LARS Technical Report 050682

# Predicting the Required Number of Training Samples

by H. M. Kalayeh and D. A. Landgrebe

Laboratory for Applications of Remote Sensing
Purdue University West Lafayette, Indiana 47906 USA
1982

LARS Technical Report 050682

# Predicting the Required Number of Training Samples

by H.M. Kalayeh and D.A. Landgrebe




Laboratory for Applications of Remote Sensing
Purdue University West Lafayette, Indiana 47906 USA
1982

LARS Technical Report 050682

# PREDICTING THE REQUIRED NUMBER OF TRAINING SAMPLES

H. M. Kalayeh and D. A. Landgrebe*

## ABSTRACT

In this paper a criterion which measures the quality of the estimate of the covariance matrix of a multivariate normal distribution is developed. Based on this criterion, the necessary number of training samples is predicted. Experimental results which are used as a guide for determining the number of training samples are included.

---

* Mr. Kalayeh is Graduate Research Assistant, Laboratory for Applications in Remote Sensing (LARS), Purdue University, West Lafayette, IN 47906-1399.

Dr. Landgrebe, formerly director of LARS, 1969-81, is Associate Dean of Engineering and Director of the Engineering Experiment Station, School of Electrical Engineering, Purdue University, West Lafayette, IN 47906-0501.

The research described in this report was sponsored in part by NASA under Contract No. NSG-5414.

## 1. INTRODUCTION

In practice, the number of training samples is frequently limited because it is expensive to collect many training samples. A typical application in which this is the case is the field of remote sensing, and we will use this application to illustrate the technique.

In remote sensing, the reflected and emitted electromagnetic energy of each pixel of a scene in several important wavelength bands is measured by a multispectral remote sensor system mounted on board an aircraft or spacecraft. The output of the sensor system is used to form a point in a q-dimensional space[6]. A commonly used pattern classification algorithm in this application is the maximum likelihood Gaussian scheme. In this instance, the classes are each characterized as a Gaussian distribution in q-space and these distributions in turn are specified by estimates of the means and covariances of each. However, we know that the performance of the estimators is dependent on the number of training samples. In the case of limited training samples, the estimates of the first and second order statistics cannot accurately depict all the information which is contained in the data. In particular, the estimate of the covariance matrix may be poor. As a result of this poor estimation, later analysis of the data (for example, classification accuracy and statistical distance measures) will be degraded. See [1] for more details. Therefore, it is important to predict how many samples will be needed in order that the performance of the estimators be statistically reasonable. In the following, a criterion is developed to measure the performance of the estimate of the covariance matrix; then the number of required samples is predicted.

## 2. PREDICTION CRITERION

Let $X_1, X_2, \ldots X_n$ be q-dimensional random sample vectors which are drawn from a normally distributed population with parameters $\theta = (M, \Sigma)$, where M is the true mean vector and $\Sigma$ the true covariance matrix. In practice, M and $\Sigma$ are not available, so they must be estimated from the observed data. The maximum likelihood estimates of M and $\Sigma$ are:

$$\hat{M} = \frac{1}{N} \sum_{i=1}^{N} X_i \tag{1}$$

$$\hat{\Sigma} = \frac{1}{N} \sum_{i=1}^{N} (X_i - \hat{M})(X_i - \hat{M})^T \tag{2}$$

For more detail, see [2].

The performance of an estimator is measured by properties, such as whether it provides (a) an unbiased estimate, (b) a consistent estimate, (c) an efficient estimate, and (d) a sufficient estimate. Now, let us study the properties of maximum likelihood estimates of M and $\Sigma$. From [2] we have:

$$E[\hat{M}] = M \tag{3}$$

$$Cov[\hat{M}] = \frac{1}{N} \Sigma \tag{4}$$

$$E[\hat{\Sigma}] = \frac{N-1}{N} \Sigma \tag{5}$$

Thus, by definition, $\hat{M}$ is an unbiased estimate of M, but $\hat{\Sigma}$ is not an unbiased estimate of $\Sigma$. However, if

$$\hat{\Sigma} = \frac{1}{N-1} \sum_{i=1}^{M} (X_i - \hat{M})(X_i - \hat{M})^T \tag{6}$$

then $E[\hat{\Sigma}] = \Sigma$ which is unbiased. The density function of $\hat{M}$ and $\hat{\Sigma}$ are:

