TR

CR 160134

JSC-12696

Lockheed Electronics Company, Inc.
A SUBSIDIARY OF LOCKHEED CORPORATION
1830 NASA Road 1, Houston, Texas 77058
Tel. 713-333-5411

Ref: 642-7137
Job Order 73-743-39
Contract NAS 9-15800

TECHNICAL MEMORANDUM

THE MEAN-SQUARE ERROR OPTIMAL LINEAR DISCRIMINANT FUNCTION AND ITS APPLICATION TO INCOMPLETE DATA VECTORS

By

H. F. Walker

(NASA-CR-160134) THE MEAN-SQUARE ERROR OPTIMAL LINEAR DISCRIMINANT FUNCTION AND ITS APPLICATION TO INCOMPLETE DATA VECTORS (Lockheed Electronics Co.) 36 p HC A03/MF A01 CSCL 09B G3/63
N79-21827
Unclas 24262

Approved By: T. C. Minter
T. C. Minter, Supervisor
Techniques Development Section

February 1979

LEC-12773

CONTENTS

TABLE

FIGURES

THE MEAN-SQUARE ERROR OPTIMAL LINEAR DISCRIMINANT FUNCTION AND ITS APPLICATION TO INCOMPLETE DATA VECTORS

1. INTRODUCTION

In many pattern recognition problems, it is desirable to map observed data vectors in n-dimensional Euclidean space R^n, $n > 1$, to R^1 in order that, hopefully, either the efficiency of classification will be increased by classifying the transformed observations in R^1 or new insights into the structure of the data will be gained by literally viewing the transformed observations in R^1. A map used for such purposes should be simple in structure while having the property that transformed observations from the different populations under consideration are as well separated as possible. Consequently, such a map is often chosen to be linear[1] and, in some sense, optimal from the point of view of separating the transformed observations of the populations at hand.

In some applications, one may expect to encounter data vectors which are incomplete in the sense that the values of one or more of the components of these vectors are unknown or missing. In such circumstances, both the choice and the implementation of a linear map from R^n to R^1 require special consideration. Indeed, once a linear map from R^n to R^1 has been chosen, it is necessary that incomplete data vectors be mapped to R^1 in such a way that their images are statistically compatible with the linear images of complete data vectors. Furthermore, if n is large and if components of data vectors can be missing at random, then it is generally impossible to prepare in advance enough "compatible" maps to meet every missing-component eventuality. Consequently, it seems advantageous in such circumstances to choose a linear map from R^n to R^1 for which one can easily obtain a "compatible" linear map appropriate for each incomplete data vector as it is encountered.

[1] Throughout this memorandum, the term "linear" is used in reference to affine maps (linear plus constant) as well as to maps which are truly linear.

When there are only two populations under consideration, the mean-square error (MSE) optimal linear discriminant function (LDF) is a linear mapping from R^n to R^1 which provides, in a certain sense, optimal separation in R^1 of the transformed observations from the two populations. It is particularly suitable for applications in which incomplete data vectors occur for two reasons.

The first reason is that, for every k, the MSE-optimal LDF from R^k to R^1 imposes certain statistical properties on transformed observations in R^1. One consequence of these properties is that the MSE-optimal LDF appropriate for each subset of the components of a data vector is automatically compatible from the point of view of classification with the MSE-optimal LDF for a full data vector. Another consequence is that a map closely related to the MSE-optimal LDF, referred to in the following as the "derived map," can be defined for each subset of the components of a data vector and is guaranteed to be compatible with the derived map for a full data vector in a way appropriate for viewing the structure of transformed observations in R^1.

The second reason is that, as each incomplete data vector is encountered, both the MSE-optimal LDF and the corresponding derived map appropriate for the known components of the vector can be determined relatively easily from a relatively small amount of stored information.

In the following sections, the MSE approach to classifier design is reviewed and specialized to the two-class case and the linear discriminant function. (Much of the material offered in this connection is either standard or an adaptation of standard material to suit the present purposes (see ref. 1).) Observing the statistical properties which the MSE-optimal LDF imposes on transformed observations in R^1, the derived maps are defined and their statistical properties are discussed. We conclude by describing three methods: a "straightforward" method, a method due to Kittler (ref. 2), and a method proposed by Golub[2] by which the MSE-optimal LDF and the corresponding

[2] Private communication.

derived map appropriate for a subset of the components of a vector can be constructed with relatively little computation and storage. The relative advantages of the three methods depend in general on the particular application at hand in a fairly complex way; this dependence is discussed in some detail.

2. THE MSE APPROACH TO CLASSIFIER DESIGN

Suppose that $x = (x_1, \cdots, x_n)^T$ is an observation in R^n known to be on one of m statistical populations $\omega_1, \cdots, \omega_m$. If the cost of misclassifying an observation is taken to be 1, then the Bayes optimal classification rule is the following: assign x to ω_i if and only if

$$P(\omega_i/x) = \max_j P(\omega_j/x)$$

where $P(\omega_i/x)$ denotes the posterior probability that x is an observation on ω_i. It is often difficult or impossible to evaluate the posterior probabilities $P(\omega_i/x)$ and therefore one must frequently deal with approximations of these probabilities in implementing the Bayes rule. Such approximations are commonly of the form

$$P(\omega_i/x) \approx a_i^T \Phi(x) \quad ; \quad i = 1, \cdots, m \tag{1}$$

where $\Phi(x) = [\phi_1(x), \cdots, \phi_r(x)]^T$ is a vector whose components are conveniently chosen linearly independent functions of x, and for each i, $a_i = (a_{i1}, \cdots, a_{ir})^T$ is a parameter vector determined so that approximation (1) is optimal in some sense. The classification rule determined by a set of such so-called discriminant functions $a_i^T \Phi$ is the following: assign x to ω_i if and only if $a_i^T \Phi(x) = \max_j a_j^T \Phi(x)$.

In the MSE approach to classifier design, one attempts to determine parameter vectors a_i which minimize the MSE of approximation (1), given by

$$J(A) = \int_{R^n} |P(x) - A^T \Phi(x)|^2 p(x) dx$$

where

$P(x) = [P(\omega_1|x), \cdots, P(\omega_m|x)]^T$
$A \quad = (a_1, \cdots, a_m)$
$p(x)$ = the unconditional probability density function of x
$|\ |$ = the Euclidean norm, i.e., $|u|^2 = u^T u$

In practice, J(A) can seldom be evaluated exactly. Typically, a labeled sample $\chi = \{x_k\}_{k=1,\cdots,N}$ of independent observations on the mixture of $\omega_1, \cdots, \omega_m$ is given, and the objective function

$$\bar{J}(A) = \sum_{k=1}^{N} \left|\alpha_k - A^T\Phi(x_k)\right|^2$$

is minimized, where $\alpha_k = (\alpha_{k1}, \cdots, \alpha_{km})^T$ is defined by

$$\alpha_{kj} = \begin{cases} 1 \text{ if } x_k \varepsilon \omega_j \\ 0 \text{ if } x_k \not\varepsilon \omega_j \end{cases} ; \quad j = 1, \cdots, m$$

For large samples, the minimizer of $\bar{J}$ should be approximately the same as the minimizer of J. Indeed, if one denotes by χ_i the subset of χ consisting of observations on ω_i and by N_i the number of observations in χ_i, then

$$\frac{1}{N}\bar{J}(A) = \frac{1}{N}\sum_{x_k \varepsilon \chi} \left|A^T\Phi(x_k)\right|^2 - 2\sum_{i=1}^{m}\frac{N_i}{N}\left[\frac{1}{N_i}\sum_{x_k \varepsilon \chi_i} a_i^T\Phi(x_k)\right] + 1$$

It follows from the strong law of large numbers (see ref. 3) that with probability 1,

$$\lim_{N\to\infty}\frac{1}{N}\bar{J}(A) = \int_{R^n} \left|A^T\Phi(x)\right|^2 p(x)dx - 2\int_{R^n} P(x)^T A^T\Phi(x)p(x)dx + 1$$

Since this expression differs from J(A) by a constant independent of A, it has the same minimizer as J.

