72

JSC-12694

79-10151

CR151886

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

Ref: 642-7138
Contract NAS 9-15200
Job Order 73-~~705~~-39
743

TECHNICAL MEMORANDUM

SAMPLE-MOMENT ESTIMATES FOR THE MEAN VECTOR AND THE COVARIANCE MATRIX FOR A SAMPLE OF INCOMPLETE DATA VECTORS

By

H. F. Walker

(E79-10151) SAMPLE-MOMENT ESTIMATES FOR THE MEAN VECTOR AND THE COVARIANCE MATRIX FOR A SAMPLE OF INCOMPLETE DATA VECTORS (Lockheed Electronics Co.) 8 p HC A02/MF A01 CSCL 12A

N79-18626

Unclas
G3/63 00151

Approved By: B L Carroll
for T. C. Minter, Supervisor
Techniques Development Section

August 1978

LEC-12771

SAMPLE-MOMENT ESTIMATES OF THE MEAN VECTOR AND THE COVARIANCE MATRIX FOR A SAMPLE OF INCOMPLETE DATA VECTORS

1. INTRODUCTION

In this memorandum, it is supposed that X is a random variable in n-dimensional Euclidean space (R^n) and that one must estimate the mean vector $\mu = E(X)$ and the covariance matrix $\Sigma = E(X - \mu)(X - \mu)^T$ on the basis of a given independent sample $\chi = \{x_k\}_{k=1,\cdots,N}$ of the observations on X. Under normal circumstances, a satisfactory solution of this frequently encountered estimation problem is given by the sample-moment estimates

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

and

$$S = \frac{1}{N}\sum_{k=1}^{N} (x_k - m)(x_k - m)^T \tag{2}$$

The estimate m is unbiased and consistent.[1] The estimate S is consistent, although it is biased. In fact, $E(S) = \frac{N - 1}{N}\Sigma$; hence, $\frac{N}{N - 1} S$ is an unbiased estimate of Σ. If X is normally distributed, then m and S are maximum-likelihood estimates of μ and Σ, respectively. The estimates m and S have the numerically attractive property that they can be evaluated with one pass through the observations in χ.

This estimation problem requires special treatment if a given sample of observations contains vectors which are incomplete in the sense that some of their components are unknown or missing. In the following section, expressions are offered for sample-moment estimates of the components of μ and Σ determined by those sample observations for which the appropriate components are known. As with the estimates given by eqs. (1 and 2) for complete data vectors, the

[1]For convenience, estimates and their associated estimators are identified.

resulting estimate of μ is unbiased and consistent, and the resulting estimate of Σ is consistent, although biased. (As with the estimate given by eq. (2), the bias is removable.) A procedure is then outlined for evaluating these estimates which requires only one pass through the observations in χ.

2. ESTIMATES OF μ AND Σ

Assume that the given independent sample $\chi = \{x_k\}_{k=1,\cdots,N}$ of observations on X contains vectors which are incomplete in the above sense, and denote the i*th* component of μ by μ_i and the ij*th* entry of Σ by σ_{ij} $(1 \leq i, j \leq n)$. Expressions are derived below for sample-moment estimates of μ_i and σ_{ij} determined by those observations in χ for which the appropriate components are known.

To obtain the expression for the sample-moment estimate of μ_i, let x_{ki} denote the i*th* component of x_k $(1 \leq k \leq N)$ and let χ_i be the set of observations x_k in χ for which x_{ki} is known. Denoting the number of observations in χ_i by N_i, the expression

$$m_i = \frac{1}{N_i} \sum_{x_k \varepsilon \chi_i} x_{ki} \tag{3}$$

is immediately obtained as the desired sample-moment estimate of μ_i. The estimate m_i is clearly unbiased and consistent.

To obtain the expression for the sample-moment estimate of σ_{ij}, set $\chi_{ij} = \chi_i \cap \chi_j$ and denote by N_{ij} the number of observations in χ_{ij}. The expression

$$s_{ij} = \frac{1}{N_{ij}} \sum_{x_k \varepsilon \chi_{ij}} (x_{ki} - m_i)(x_{kj} - m_j) \tag{4}$$

is offered as the desired sample-moment estimate of σ_{ij}. In this expression, m_i and m_j are, respectively, the estimates of μ_i and μ_j given (mutatis mutandis) by eq. (3). The estimate s_{ij} is easily seen to be consistent, although it is biased. Indeed,

$$E(s_{ij}) = \left\{1 - \frac{1}{N_i} - \frac{1}{N_j} + \frac{N_{ij}}{N_i N_j}\right\}\sigma_{ij}$$

Note that the bias of s_{ij} is removable in the sense that one may easily obtain from s_{ij} the unbiased estimate

$$\left\{1 - \frac{1}{N_i} - \frac{1}{N_j} + \frac{N_{ij}}{N_i N_j}\right\}^{-1} s_{ij}$$