$$p(\hat{M}) = \frac{1}{(2\pi)^{\frac{q}{2}} |\frac{1}{N}\Sigma|^{\frac{1}{2}}} \exp\{-\tfrac{1}{2}(\hat{M}-M)^T N\Sigma^{-1}(\hat{M}-M)\} \tag{7}$$

$$p(\hat{\Sigma}) = \frac{(N-1)^q |\Sigma|^{(N-q-2)/2} \exp\{-\tfrac{1}{2}(N-1)\mathrm{tr}\Sigma^{-1}\hat{\Sigma}\}}{2^{(N-1)q/2} \pi^{q(q-1)/4} |\Sigma|^{(N-1)/2} \prod_{i=1}^{q} \Gamma[\tfrac{1}{2}(N-i)]} \tag{8}$$

That is, $\hat{M} \sim N(M, \frac{1}{N}\Sigma)$, a normal distribution and $\hat{\Sigma} \sim W(\Sigma, N)$, a wishart distribution. For more details of other properties of these estimators, see [2,3] and for various properties of the wishart distribution see [4].

Though the distribution of $\hat{\Sigma}$ is complex, the performance of the estimates of the covariance matrix which are of interest can be measured by the variance of the diagonal components of $\hat{\Sigma}$, as follows:

$$\hat{\sigma}_{kk} = \frac{1}{N-1} \sum_{i=1}^{N} (X_{i_k} - \hat{m}_k)^2 \tag{9}$$

In [3] it is shown that $(N-1)\frac{\hat{\sigma}_{kk}}{\sigma_{kk}}$ has a chi-square distribution with (N-1) degrees of freedom. And

$$E[\hat{\sigma}_{kk}] = \sigma_{kk} \tag{10}$$

$$E\left[\frac{\hat{\sigma}_{kk}}{\sigma_{kk}}\right] = 1 \tag{11}$$

$$\text{var}\left[\hat{\sigma}_{kk}\right] = \frac{2\sigma_{kk}^{2}}{N-1} \tag{12}$$

$$\text{var}\left[\frac{\hat{\sigma}_{kk}}{\sigma_{kk}}\right] = \frac{2}{N-1} \tag{13}$$

Now let $Y = \Lambda^{-\frac{1}{2}}\phi^{T}X$ where $\phi$ and $\Lambda$ are the eigenvector matrix and the eigenvalue matrix, respectively, of the covariance matrix, $\text{Cov}(Y) = I$, and in practice $\phi$, $\Lambda$ are the eigenvector matrix and the eigenvalue matrix of $\hat{I}$. Therefore, $Y = \hat{\Lambda}^{-\frac{1}{2}}\hat{\phi}^{T}X$ and $\text{cov}(Y) = \hat{I}$ and let the diagonal element of this matrix be $\hat{\gamma}_{kk}$. Because of the orthnormal transformation, the features in the new space are independent; therefore, $(N-1)\hat{\gamma}_{kk}$ has chi-square distribution with (N-1) degrees of freedom. For brevity, let:

$$(N-1)\hat{\gamma}_{kk} \sim \chi^{2}(N-1) \tag{14}$$

and

$$Q = [\hat{\gamma}_{11} + \ldots + \hat{\gamma}_{qq}] \tag{15}$$

then

$$(N-1)\hat{Q} \sim \chi^{2}(q(N-1)) \tag{16}$$

$$E[(N-1)\hat{Q}] = q(N-1)) \tag{17}$$

$$E[\hat{Q}] = q \tag{18}$$

$$\text{var}[(N-1)\hat{Q}] = 2q(N-1) \tag{19}$$

$$\text{var}[\hat{Q}] = \frac{2q}{N-1} \tag{20}$$

A logical choice for our prediction criterion is $\text{var}(\hat{Q})$ because it measures the dispersion of the estimate of the covariance matrix.

To see how to apply the criterion, suppose it is desired that $\text{var}(\hat{Q}) \leq \alpha$. Therefore, from (20)

$$N \geq 1 + \frac{2q}{\alpha} \tag{21}$$

Note that the minimum value of N is q + 1, because if N is less than q + 1, then the covariance matrix will be singular. So,

$$\text{var}(\hat{Q})\text{max} = \frac{2q}{N_{\text{min}}-1} = 2 \tag{22}$$

A plot of the $\text{var}(\hat{Q})$ as a function of N with q as a parameter is shown in Figure 1. Now, if for example $\alpha = 0.2$, then $N \geq 1 + 10q$.