A necessary condition for A to be a minimizer of $\bar{J}$ is that all partial derivatives of $\bar{J}$ with respect to the entries of a vanish. This is equivalent to the condition that SA = B, where

$$S = \sum_{k=1}^{N} \Phi(x_k)\Phi(x_k)^T$$

and

$$B = \sum_{k=1}^{N} \Phi(x_k)\alpha_k^T$$

If S is nonsingular, this condition is sufficient as well as necessary, and $\bar{J}$ has a unique minimizer,

$$A = S^{-1}B \tag{2}$$

Since the functions $\phi_1, \cdots, \phi_r$ are linearly independent, the matrix

$$\int_{R^n} \Phi(x)\Phi(x)^T p(x)dx$$

is nonsingular, and it follows that, with probability 1, S is nonsingular for sufficiently large N. In fact, if $\phi_1, \cdots, \phi_r$ are real-analytic as well as linearly independent, S is nonsingular with probability 1 whenever $N \geq r$. (See Appendix 2 of ref. 4.) Thus it is reasonable to assume in the following that S is nonsingular and that the unique minimizer of $\bar{J}$ is given by eq. (2). The discriminant functions $a_i^T\Phi$ determined by eq. (2) are referred to as the MSE-optimal discriminant functions.

3. THE MSE-OPTIMAL LDF IN THE TWO-CLASS CASE

Suppose that there are only two statistical populations under consideration. The classification rule determined by discriminant functions $a_1^T\Phi$ and $a_2^T\Phi$ can be phrased in the following way: assign x to ω_1 if and only if $a^T\Phi(x) > 0$, where $a = a_1 - a_2$. If $a_1^T\Phi$ and $a_2^T\Phi$ are the MSE-optimal discriminant functions determined by eq. (2) on the basis of some labeled sample, then a can be obtained by right-multiplying both sides of eq. (2) by $(1, -1)^T$ to yield

$$a = S^{-1}b \tag{3}$$

where

$$b = \sum_{k=1}^{N} \Phi(x_k)\beta_k$$

and

$$\beta_k = \begin{cases} +1 \text{ if } x_k \varepsilon \omega_1 \\ -1 \text{ if } x_k \varepsilon \omega_2 \end{cases}$$

For a so defined, $a^T\Phi$ is referred to as the MSE-optimal discriminant function.

The MSE-optimal LDF is the MSE-optimal discriminant function $a^T\Phi$ obtained by defining $\Phi(x) = \binom{1}{x}$. (Vectors and matrices are expressed in partitioned forms whose meanings should be clear from the context.) In this section, an explicit expression is derived for the MSE-optimal LDF in terms of the observations in a given labeled sample.

One easily obtains

$$b = \begin{bmatrix} N_1 - N_2 \\ \sum\limits_{x_k \varepsilon X_1} x_k - \sum\limits_{x_k \varepsilon X_2} x_k \end{bmatrix} = \begin{bmatrix} N_1 - N_2 \\ N_1 m_1 - N_2 m_2 \end{bmatrix}$$

$$S = \begin{bmatrix} N & \sum_{k=1}^{N} x_k^T \\ \sum_{k=1}^{N} x_k & \sum_{k=1}^{N} x_k x_k^T \end{bmatrix} = \begin{bmatrix} N & Nm^T \\ Nm & S_W + N_1 m_1 m_1^T + N_2 m_2 m_2^T \end{bmatrix}$$

where

$$m_i = \frac{1}{N_i} \sum_{x_k \varepsilon X_i} x_k \quad ; \quad i = 1, 2$$

$$m = \frac{1}{N} \sum_{k=1}^{N} x_k$$

and

$$S_W = \sum_{k=1}^{N} x_k x_k^T - N_1 m_1 m_1^T - N_2 m_2 m_2^T$$

The matrix S_W is called the "within-class scatter matrix" and can be written as

$$S_W = S_1 + S_2$$

where

$$S_i = \sum_{x_k \varepsilon X_i} (x_k - m_i)(x_k - m_i)^T$$

is the scatter matrix for ω_i.

Setting $a = \begin{pmatrix} a_0 \\ a' \end{pmatrix}$, where a_0 is a scalar and $a' \varepsilon R^n$, one sees that eq. (3) is equivalent to

$$\begin{bmatrix} N & Nm^T \\ Nm & S_W + N_1m_1m_1^T + N_2m_2m_2^T \end{bmatrix} \begin{bmatrix} a_0 \\ a' \end{bmatrix} = \begin{bmatrix} N_1 - N_2 \\ N_1m_1 - N_2m_2 \end{bmatrix}$$

This equation yields

$$a_0 = \frac{1}{N}(N_1 - N_2) - m^Ta' \qquad (4)$$

$$S_Wa' = N_1m_1 - N_2m_2 - Nma_0 - (N_1m_1m_1^T + N_2m_2m_2^T)a' \qquad (5)$$

Substituting eq. (4) into eq. (5), one obtains

$$S_W a' = N_1m_1 - N_2m_2 - (N_1 - N_2)m + (Nmm^T - N_1m_1m_1^T - N_2m_2m_2^T)a'$$

$$= 2\frac{N_1N_2}{N}(m_1 - m_2) - \frac{N_1N_2}{N}(m_1 - m_2)(m_1 - m_2)^Ta'$$

$$= \frac{N_1N_2}{N}\left[2 - (m_1 - m_2)^Ta'\right](m_1 - m_2) \qquad (6)$$

One sees from eq. (6) that, except for an unimportant scale factor, a' is the Fisher linear discriminant $S_W^{-1}(m_1 - m_2)$.

Writing $a' = \lambda S_W^{-1}(m_1 - m_2)$ and substituting this expression in eq. (6), one obtains

$$\lambda(m_1 - m_2) = \frac{N_1N_2}{N}\left[2 - \lambda||m_1 - m_2||^2\right](m_1 - m_2)$$

where the vector norm $||\ ||$ is defined by $||u||^2 = u^TS_W^{-1}u$. It follows from this equation that

$$\lambda\left(1 + \frac{N_1N_2}{N}||m_1 - m_2||^2\right) = \frac{2N_1N_2}{N}$$

or

$$\lambda = \frac{2N_1N_2}{\left(N + N_1N_2||m_1 - m_2||^2\right)}$$

$$= \frac{2}{\left(\frac{1}{N_1} + \frac{1}{N_2} + ||m_1 - m_2||^2\right)}$$

Thus

$$a' = \frac{2}{\left(\frac{1}{N_1} + \frac{1}{N_2} + ||m_1 - m_2||^2\right)} S_W^{-1}(m_1 - m_2) \tag{7}$$

Substituting eq. (7) into eq. (4)

$$a_0 = \frac{1}{N}(N_1 - N_2) - m^T\left[\frac{2}{\left(\frac{1}{N_1} + \frac{1}{N_2} + ||m_1 - m_2||^2\right)} S_W^{-1}(m_1 - m_2)\right]$$

Using algebra, one obtains the simpler expression

$$a_0 = \frac{\left(\frac{1}{N_2} + ||m_2||^2\right) - \left(\frac{1}{N_1} + ||m_1||^2\right)}{\left(\frac{1}{N_1} + \frac{1}{N_2} + ||m_1 - m_2||^2\right)} \tag{8}$$

From eqs. (7) and (8), the MSE-optimal LDF is seen to be

$$a^T\Phi(x) = a_0 + a'^Tx$$

$$= \frac{\left(\frac{1}{N_2} + ||m_2||^2\right) - \left(\frac{1}{N_1} + ||m_1||^2\right)}{\left(\frac{1}{N_1} + \frac{1}{N_2} + ||m_1 - m_2||^2\right)}$$

$$+ \frac{2}{\left(\frac{1}{N_1} + \frac{1}{N_2} + ||m_1 - m_2||^2\right)} (m_1 - m_2)^T S_W^{-1} x \tag{9}$$

4. STATISTICAL PROPERTIES IMPOSED BY THE MSE-OPTIMAL LDF AND THE DERIVED MAP

Once a map from R^n to R^1 has been chosen, it is necessary to map incomplete data vectors to R^1 in such a way that their images are statistically compatible with the images of complete data vectors. The phrase "statistically compatible" should be interpreted in a way appropriate for the intended objective of mapping data vectors in R^n to R^1, whether the objective is effecient classification or gaining new insights into the structure of the data set. Certain statistical properties imposed by the MSE-optimal LDF on transformed observations in R^1 and the implications of these properties in determining "statistically compatible" families of maps will now be discussed.

