A NOTE ON CLASSIFICATION

ASHOK K. SAXENA
Lucknow University

Consider the multivariate complex Gaussian distribution Π_j $(j = 1, 2)$ as defined by Goodman [3].

$$\Pi_j: p_j(\xi) = \Pi^{-p} |\Sigma|^{-1} \exp [-(\overline{\xi - \mu_j})'\Sigma^{-1}(\xi - \mu_j)],$$

where $E(\xi) = \mu_j$ and cov $\xi = \Sigma$ (Hermitian positive definite complex covariance matrix). Any observation ξ will be a point in the space R^2, where R is the p-dimensional space. Partition R^2 into two subspaces R_1 and R_2 such that R_j identifies Π_j. Now if q_j is the a priori probability of drawing ξ from Π_j, the conditional probability (after the individual is drawn) of the same will be $q_j p_j(\xi)/\sum_{j=i}^{2} q_j p_j(\xi)$. The expected loss to be minimized (which is also the probability of misclassification when the costs of misclassification are unity) is

$$q_1 \int_{R_2} p_1(\xi)\, d\xi + q_2 \int_{R_1} p_2(\xi)\, d\xi.$$

Then the Bayes solution, which consists of assigning the individual to the population with higher conditional probability, gives the subspaces as

$$R_1: q_1 p_1(\xi) > q_2 p_2(\xi),$$

and

$$R_2: q_1 p_1(\xi) \leqq q_2 p_2(\xi).$$

When the costs of misclassification are not unity, these will be modified as

$$R_1: [q_1 C(2/1)]p_1(\xi) > [q_2 C(1/2)]p_2(\xi),$$

$$R_2: [q_1 C(2/1)]p_1(\xi) \leqq [q_2 C(1/2)]p_2(\xi).$$

Where $C(j/i)$; $(i, j = 1, 2)$ is the cost of misclassification of the individual from the ith population. Using Π_j as defined above

$$R_1: U > \log k$$

$$R_2: U \leqq \log k.$$

Where $U = \bar{\xi}'\Sigma^{-1}(\mu_1 - \mu_2) + (\overline{\mu_1 - \mu_2})'\Sigma^{-1}\xi - \bar{\mu}_1'\Sigma^{-1}\mu_1 + \bar{\mu}_2'\Sigma^{-1}\mu_2$ and k is a constant depending upon q_j and $C(j/i)$. It is easily seen that U is real valued. The distribution of U is ordinary univariate normal with $E(U) = (-)^{j+1}\nu$ and var $(U) = 2\nu$ where $\nu = (\overline{\mu_1 - \mu_2})'\Sigma^{-1}(\mu_1 - \mu_2')$, according as $\xi \varepsilon \Pi_j$, (see [1]). $\nu = (\overline{\mu_1 - \mu_2})'\Sigma^{-1}(\mu_1 - \mu_2)$ is termed as the "distance" between the two populations. If the parameters are estimated from sample of size N_j from Π_j, then

Received 3 February 1967; revised 26 May 1967.

the estimated "distance" will be[1]

$$(\overline{\alpha_1 - \alpha_2})' S^{-1} (\alpha_1 - \alpha_2).$$

This is simply related to the complex analogue of Hotelling's-T^2 (see [4]) given by

$$T_c^2 = N_1 N_2 (N_1 + N_2)^{-1} (\overline{\alpha_1 - \alpha_2})' S^{-1} (\alpha_1 - \alpha_2)$$

Further it is known that

$$(N_1 + N_2 - p - 1)/p \cdot T_c^2/(N_1 + N_2 - 2)$$

is distributed as a non-central F with degrees of freedom $[2p, 2(N_1 + N_2 - p - 1)]$ and non-centrality parameter

$$2N_1 N_2 (N_1 + N_2)^{-1} (\overline{\mu_1 - \mu_2})' \Sigma^{-1} (\mu_1 - \mu_2).$$

LEMMA. *T_c^2 is almost invariant in the space of sufficient statistics (α_1, α_2, S) under the full linear group G of $p \times p$ non-singular complex matrices under multiplication.*

For the proof the reader is referred to [1] and [2].

An extension to the case of more than two populations or when the sample consists of more than a single observation is obvious.

REFERENCES

[1] ANDERSON, T. W. (1958). *Introduction to Multivariate Statistical Analysis.* John Wiley & Sons.

[2] GIRI, N. (1965). On the complex analogue of T^2 and R^2-tests *Ann. Math. Statist.* **36**. 664–670.

[3] GOODMAN, N. R. (1963). Statistical analysis based on a certain multivariate complex Gaussian distribution (An Introduction). *Ann. Math. Statist.* **34**. 152–176.

[4] SAXENA, A. K. (1966). On the complex analogue of T^2 for two populations. *J. Indian Statist. Assoc.* **4**. 99–102.

[1] α_1 and α_2 are the maximum likelihood estimates of μ_1 and μ_2 and S is the standard unbiassed estimate of Σ based on $N_1 + N_2 - 2$ degrees of freedom.