The next question to be addressed is how does one choose a reasonable value for $\alpha$. To answer this question, let us consider the following. As shown in Figure 1, if $N > 1 + 10q$, then $\text{var}(\hat{Q})$ is decreasing very slowly and its slope is small, less than $-.02/q$. This suggests that if $N = 1 + 10q$, then the statistical distance between the true probability density and the estimated one may be close to zero. The transformed divergence[5,6] is a useful statistical distance measure and is given by

$$D_T = 2000[1 - \exp(-D/8)], \tag{23}$$

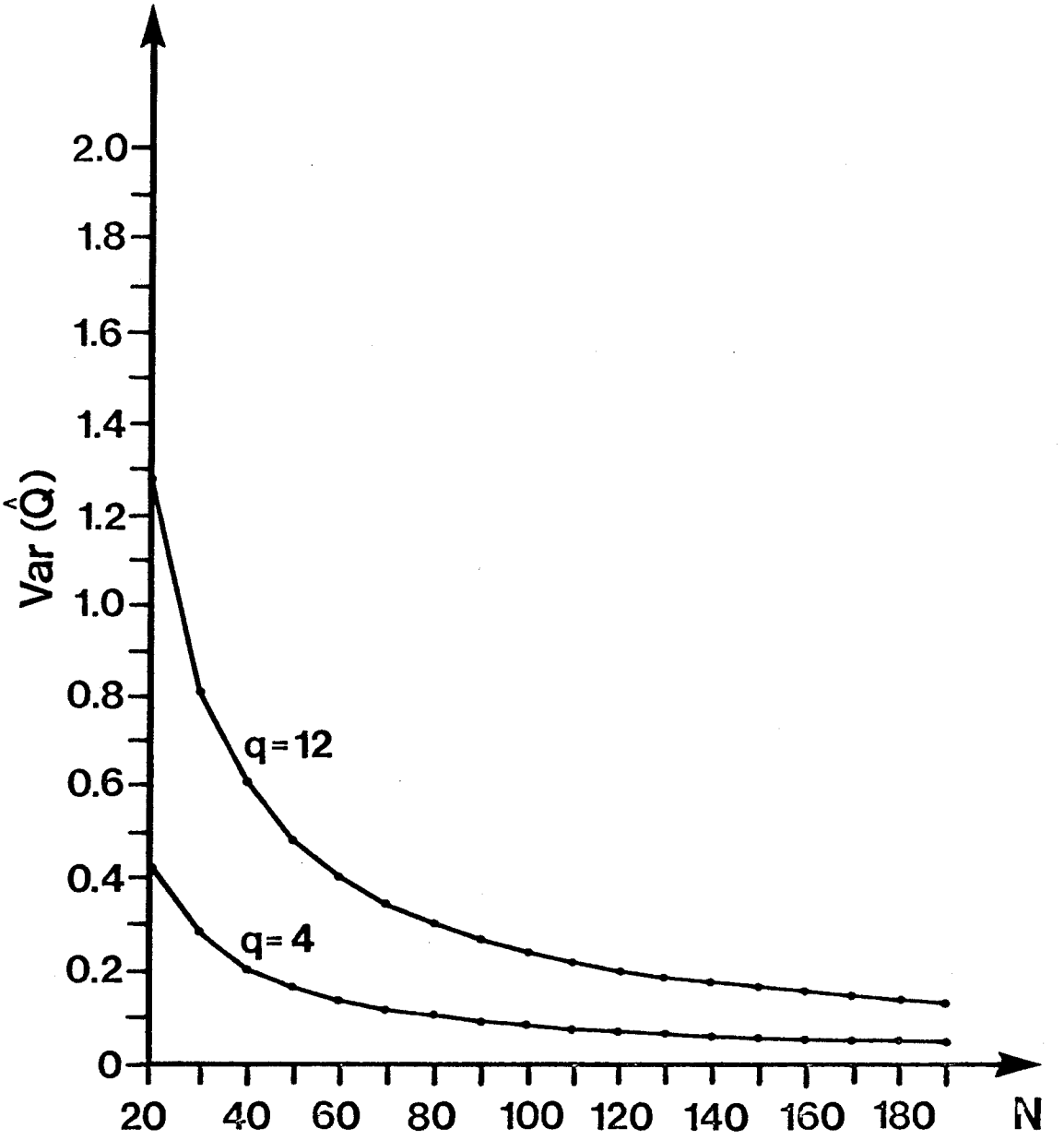


Figure 1. Variance of $\hat{Q}$ as a function of number of training samples N.