An alternate sample-moment expression is

$$s_{ij} = \frac{1}{N_{ij}} \sum_{x_k \varepsilon X_{ij}} \left[x_{ki} - m_i^{(j)}\right]\left[x_{kj} - m_j^{(i)}\right] \tag{5}$$

where

$$m_i^{(j)} = \frac{1}{N_{ij}} \sum_{x_k \varepsilon X_{ij}} x_{ki}$$

and

$$m_j^{(i)} = \frac{1}{N_{ij}} \sum_{x_k \varepsilon X_{ij}} x_{kj}$$

One might consider the estimate given by eq. (4) to be more appealing than that given in eq. (5). For one thing, if s_{ij} is given by eq. (5), then

$$E(s_{ij}) = \left(1 - \frac{1}{N_{ij}}\right)\sigma_{ij}$$

Since

$$1 - \frac{1}{N_{ij}} \leq 1 - \frac{1}{N_i} - \frac{1}{N_j} + \frac{N_{ij}}{N_i N_j} < 1$$

it follows that the bias of the estimate given by eq. (4) is less than that of the estimate given by eq. (5). Of course, this is not a very serious consideration, since the bias of both estimates is removable. A more serious consideration is that the estimate given by eq. (4) seems likely to have smaller variance than that given by eq. (5). This has not been proved; however, the following heuristic arguments are offered: Since m_i and m_j have smaller variances than $m_i^{(j)}$ and $m_j^{(i)}$, respectively, m_i and m_j should contribute less to the variance of the sum in eq. (4) than $m_i^{(j)}$ and $m_j^{(i)}$ contribute to the variance of the sum in eq. (5). Also, if X is normally distributed and μ_i and μ_j are known, then the maximum-likelihood estimate of σ_{ij}, based on χ_{ij}, is

$$\hat{\sigma}_{ij} = \frac{1}{N_{ij}} \sum_{x_k \varepsilon \chi_{ij}} (x_{ki} - \mu_i)(x_{kj} - \mu_j)$$

This estimate is unbiased and, therefore, has minimum variance among all unbiased estimates. Since m_i and m_j have smaller variances than $m_i^{(j)}$ and $m_j^{(i)}$, respectively, the estimate given by eq. (4) should differ less than that given by eq. (5) from the minimum-variance estimator $\hat{\sigma}_{ij}$.

3. A ONE-PASS EVALUATION PROCEDURE

A procedure is now outlined for evaluating the sample-moment estimates given by eqs. (3) and (4) for all i and j between 1 and n. This procedure requires only one pass through the observations in the sample set χ. Assume that the sets χ_i and χ_j have been determined, and rewrite eq. (4) as

$$s_{ij} = \frac{1}{N_{ij}} \sum_{x_k \varepsilon \chi_{ij}} x_{ki}x_{kj} - m_i m_j^{(i)} - m_i^{(j)} m_j + m_i m_j \qquad (6)$$

Since $s_{ij} = s_{ji}$, it is necessary to evaluate the quantities s_{ij} only for $j \geq i$. Note that

$$s_{ii} = \frac{1}{N_i} \sum_{x_k \varepsilon X_i} x_{ki}^2 - m_i^2 \tag{7}$$

For convenience in describing the procedure, set

$$T_{ij} = \frac{1}{N_{ij}} \sum_{x_k \varepsilon X_{ij}} x_{ki} x_{kj}$$

PROCEDURE:

1. Initialize

   (a) $N_i = 0$, $m_i = 0$ for $1 \leq i \leq n$

   (b) $T_{ij} = 0$ for $1 \leq i \leq j \leq n$

   (c) $N_{ij} = 0$, $m_i^{(j)} = 0$, $m_j^{(i)} = 0$ for $1 \leq i < j \leq n$

2. Carry out the algorithm described in figure 1. This algorithm produces m_i and T_{ii} for $1 \leq i \leq n$ and $m_i^{(j)}$, $m_j^{(i)}$, and T_{ij} for $1 \leq i < j \leq n$.

3. For $1 \leq i \leq n$, obtain m_i as a direct output of the algorithm and obtain s_{ii} from eq. (7). For $1 \leq i < j \leq n$, obtain s_{ij} from eq. (6).

For $k = 1, \cdots, N$, do the following:

For $i = 1, \cdots, n$, do the following:

If $x_k \varepsilon X_i$, continue; otherwise, return.

Do the following updates:

$$N_i = N_i + 1$$

$$m_i = \frac{N_i - 1}{N_i} m_i + \frac{1}{N_i} x_{ki}$$

$$T_{ii} = \frac{N_i - 1}{N_i} T_{ii} + \frac{1}{N_i} x_{ki}^2$$

If $i < n$, continue; otherwise, return.

For $j = i + 1, \cdots, n$, do the following:

If $x_k \varepsilon X_{ij}$, continue; otherwise, return.

Do the following updates:

$$N_{ij} = N_{ij} + 1$$

$$m_i^{(j)} = \frac{N_{ij} - 1}{N_{ij}} m_i^{(j)} + \frac{1}{N_{ij}} x_{ki}$$

$$m_j^{(i)} = \frac{N_{ij} - 1}{N_{ij}} m_j^{(i)} + \frac{1}{N_{ij}} x_{kj}$$

$$T_{ij} = \frac{N_{ij} - 1}{N_{ij}} T_{ij} + \frac{1}{N_{ij}} x_{ki} x_{kj}$$

Figure 1.— Algorithm.