Suppose that efficient classification is the objective of mapping data vectors in R^n to R^1. Let $L(x) = a^T\Phi(x)$ denote the MSE-optimal LDF as determined by eq. (9) on the basis of a labeled sample of observations on two populations ω_1 and ω_2. According to the classification rule associated with L, zero is the threshold for discriminating between transformed observations on ω_1 and transformed observations on ω_2. Since this is true independent of n, it follows that the MSE-optimal LDF appropriate for each subset of the components of a data vector has the same associated classification regions in R^1 as the MSE-optimal LDF for a full data vector. Thus the family of all MSE-optimal LDF's appropriate for subsets of components of data vectors may be regarded as being statistically compatible in a way appropriate for this objective.

Now suppose that the primary objective of mapping data vectors in R^n to R^1 is viewing the data. It is easily verified that

$$L(m_1) = \frac{\frac{1}{N_2} - \frac{1}{N_1} + ||m_1 - m_2||^2}{\frac{1}{N_1} + \frac{1}{N_2} + ||m_1 - m_2||^2} \qquad (10)$$

and

$$L(m_2) = \frac{\frac{1}{N_2} - \frac{1}{N_1} - ||m_1 - m_2||^2}{\frac{1}{N_1} + \frac{1}{N_2} + ||m_1 - m_2||^2} \tag{11}$$

From eqs. (10) and (11), one sees that $L(m_1)$ and $L(m_2)$ are in (-1, 1) and lie symmetrically to the left and right, respectively, of

$$\frac{\frac{1}{N_2} - \frac{1}{N_1}}{\frac{1}{N_1} + \frac{1}{N_2} + ||m_1 - m_2||^2}$$

Note that if N_1/N_2 is large, then both $L(m_1)$ and $L(m_2)$ are near +1. If this ratio is small, then both $L(m_1)$ and $L(m_2)$ are near -1. If $N_1 = N_2$, then $L(m_1) = -L(m_2)$.

As an interesting aside, explore the limiting behavior of $L(m_1)$ and $L(m_2)$ for increasingly large samples. Suppose that N_1 and N_2 grow large in such a way that

$$\lim_{N\to\infty} \frac{N_1}{N} = \alpha_1$$

and

$$\lim_{N\to\infty} \frac{N_2}{N} = \alpha_2$$

for some nonnegative α_1 and α_2 satisfying $\alpha_1 + \alpha_2 = 1$. Denote by μ_i and Σ_i the mean vector and covariance matrix, respectively, for observations on ω_i in R^n, and set $\Sigma = \alpha_1\Sigma_1 + \alpha_2\Sigma_2$. Using algebra, it follows from the strong law of large numbers that, with probability 1,

$$\lim_{N\to\infty} L(m_1) = \frac{\frac{1}{\alpha_2} - \frac{1}{\alpha_1} + ||\mu_1 - \mu_2||^2}{\frac{1}{\alpha_1} + \frac{1}{\alpha_2} + ||\mu_1 - \mu_2||^2}$$

$$\lim_{N\to\infty} L(m_2) = \frac{\frac{1}{\alpha_2} - \frac{1}{\alpha_1} - ||\mu_1 - \mu_2||^2}{\frac{1}{\alpha_1} + \frac{1}{\alpha_2} + ||\mu_1 - \mu_2||^2}$$

where the norm $|| \ ||$ is now defined by

$$||u||^2 = u^T\Sigma^{-1}u \text{ for } u\varepsilon R^n$$

It is evident from eqs. (10) and (11) that L maps the sample means in R^n to points in R^1 which depend on the sample means and variances in R^n. From L, however, one can easily obtain . map, called the "derived map" and denoted by L', for which this is not the case. Specifically, we define

$$L'(x) = \frac{\left(\frac{1}{N_1} + \frac{1}{N_2} + ||m_1 - m_2||^2\right)}{||m_1 - m_2||^2} L(x) + \frac{\frac{1}{N_1} - \frac{1}{N_2}}{||m_1 - m_2||^2}$$

$$= \frac{1}{||m_1 - m_2||^2}\left[||m_2||^2 - ||m_1||^2 + 2(m_1 - m_2)^T S_W^{-1} x\right] \qquad (12)$$

One verifies immediately that $L'(m_1) = +1$ and $L'(m_2) = -1$, independent of n and the sample statistics in R^n. It follows that for each subset of the components of a data vector, the derived map appropriate for that component subset maps the sample mean vectors of ω_1 and ω_2 in that component subset to +1 and -1, respectively. Consequently, one may regard the family of derived maps associated with subsets of components of data vectors as being statistically compatible for the objective of viewing the data to the extent that sample means of ω_1 and ω_2 in subsets of components of data vectors are compatibly mapped.

If both viewing the data and classification are objectives of mapping data vectors in R^n to R^1, one may easily associate a classification rule with L' identical to that associated with L by observing that $L(x) > 0$ if and only if

$$L'(x) > \frac{\frac{1}{N_1} - \frac{1}{N_2}}{||m_1 - m_2||^2}$$

5. CONSTRUCTION OF THE MSE-OPTIMAL LDF AND THE DERIVED MAP FOR INCOMPLETE DATA VECTORS

To successfully employ the MSE-optimal LDF or the derived map in mapping incomplete data vectors to R^1, it may be essential to construct appropriate maps for incomplete data vectors as efficiently and accurately as possible. Three methods are offered for constructing the MSE-optimal LDF or the derived map appropriate for a subset of the components of a full data vector: a "straightforward" method, Kittler's method, and Golub's method. These methods require relatively little computation and storage, and require no "retraining", i.e., no direct dealing with the original labeled sample; hence they are well-suited for applications in which it is desirable to construct the appropriate map for each incomplete data vector as it is encountered.

The underlying algebraic problem which must be solved to obtain the MSE-optimal LDF or the derived map appropriate for an incomplete data vector is described in the following section. Three procedures for solving this problem are offered and three methods are derived for constructing the desired MSE-optimal LDF and the derived map. A discussion of the relative advantages of these methods in applications, focusing on the relative efficiency and accuracy of the methods, concludes the section.

5.1 THE ALGEBRAIC PROBLEM

Suppose that A is a positive-definite symmetric $k \times k$ matrix and that u and v are vectors in R^k satisfying $Au = v$. For given indices $i_1, \cdots, i_\ell$, $\ell < k$, denote by $\hat{v}$ the vector in $R^{k-\ell}$ obtained by deleting components $i_1, \cdots, i_\ell$ from v, and denote by $\hat{A}$ the $(k - \ell) \times (k - \ell)$ matrix obtained by deleting rows and columns $i_1, \cdots, i_\ell$ from A. Consider the following problem: Find $u^* \varepsilon R^{k-\ell}$ which satisfies $\hat{A}u^* = \hat{v}$. This algebraic problem is the fundamental problem which must be solved in constructing the MSE-optimal LDF or the derived map appropriate for an incomplete data vector. Three procedures for solving this problem follow; each procedure assumes that some information associated with the equation $Au = v$ is initially available.

In the first procedure, it is assumed that A and v are available. The procedure consists of simply forming $\hat{A}$ and $\hat{v}$ and solving $\hat{A}u^* = \hat{v}$ in a straightforward manner. In formulating this procedure, it is specified that the equation is to be solved by first obtaining the Cholesky decomposition of $\hat{A}$ and then solving the resulting triangular systems. This method is not only stable but also faster than competing methods such as Gaussian elimination. For a full discussion of this method and related methods, see references 5 and 6.

PROCEDURE 1:

a. Form $\hat{v}$ and $\hat{A}$ by deleting the components of v and the rows and columns of A indexed by i_j; $j = 1, \cdots, \ell$.

b. Obtain in the usual way the $(k - \ell) \times (k - \ell)$ upper-triangular Cholesky factor R* satisfying $R^{*T}R^* = \hat{A}$.

c. Obtain u* by solving in order the triangular systems $R^{*T}z^* = \hat{v}$ and $R^*u^* = z^*$.

In the second procedure, it is assumed that A^{-1} and u are available. The basis of this procedure is the following observation (ref. 2): If $\ell = 1$ and i_1 is denoted simply by i, then

$$\hat{A}^{-1} = D - \frac{1}{d} rr^T \tag{13}$$

where

d is the *ith* diagonal element of A^{-1}

D is the $(k - 1) \times (k - 1)$ matrix obtained by deleting the *ith* row and column from A^{-1}

r is the $(k - 1)$-dimensional vector obtained from the *ith* column of A^{-1} by deleting d from it

It follows that u* is given by

$$u^* = \hat{u} - \frac{u_i}{d} r \tag{14}$$

where u_i is the *ith* component of u and $\hat{u}$ is obtained by deleting u_i from u. For general ℓ, u* can be obtained from u and A^{-1} by repeated applications of eqs. (13) and (14). For convenience in describing Procedure 2, assume that $i_1 > i_2 > \cdots > i_\ell$.

