% confidence interval for $d_M^2(X)$ as the one-sided interval

$$P[0 \leq d_M^2(X) \leq d_*^2] = .99$$

where

$$d_*^2 = \left(\frac{M+1}{M}\right) N\left(\frac{M-1}{M-N}\right) F_{N,M-N}(\alpha=.01)$$

Again using M=100, N=6 we have

$$d_*^2 = 6.38\ F_{6.94}(.01) = 19.156$$

Thus 99% of the time an independent random sample drawn from a population characterized by 6 features and 100 samples will be a distance squared less than 19.156 from the sample cluster. This number, 19.156, is independent of the true mean μ and the true covariance matrix Σ thus it is the same for all clusters formed from 100 samples in a six dimensional space. It is thus a reasonable choice for the threshold θ.

IV. SUMMARY

Using the F-distributed T^2 statistic we have determined a relation for the number of samples M needed to characterize a multivariate Gaussian process with dimensionality N. The characterization of the cluster has been in the useful form of a distance metric which may be interpreted as either a signal-to-noise ratio or a pattern classification rule. Further, in the pattern recognition context, we have shown that the results can be interpreted as to how to set a distance threshold θ beyond which all pattern class labels would be rejected.

REFERENCES

1. Reed, I. S., et al., "Rapid Convergence Rate in Adaptive Arrays," IEEE Trans. Aerospace Electron Systems AES-10, 853-863 (1974).

2. Cover, T. M., "Geometrical and Statistical Properties of Linear Inequalities with Applications in Pattern Recognition," IEEE Trans. Electron. Computers EC-14, 326-334 (1965).

3. Foley, D. H., "Considerations of Sample and Feature Size," IEEE Trans. Inf. Theory IT-18, 618-626 (1972).

4. Burrington, R. S. and May, D. C., Handbook of Probability and Statistics with Tables (Handbook Publishers, Sandusky, Ohio, 1953).

5. Anderson, T. W., An Introduction to Multivariate Statistical Analysis (Wiley, New York, 1958).

6. Fukunaga, K. and Kessell, D. L., "Error Evaluation and Model Validation in Statistical Pattern Recognition," Purdue University Electrical Engineering Technical Report, TR-EE 72-23, Lafayette, Indiana, (1972).

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: ESD-TR-78-80	2. GOVT ACCESSION NO.
3. RECIPIENT'S CATALOG NUMBER	
4. TITLE (and Subtitle): Further Consideration of Sample and Feature Size	5. TYPE OF REPORT & PERIOD COVERED: Technical Note
6. PERFORMING ORG. REPORT NUMBER: Technical Note 1978-13	
7. AUTHOR(S): Ian T. Young	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: Program Element No. 63431F Project No. 1227
11. CONTROLLING OFFICE NAME AND ADDRESS: Air Force Systems Command, USAF Andrews AFB Washington, DC 20331	12. REPORT DATE: 28 April 1978
	13. NUMBER OF PAGES: 18
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)

correlation matrix | multivariate Gaussian-process | signal-to-noise ratio

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

In this report it is shown that in the context of a specific pattern classification decision metric the number of samples M needed to characterize a cluster described by N features is:

$$M \geq (1 + \beta^{-1})(N + 2) \qquad \text{(i)}$$

where β represents an interval width. The distance metric

$$d^2(X) = (X - \hat{\mu}_X)^t S_X^{-1} (X - \hat{\mu}_X) \qquad \text{(ii)}$$

is shown to have an F distribution which leads to result (i). An additional application of the distribution of (ii) is discussed in terms of a specific type of pattern classifier.

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