where

$$D = \tfrac{1}{2}tr(\Sigma-\hat{\Sigma})(\hat{\Sigma}^{-1}-\Sigma^{-1}) + \tfrac{1}{2}tr(\Sigma^{-1}+\hat{\Sigma}^{-1})(M-\hat{M})(M-\hat{M})^T \qquad (24)$$

We will use it to experimentally measure the quality of the estimates of the parameters and also as a guide to choosing $\alpha$ or N. The following procedure provides a practical means for doing so:

1. Assume that the true probability density of the data is normal with mean vector M and covariance matrix $\Sigma$.

2. Based on the true parameters of the distribution, $N_i$ data points are randomly generated.

3. The parameters of the distribution are estimated based on the $N_i$ randomly generated samples and then, using transformed divergence, the statistical distance between the true probability density and the estimated one is computed.

4. Step 3 is repeated five times and the average transformed divergence is calculated.

5. The average transformed divergence for different values of $var(\hat{Q})$ is computed and shown in Figure 2.

The result in Figure 2 shows almost a linear relationship between $D_T$ and $var(\hat{Q})$. This implies that when $var(\hat{Q})=var(\hat{Q})max = 2$, then $D_T=(D_T)max = 2000$. This indicates that the quality of the estimates of the parameters is very poor. However, if $var(\hat{Q}) = 0.2$, then $D_T = 175$, which suggests that the estimated probability density is very close to the true one. In practice, however, the true parameters of the distri-