PROCEDURE 2:

a. Set $u(0) = u$ and $A^{-1}(0) = A^{-1}$.

b. For $j = 1, \cdots, \ell$, do the following:

1. Set

$$u(j) = \hat{u}(j - 1) - \frac{u_{i_j}(j - 1)}{d(j - 1)} r(j - 1)$$

where

$u_{i_j}(j - 1)$ is the component of $u(j - 1)$ indexed by i_j

$\hat{u}(j - 1)\varepsilon R^{k-j}$ is obtained by deleting $u_{i_j}(j - 1)$ from $u(j - 1)$

$d(j - 1)$ is the diagonal entry of $A^{-1}(j - 1)$ indexed by i_j

$r(j - 1)\varepsilon R^{k-j}$ is obtained by deleting $d(j - 1)$ from the column of $A^{-1}(j - 1)$ indexed by i_j

2. If $j = \ell$, stop; otherwise, set

$$A^{-1}(j) = D(j - 1) - \frac{1}{d(j - 1)} r(j - 1)r(j - 1)^T$$

where

$d(j - 1)$ and $r(j - 1)$ are defined above

$D(j - 1)$ is the matrix obtained by deleting the i_j*th* row and column from $A^{-1}(j - 1)$

c. Set $u^* = u(\ell)$.

In the third procedure, it is assumed that one has available the upper-triangular Cholesky factor R satisfying $R^TR = A$ and the vector z satisfying $R^Tz = v$. Of course, u is determined by the equation $Ru = z$, which is in upper-triangular form. The third procedure is based on the fact that, from R and z, one can obtain in an relatively efficient and stable manner a $(k - \ell)$-dimensional equation in upper-triangular form whose solution is u*. Indeed, if $\hat{R}$ denotes the $k \times (k - \ell)$ matrix obtained by deleting columns $i_1, \cdots, i_\ell$ from R, then $\hat{R}^T\hat{R} = \hat{A}$, $\hat{R}^Tz = \hat{v}$, and $\hat{R}^T\hat{R}u^* = \hat{v} = \hat{R}^Tz$. Now one can find a $k \times k$ orthogonal matrix P for which $\hat{R} = P\binom{R^*}{0}$, where R* is a $(k - \ell) \times (k - \ell)$ upper-triangular matrix. (An efficient and stable way to obtain P is by composing appropriate Householder transformations. See ref. 7 for details.) Then $\hat{R}^T\hat{R} = R^{*T}R^*$, and by setting $P^Tz = \binom{z^*}{z^{**}}$, where $z^*\varepsilon R^{k-\ell}$, one obtains

$$R^{*T}R^*u^* = \binom{R^*}{0}^T P^Tz = R^{*T}z^*$$

It follows that u* is given by the equation $R^*u^* = z^*$, which is in upper-triangular form.

PROCEDURE 3:

a. Form $\hat{R}$ by deleting the columns of R indexed by i_j; $j = 1, \cdots, \ell$.

b. Obtain a $k \times k$ orthogonal matrix P for which the factorization $\hat{R} = P\binom{R^*}{0}$ holds, where R* is a $(k - \ell) \times (k - \ell)$ upper-triangular matrix; set $P^Tz = \binom{z^*}{z^{**}}$, where $z^*\varepsilon R^{k-\ell}$.

c. Obtain u* by solving the upper-triangular system $R^*u^* = z^*$.

5.2 THE THREE METHODS

To demonstrate the relation of the algebraic problem just described to the problem of obtaining the MSE-optimal LDF or the derived map appropriate for an incomplete data vector, recall that the MSE-optimal LDF L for vectors in R^n is given by $L = a^T\Phi$, where $\Phi(x) = \binom{1}{x}$ and a is given by eq. (3). Since this is true for general n, one sees that if $\hat{x}\varepsilon R^{n-\ell}$ denotes the vector obtained by deleting components $i_1, \cdots, i_\ell$ from a vector $x\varepsilon R^n$, then the MSE-optimal LDF appropriate for $\hat{x}$, denoted by $\hat{L}$, is given by

$$\hat{L} = a^{*T}\hat{\Phi}$$

where

$$\hat{\Phi}(\hat{x}) = \begin{pmatrix} 1 \\ \hat{x} \end{pmatrix}$$

and a* satisfies

$$\hat{S}a^* = \hat{b} \tag{15}$$

In this expression,

$$\hat{b} = \sum_{k=1}^{N} \hat{\Phi}(\hat{x}_k)\beta_k$$

and

$$\hat{S} = \sum_{k=1}^{N} \hat{\Phi}(\hat{x}_k)\hat{\Phi}(\hat{x}_k)^T$$

For convenience, write

$$b = \begin{bmatrix} b_0 \\ b_1 \\ \cdot \\ \cdot \\ \cdot \\ b_n \end{bmatrix}$$

and

$$S = \begin{bmatrix} S_{00} & S_{01} & \cdots & S_{0n} \\ S_{10} & S_{11} & \cdots & S_{1n} \\ \cdot & \cdot & & \\ \cdot & \cdot & & \\ \cdot & \cdot & & \\ S_{n0} & S_{n1} & \cdots & S_{nn} \end{bmatrix}$$

It is clear that $\hat{b}$ is obtained from b by deleting from b the components indexed by i_j for $j = 1, \cdots, \ell$; similarly, $\hat{S}$ is obtained from S by deleting from S the rows and columns of S indexed by i_j for $j = 1, \cdots, \ell$. Thus the problem of determining $\hat{L}$ is fundamentally that of obtaining the solution a* of eq. (15), which is an algebraic problem of the type just described.

Similarly, taking eq. (12) as the definition of the derived map for general n, observe that the derived map for $\hat{x}$, denoted by $\hat{L}'$, is given by

$$\hat{L}'(\hat{x}) = \frac{1}{||\hat{m}_1 - \hat{m}_2||^2}\left[||\hat{m}_2||^2 - ||\hat{m}_1||^2 + 2(\hat{m}_1 - \hat{m}_2)^T\hat{S}_W^{-1}\hat{x}\right] \qquad (16)$$

where

$$\hat{m}_i = \frac{1}{N_i}\sum_{x_k \varepsilon X_i} \hat{x}_k \quad ; \quad i = 1, 2$$

$$\hat{S}_W = \sum_{x_k \varepsilon X} \hat{x}_k\hat{x}_k^T - N_1\hat{m}_1\hat{m}_1^T - N_2\hat{m}_2\hat{m}_2^T$$

and the norm $||\ ||$ is now given by

$$||\hat{u}||^2 = \hat{u}^T\hat{S}_W^{-1}\hat{u} \text{ for } \hat{u}\varepsilon R^{n-\ell}$$

Clearly, $\hat{m}_1$ and $\hat{m}_2$ are obtained from m_1 and m_2, respectively, by deleting components $i_1, \cdots, i_\ell$ from m_1 and m_2; similarly, $\hat{S}_W$ is obtained from S_W by deleting rows and columns $i_1, \cdots, i_\ell$ from S_W. Since $\hat{L}'$ is easily specified once the vectors $\hat{S}_W^{-1}\hat{m}_1$ and $\hat{S}_W^{-1}\hat{m}_2$ are known, one sees that the problem of determining $\hat{L}'$ is fundamentally that of solving the equation

$$\hat{S}_W u_i^* = \hat{m}_i \quad ; \quad i = 1, 2 \qquad (17)$$

Once the role played by the algebraic problem of section 5.1 in the determination of $\hat{L}$ and $\hat{L}'$ has been established, three methods of obtaining these maps are immediately suggested. These methods can be formulated as follows.

THE STRAIGHTFORWARD METHOD OF OBTAINING $\hat{L}$:

a. Recall b and S from storage.

b. Determine $a^* = \hat{S}^{-1}\hat{b}$ by using Procedure 1 to solve eq. (15), taking $k = n + 1$, $v = b$, $A = S$, and $u^* = a^*$.

c. Obtain $\hat{L} = a^{*T}\hat{\Phi}$.

THE STRAIGHTFORWARD METHOD FOR OBTAINING $\hat{L}'$:

a. Recall m_1, m_2, and S_W from storage.

b. For $i = 1, 2$, determine $u_i^* = \hat{S}_W^{-1}\hat{m}_i$ by using Procedure 1 to solve eq. (17), taking $k = n$, $v - m_i$, $A = S_W$ and $u^* = u_i^*$.

c. Using $u_i^* = \hat{S}_W^{-1}\hat{m}_i$, $i = 1, 2$, determine $||\hat{m}_1||^2$, $||\hat{m}_2||^2$, $\hat{S}_W^{-1}(\hat{m}_1 - \hat{m}_2)$, and $||\hat{m}_1 - \hat{m}_2||^2$.

d. Obtain $\hat{L}'$, as given by eq. (16).

THE KITTLER METHOD FOR OBTAINING $\hat{L}$:

a. Recall $a = S^{-1}b$ and S^{-1} from storage.

b. Determina $a^* = \hat{S}^{-1}\hat{b}$ by using Procedure 2 to solve eq. (15), taking $k = n + 1$, $u = a$, $A^{-1} = S^{-1}$, and $u^* = a^*$.

c. Obtain $\hat{L} = a^{*T}\hat{\Phi}$.

THE KITTLER METHOD FOR OBTAINING $\hat{L}'$:

a. Recall $S_W^{-1}m_1$, $S_W^{-1}m_2$, and S_W^{-1} from storage.

b. For $i = 1, 2$, determine $u_i^* = \hat{S}_W^{-1}\hat{m}_i$ by using Procedure 2 to solve eq. (17), taking $k = n$, $u = S_W^{-1}m_i$, $A^{-1} = S_W^{-1}$, and $u^* = u_i^*$.

c. Using $u_i^* = \hat{S}_W^{-1}\hat{m}_i$, $i = 1, 2$, determine $||\hat{m}_1||^2$, $||\hat{m}_2||^2$, $\hat{S}_W^{-1}(\hat{m}_1 - \hat{m}_2)$, and $||\hat{m}_1 - \hat{m}_2||^2$.

d. Obtain $\hat{L}'$, as given by eq. (16).

THE GOLUB METHOD FOR OBTAINING $\hat{L}$:

a. Recall from storage the upper-triangular Cholesky factor R of S and the solution z of $R^T z = b$.

b. Determine $a^* = \hat{S}^{-1}\hat{b}$ by using Procedure 3 to solve eq. (15), taking $k = n + 1$ and $u^* = a^*$.

c. Obtain $\hat{L} = a^{*T}\hat{\Phi}$.

THE GOLUB METHOD FOR OBTAINING $\hat{L}'$:

a. Recall from storage the upper-triangular Cholesky factor R of S_W and the solutions z_i of $R^T z_i = m_i$ for $i = 1, 2$.

b. For $i = 1, 2$, determine $u_i^* = \hat{S}_W^{-1}\hat{m}_i$ by using Procedure 3 to solve eq. (17), taking $k = n$, $z = z_i$, and $u^* = u_i^*$.

c. Using $u_i^* = \hat{S}_W^{-1}\hat{m}_i$, $i = 1, 2$, determine $||\hat{m}_1||^2$, $||\hat{m}_2||^2$, $\hat{S}_W^{-1}(\hat{m}_1 - \hat{m}_2)$, and $||\hat{m}_1 - \hat{m}_2||^2$.

d. Obtain $\hat{L}'$, as given by eq. (16).

5.3 RELATIVE ADVANTAGES OF THE THREE METHODS

In discussing the relative advantages of the three methods formulated in section 5.2, the focus is on efficiency and accuracy. None of the three methods presents any particularly stringent requirements of storage or preparatory computation.

The efficiency of a method is usually reflected in the number of arithmetic operations required to implement it. Since the relative numbers of arithmetic operations required by the three methods are determined directly by the relative numbers of arithmetic operations required by the three procedures in section 5.1, consider the arithmetic operations necessary to implement these procedures. In particular, since the number of additions required by procedures of this type is approximately the same as the number of multiplications required, the following formulas are offered which specify the numbers of multiplications required by Procedures 1, 2, and 3, respectively, for any k. (When these procedures are used in the methods of section 5.2, k is either n or (n + 1).)

$$\frac{(k - \ell)^3}{6} + \frac{3}{2}(k - \ell)^2 + \frac{1}{3}(k - \ell) \tag{18}$$

$$\frac{(\ell - 1)(k - \ell)^2}{2} + \frac{(\ell - 1)^2(k - \ell)}{2} + \frac{(\ell - 1)^3}{6} + 3(\ell - 1)(k - \ell) + \frac{3}{2}(\ell - 1)^2 + (k - \ell) + \frac{7}{3}(\ell - 1) + 1 \tag{19}$$

$$\frac{3}{2}(\ell - 1)(k - \ell)^2 + \frac{9}{2}(k - \ell)^2 + \frac{9}{2}(\ell - 1)(k - \ell) + \frac{25}{2}(k - \ell) \tag{20}$$

It is apparent from the highest-order terms in these formulas that if k is very large and if ℓ is small relative to k, then the Kittler methcd requires the fewest multiplications of the three, followed by the Golub method and the straightforward method in that order. In fact, in these circumstances, if $\ell = 1$, then the numbers of multiplications required by the three methods are $O(k)$, $O(k^2)$, and $O(k^3)$, respectively. If $\ell > 1$ but is still small relative to k, then both the Kittler method and the Golub method require $O(k - \ell)^2$ multiplications, while the straightforward method requires $O(k - \ell)^3$ multiplications. If $\ell = 2$, then the Kittler method requires fewer multiplications than the Golub method by about a factor of 12; however, as ℓ grows, this advantage quickly drops to a factor slightly greater than 3.

If k is not too large or if the size of ℓ is significant relative to k, then the order of the numbers of multiplications required by the three methods changes. To illustrate the variance of this order with different values of k and ℓ, see table I, in which the order is indicated on the left ("K," "G," and "S" represent "Kittler," "Golub," and "straightforward," respectively), and the values of k are listed across the top. The lower right number in each entry is the value of ℓ for which the order on the left first occurred with the value of k given above; the upper left number in each entry is the fraction ℓ/k. It should be noted that, except for fairly small k, the orders (K, G, S), (K, S, G), (S, K, G), and (S, G, K) appear in order as ℓ ranges from 1 to (k - 1). Furthermore, if k is large, then these orders appear at fairly well-specified values of ℓ/k. Specifically, the order (K, S, G) replaces the order (K, G, S) when ℓ is about 10 percent of k; the order (S, K, G) replaces the order (K, S, G) when ℓ is about 21 percent of k; and the order (S, G, K) replaces the order (S, K, G) when ℓ is about 58 percent of k. Figures 1 through 5 indicate for five different values of k the relative sizes of the numbers of multiplications required by the three methods as ℓ ranges from 1 to (k - 1).

In gaging the accuracy of a method, one should consider not only the number of arithmetic operations required by the method but also the stability of the method. Loosely speaking, a method is said to be stable if small errors introduced in the course of the computation do not compound themselves to an unreasonable degree as the computation proceeds; otherwise, the method is said to be unstable. Somewhat more strictly speaking, stable methods generally exhibit error growth which is linear in the number of arithmetic operations performed, while the error growth associated with unstable methods is exponential in the number of arithmetic operations performed. To quote reference 8, "linear growth is normal, usually unavoidable, and not dangerous; exponential growth may, however, be disastrous and should be avoided at all costs."

The instability of a method may become especially serious when there are bad features inherent in a particular problem to which the method is applied. Here, the basic problem under consideration is that posed in section 5.1, that of solving the linear equation $\hat{A}u^* = \hat{v}$ using information about the equation $Au = v$. A linear equation involving a positive-definite symmetric matrix is said to be ill-conditioned if the condition number, defined to be the ratio of the largest eigenvalue to the smallest[1], is large. The practical

[1] For a definition of the condition number for a general linear system, see reference 6.

importance of ill-conditioning is that small errors incurred in computing the approximate solution of an ill-conditioned linear equation may result in large errors in the approximate solution. Thus, procedures for solving an ill-conditioned linear equation must be chosen with care in order that they yield approximate solutions which are meaningful. In particular, unstable procedures should be avoided in solving ill-conditioned linear equations.