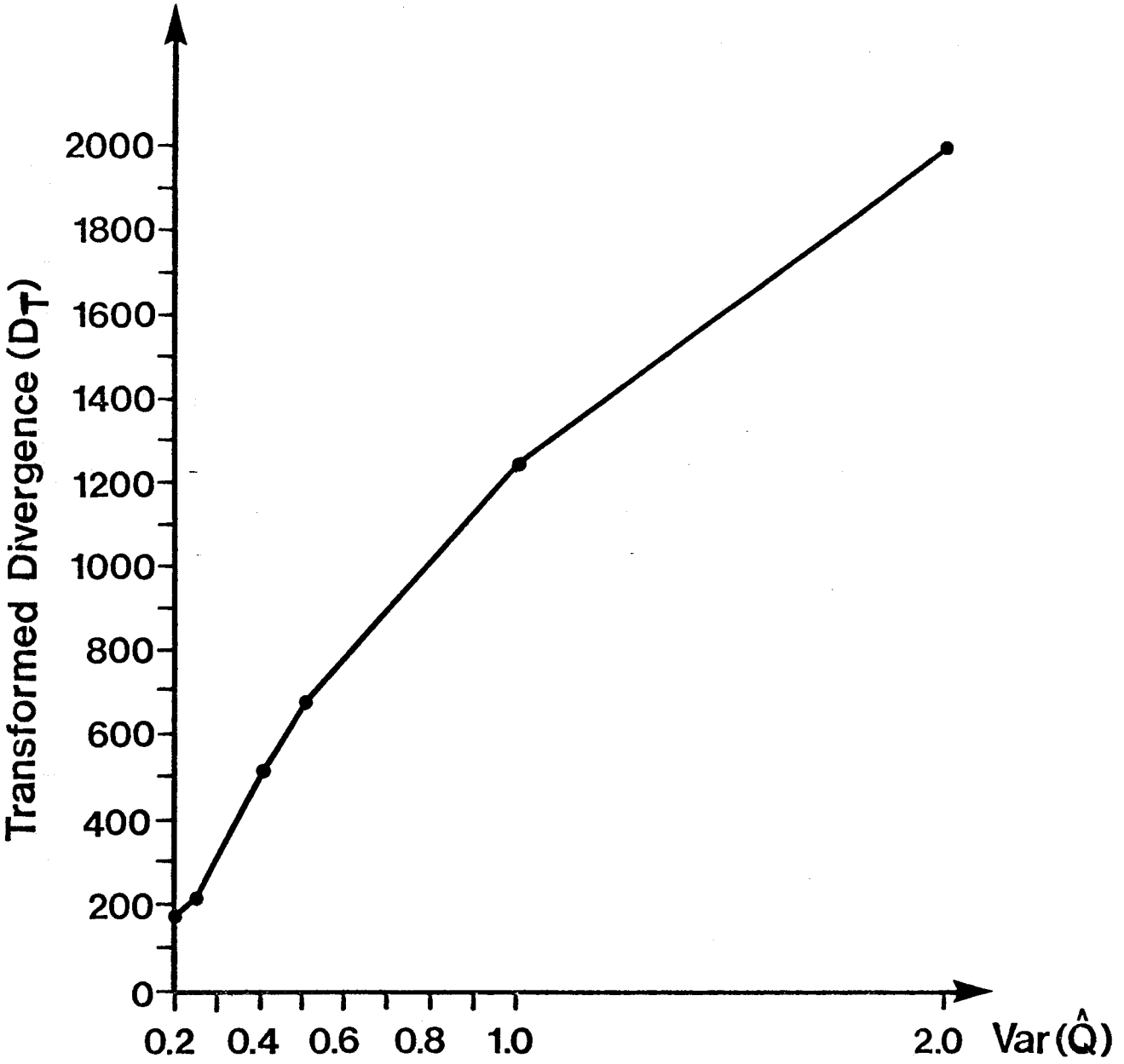


Figure 2. The average transformed divergence as a function of variance of $\hat{Q}$.

bution are not available and neither is the transformed divergence. As mentioned earlier, a logical choice for our prediction criterion is $\text{var}(\hat{Q})$ because it measures the dispersion of the estimate.

We have found that $D_T = 500$, or equivalently, $\alpha = 0.4$ is a logical threshold to decide whether the estimates of the parameters are good or not. This choice implies that the number of training samples should not be less than $1 + 5q$. However, we believe by using information given in Table 1, one should be able to establish an upperbound on $\text{var}(\hat{Q})$ and consequently estimate the required number of training samples.

## 3. CONCLUSION

The main purpose of this paper was to develop a criterion to measure the dispersion of the estimate of the covariance matrix of a multivariate normal distribution and, based on this criterion, to be able to predict the necessary number of training samples. To accomplish this, the variance of $\hat{Q} = \text{tr}(\hat{I} = \hat{\Lambda}^{-\frac{1}{2}}\hat{\phi}^T\hat{\Sigma}\hat{\phi}\hat{\Lambda}^{-\frac{1}{2}})$ was chosen as the predictor criterion. It was theoretically shown that variance of $\hat{Q}$ is equal to $\frac{2q}{N-1}$ with maximum value of 2. Also, the divergence between the true distribution and the estimated one for different values of variance of $\hat{Q}$ was experimentally computed and used to establish an upperbound on the variance of $\hat{Q}$. It was suggested that the required training samples should be about five times the number of features.

Table 1. Distance between the true distribution and estimated one as a function of $\text{var}(\hat{Q})$ or number of training samples.

| $\text{var}(\hat{Q})$ | $D_T$ | D | N |
|---|---|---|---|
| 1.00 | 1250 | 7.85 | 1 + 2q |
| 0.50 | 675 | 3.40 | 1 + 4q |
| 0.40 | 500 | 2.30 | 1 + 5q |
| 0.25 | 210 | 0.80 | 1 + 8q |
| 0.20 | 175 | 0.70 | 1 + 10q |

## 4. REFERENCES

[1] M.A. Muasher and D.A. Landgrebe. Multistage classification of multispectral earth observational data: The design approach. School of Electrical Engineering, Technical Report TR-EE 81-41, and Laboratory for Applications of Remote Sensing, Technical Report 101381, Purdue University, West Lafayette, IN 47907-0501. Dec. 1981.

[2] K. Fukunaga. Introduction to statistical pattern recognition. Academic Press, New York, 1972.

[3] J.P. Bickel and A.K. Docksum. Mathematical statistics: basic ideas and selected topics. Holden-Day, Inc., San Francisco, 1977.

[4] T.W. Anderson. Introduction to multivariate statistical analysis. Wiley, New York, 1958.

[5] P.H. Swain and R.C. King. Two Effective Feature Selection Criteria for Multispectral Remote Sensing. Laboratory for Applications of Remote Sensing, Purdue University, West Lafayette, IN 47906-1399. LARS Technical Report 042673, Apr. 1973.

[6] P.H. Swain and S.M. Davis, eds. Remote Sensing: The Quantitative Approach. McGraw Hill, Inc., New York, 1978.