The linear equations (15) and (17), which must be solved by the procedures of section 5.1 to obtain $\hat{L}$ and $\hat{L}'$, are likely to be ill-conditioned in many applications. (The same is true of the "parent" equations $Sa = b$ and $S_W u_i = m_i$, $i = 1, 2$.) Indeed, the matrices $\hat{S}$ and $\hat{S}_W$ will be ill-conditioned if the "incomplete" labeled training vectors $\hat{x}_k$, $k = 1, \cdots, N$, "nearly" lie in some proper subspace of $R^{n-\ell}$. This will occur, for example, if two or more of the components of the "incomplete" vector random variable $\hat{X}$ are highly correlated. Consequently, potentially unstable procedures should be used in the solution of eqs. (15) and (17) only if it is ascertained that these equations (and, perhaps, their "parent" equations as well) are not too ill-conditioned.

Procedures 1 and 3 of section 5.1 are known to be stable. Indeed, Procedure 1 involves only Cholesky decomposition and the solution of two triangular systems. Both Cholesky decomposition and the solution of triangular systems are stable, and the latter can usually be carried out with a high degree of accuracy (ref. 7). In Procedure 3, the formation of the initial Cholesky factor R of A is stable, as is the solution of the upper-triangular system $R^*u^* = z^*$. Also, it was observed earlier that the factorization $\hat{R} = P\binom{R^*}{0}$ can be obtained in an efficient and stable way by constructing P as a composition of Householder transformations (ref. 9).

On the other hand, Procedure 2 appears to be potentially unstable. The basis of Procedure 2 is the repeated application of the matrix-inverse-update formula of step 2, which is derived from the Sherman-Morrison formula (ref. 9), for updating the inverse of a matrix following a rank 1 change. The Sherman-Morrison formula is known to exhibit instability in many applications.

Indeed, the potential for difficulties in Procedure 2 is apparent in the formula of step 2: if the subtraction called for by the formula is subject to numerical error, the resulting approximate inverse can actually fail to be positive-definite! If only one component of a data vector is missing ($\ell = 1$), then the inverse is not updated and the potential instabilities do not materialize.

The salient points of this discussion are summarized and conclusions are drawn in the following:

a. The straightforward method is always stable. It is preferred over the other two methods for reasons of efficiency as well as accuracy when more than about 21 percent of the vector components are missing. It is more efficient than the Golub method when more than about 10 percent of the components are missing. If the number of components of complete vectors is not large (no greater than 30 to 40) then the straightforward method is competitive in efficiency with the Golub method for any number of missing components.

b. The Kittler method is more efficient than either the straightforward method or the Golub method when no more than about 21 percent of the vector components are missing. However, because of the potential instabilities inherent in this method, it should be used only when speed of computation is of overriding concern and eqs. (15) and (17) and their "parent" equations are well-conditioned. When only one component is missing, the Kittler method can be safely used and results in considerable savings in computation time.

c. The Golub method is always stable. As the number of.components of complete vectors becomes greater than 40, it offers increasingly significant benefits in efficiency over the straightforward method when no more than about 10 percent of the vector components are missing. After more than a few components are missing, the advantage in efficiency of the Kittler method over the Golub method drops to a factor slightly greater than 3. The Golub method becomes more efficient then the Kittler method after about 58 percent of the components are missing. Of course, the straightforward method is by far the most efficient method at this point.

TABLE I.– VARIANCE OF ORDER

Order \ k	4	8	12	16	20	24	28	32	36
K, G, S						1	1	1	1
K, S, G	1	1	1	1	1	.08 2	.071 2	.063 2	.083 3
S, K, G	.5 2	.375 3	.333 4	.313 5	.300 6	.292 7	.286 8	.25 8	.25 9
S, G, K		.75 6	.75 9	.688 11	.65 13	.667 16	.643 18	.625 20	.639 23

Order \ k	48	60	72	84	96	108	120	132	144
K, G, S	1	1	1	1	1	1	1	1	1
K, S, G	.083 4	.083 5	.083 6	.095 8	.094 9	.093 10	.092 11	.091 12	.097 14
S, K, G	.25 12	.233 14	.236 17	.226 19	.229 22	.222 24	.225 27	.220 29	.222 32
S, G, K	.625 30	.617 37	.611 44	.607 51	.604 58	.602 65	.592 71	.591 78	.590 85

Order \ k	200	300	400	500	600	700	800	900	1000
K, G, S	1	1	1	1	1	1	1	1	1
K, S, G	.095 19	.097 29	.098 39	.098 49	.098 59	.099 69	.099 79	.099 89	.099 99
S, K, G	.215 43	.213 64	.21 84	.21 105	.21 126	.209 146	.209 167	.209 188	.208 208
S, G, K	.59 118	.587 176	.583 233	.582 291	.582 349	.581 407	.581 465	.581 523	.581 581

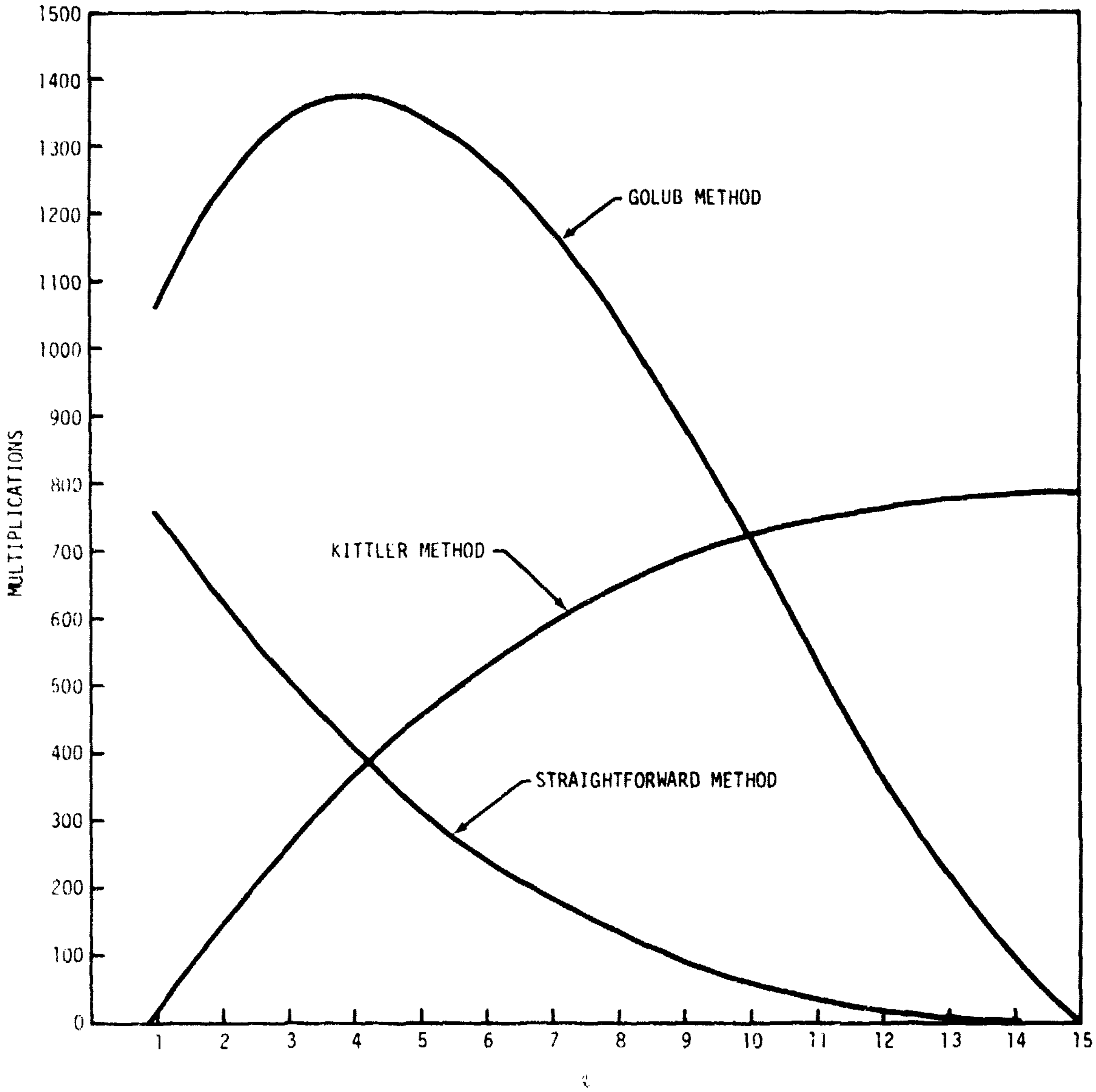

Figure 1.– Comparison of multiplications when k = 15.

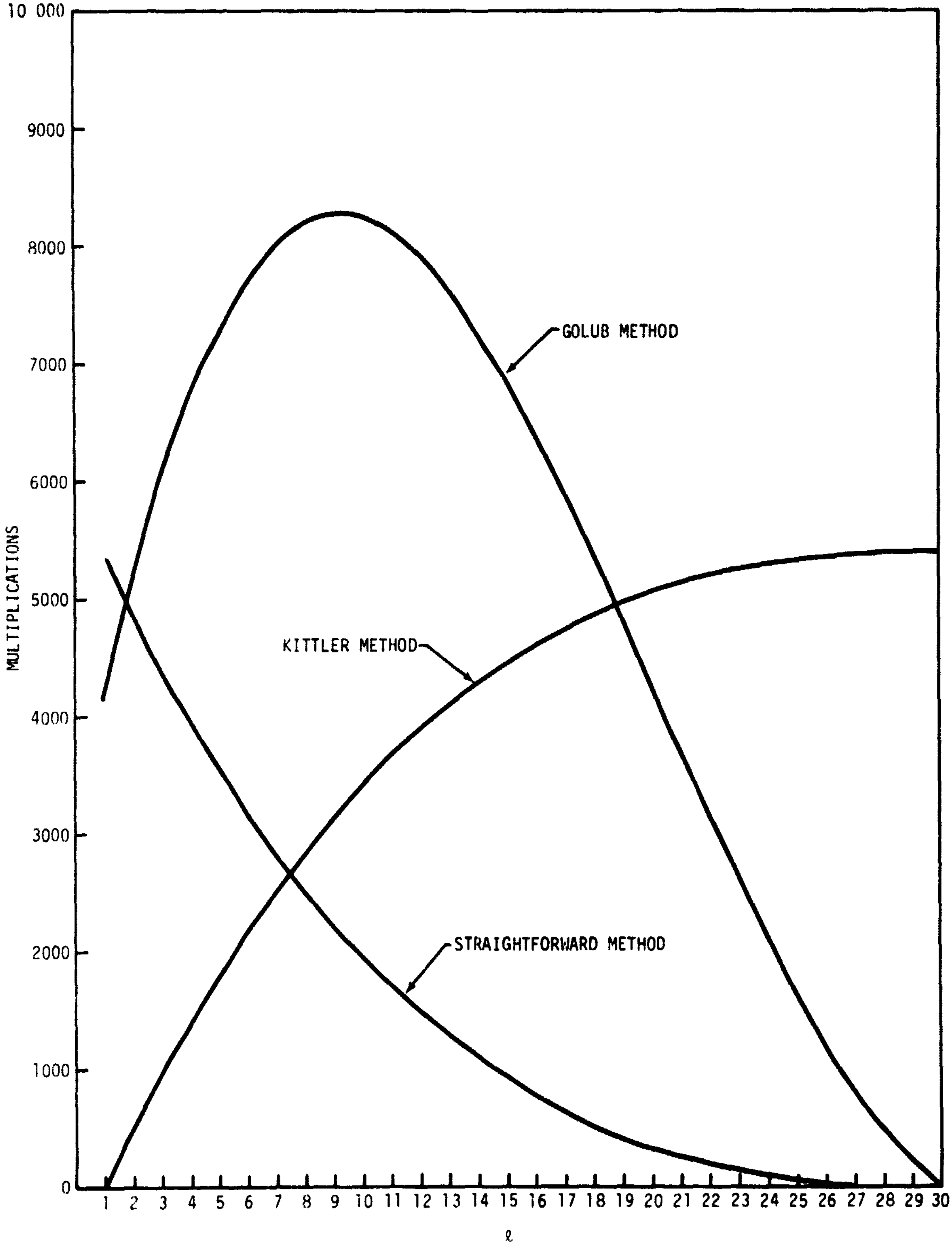

Figure 2.– Comparison of multiplications when k = 30.

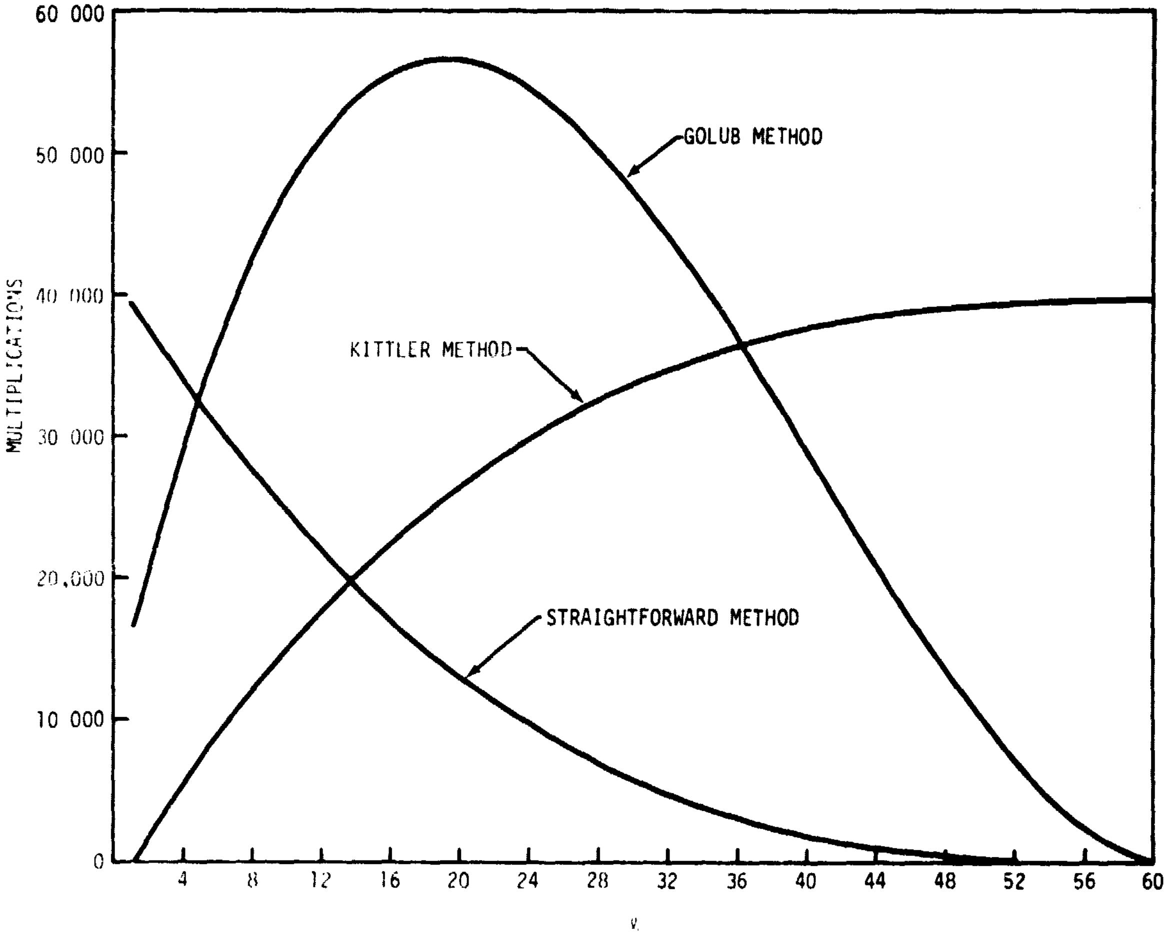

Figure 3.— Comparison of multiplications when k = 60.

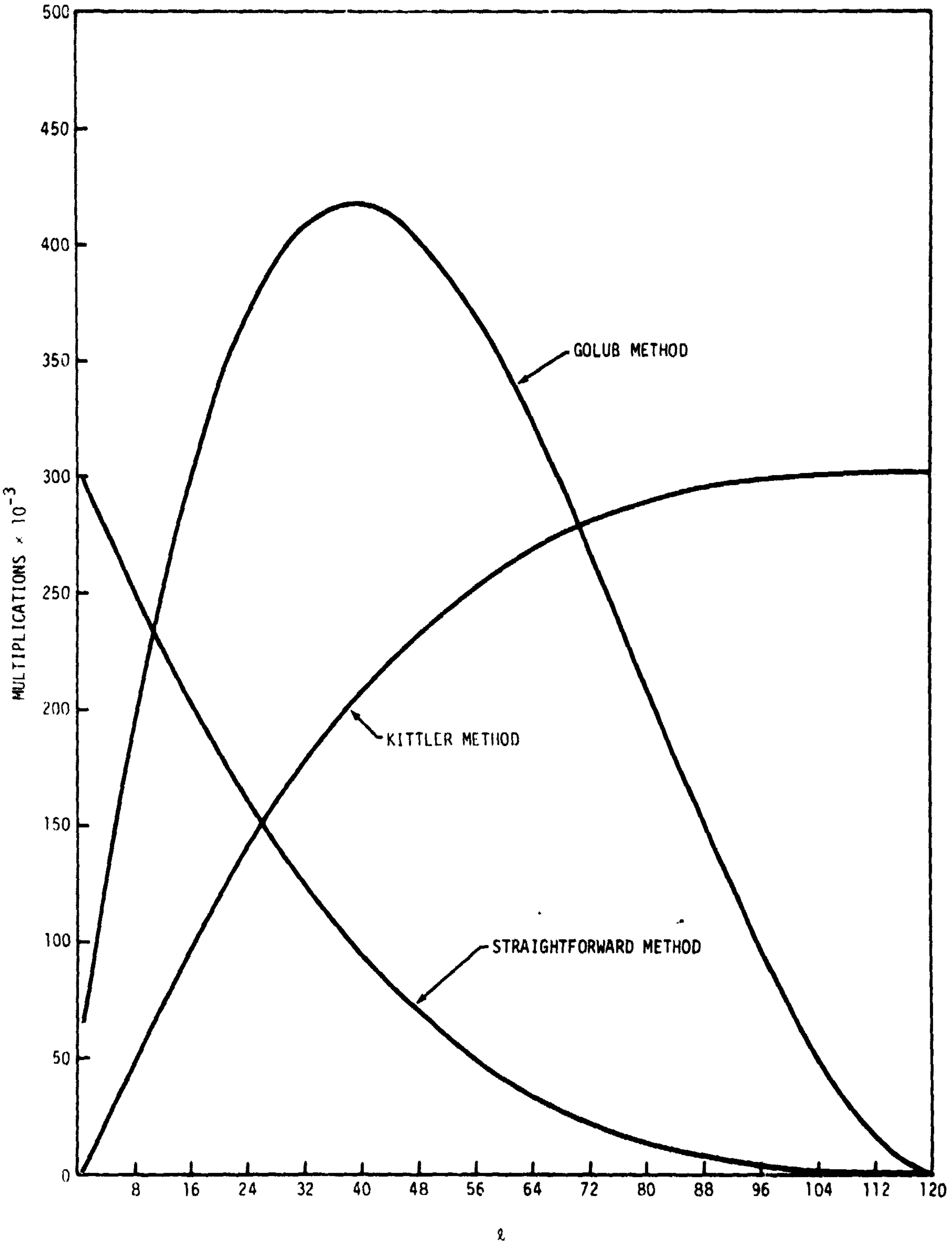

Figure 4.— Comparison of multiplications when k = 120.

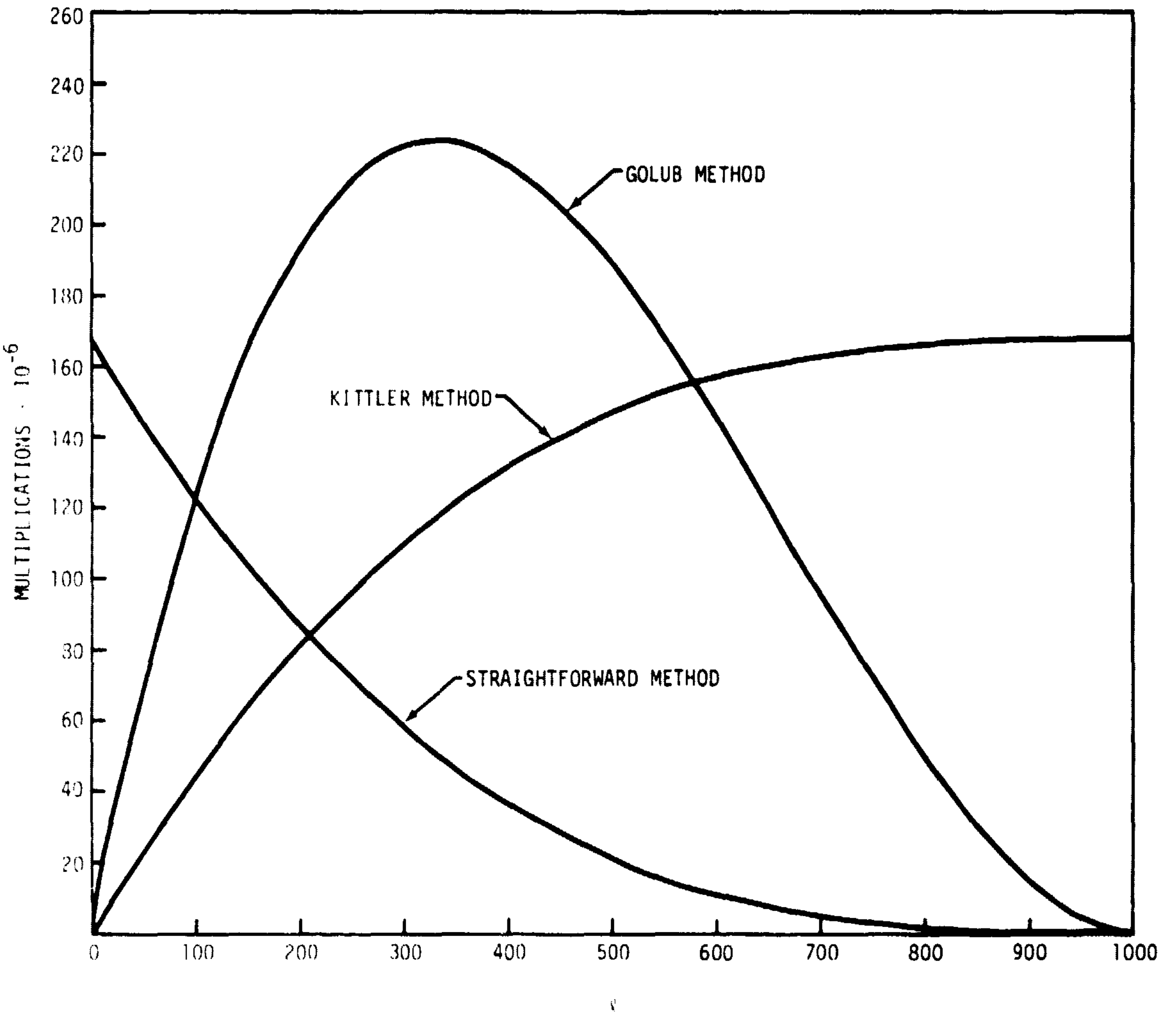

Figure 5.– Comparison of multiplications when k = 1000.

6. REFERENCES

1. Duda, R. O.; and Hart, P. E.: Pattern Classification and Scene Analysis. John Wiley and Sons, Inc. (New York), 1973.

2. Kittler, J.: Classification of Incomplete Pattern Vectors Using Modified Discriminant Functions. IEEE Trans. Comput., Vol. C-27, April 1978, pp. 367-375.

3. Loéve, M.: Probability Theory. D. Van Nostrand Co. (New York), 1963.

4. Peters, B. C.; and Walker, H. F.: An Iterative Procedure for Obtaining Maximum-Likelihood Estimates of the Parameters for a Mixture of Normal Distributions. SIAM J. Appl. Math. (unpublished).

5. Seber, G. A. F.: Linear Regression Analysis. John Wiley and Sons, Inc. (New York), 1977.

6. Stewart, G. W.: Introduction to Matrix Computations. Academic Press, Inc. (New York), 1973.

7. Golub, G. H.; and Styan, G. P. H.: Numerical Computations for Univariate Linear Models. J. Statist. Comput. Simul., vol. 2, 1973, pp. 253-274.

8. Conte, S. D.; and de Boor, C.: Elementary Numerical Analysis: An Algorithmic Approach. McGraw-Hill, Inc. (New York) 1972.

9. Sherman, J.; and Morrison, W. J.: Adjustment of an Inverse Matrix Corresponding to Changes in the Elements of a Given Column or a Given Row of the Original Matrix. Ann. Math. Statist., vol. 20, 1949, p. 621.