N71-35524

RF Project 2241

Report No. Final

FINAL

REPORT

By

THE OHIO STATE UNIVERSITY
RESEARCH FOUNDATION

1314 KINNEAR RD.
COLUMBUS, OHIO 43212

To NATIONAL AERONAUTICAL & SPACE ADMINISTRATION
Office of University Affairs
Washington D.C. 20546

On RANKING PROBLEMS IN MULTIVARIATE NORMAL (STATISTICAL) POPULATIONS (First-year work)
NONPARAMETRIC RANGING AND SELECTION PROCEDURES (Second-year work)

For the period July 1, 1966 - September 30, 1969

Submitted by Dr. M. Haseeb Rizvi

Department of Mathematics

Date June 17, 1971

OPTIMAL SELECTION OF AUTOMATION SYSTEMS UNDER MULTIVARIATE NORMAL MODEL

M. Haseeb Rizvi

Stanford University, Stanford, California

0. SUMMARY AND APPLICATIONS

Suppose we have several (say $k \geq 2$) alternative automation systems $\Pi_i (i = 1,\ldots,k)$ and we are interested in selecting a certain number $t(< k)$ of best systems in terms of reliability, feasibility and economy; the case $t = 1$ corresponds to the selection of the best automation system. Let these k automation systems be operating under k independent p-variate normal distributions with column vector means μ_i and covariance matrices $\Sigma_i (i = 1,\ldots,k)$. Assume that the ranking criterion which incorporates the various considerations of reliability, feasibility and economy is given by the parametric function $\theta_i = \mu_i' \Sigma_i^{-1} \mu_i$ for $\Pi_i (i = 1,\ldots,k)$; thus we assume that the larger the θ-value of a system Π the better is the system. A typical parametric function θ represents the Mahalanobis distance between two p-variate normal distributions, one with p-vector mean μ and covariance matrix Σ and another with mean null p-vector and the same covariance matrix Σ. Mahalanobis distances are commonly employed for purposes of comparisons in multivariate analysis. Within this set-up we require a selection procedure R (optimal in the sense of economizing on the sample size to be used) which makes a correct selection with a probability no smaller than P^*, a preassigned quantity, wherever the t largest θ-values are (i) at least δ_1

larger than the rest of θ-values, and are simultaneously (ii) at least as large as δ_2 times the largest of the rest of θ-values. Like P^*, δ_1 and δ_2 are also specified in advance by the experimenter.

A selection procedure $R_1(R_2)$ is proposed for case 1 (case 2) when $\Sigma_1,\ldots,\Sigma_k$ are all known (unknown) and the common number of observations (needed from each of the k automation systems) is obtained so that the probability of a correct selection is no less than P^*. Some tables are provided for determination of the common sample size for various values of the constants involved.

1. INTRODUCTION

Alam and Rizvi [1] considered the problems of selection of the t largest non-centrality parameters of the k non-central chi-squared distributions as well as of the k non-central F distributions and obtained the mathematical results concerning the "least favorable configurations" of the parameter space (of k non-centrality parameters) within a specified parametric subspace. The least favorable configuration of the parameters is defined to be that configuration for which the probability of a correct selection for a given selection procedure is minimum. Thus the probability of a correct selection evaluated at the least favorable configuration of parameters can be obtained as an integral that depends on the common sample size n. This integral can then be equated to the pre-assigned probability P^* and a solution for n obtained. The ranking of k p-variate normal distributions in terms of Mahalanobis distance functions $\theta_i = \mu_i'\Sigma_i^{-1}\mu_i$ can be reduced to ranking of non-centrality parameters of k non-central chi-squared distributions (F distributions) if the selection procedure is based on the natural ordering of some statistic $nU_i(nV_i)$ from Π_i that has a non-central chi-squared (F) distribution with non-centrality parameter $n\theta_i$. Using this approach the present paper adapts the procedures of [1] for the selection of t best of the k automation systems (operating under independent p-variate normal distributions) on the basis of Mahalanobis distances and provides some tables for determination of the most-economical value of the common sample size n.

When $p = 1$ and the common variance σ^2 of the k univariate normal distributions is unity, the Mahalanobis distances clearly reduce to μ_i^2; the ranking criterion thus is μ_i^2 or equivalently $|\mu_i|$. In this special situation, the solution of the ranking problem with a much larger "preference zone" of the parameter space than that of [1] when $p = 1$ is possible and has been considered by Rizvi [4]. Whereas a more stringent characterization of the preference zone as in [1] is necessary for $p > 1$, the univariate problem is solved with a reasonably general preference zone in [4]. It should be pointed out here that the measurement signal-to-noise ratio $|\mu|/\sigma$, where μ is the mean and σ^2 the variance of a normal random variable, plays a basic role in the evaluation of modern electronic equipment. An electronic device is considered superior if it has a larger signal-to-noise ratio. Thus if we have k electronic devices to compare and they all have a known common variance, we really are interested in ranking k independent normal distributions with unknown means and a common known variance, say unity, according to the unknown ordering of the absolute values of the means. This is the problem treated extensively in [4].

It follows from the general treatment of Hall [2] that the decision rules R_1 and R_2 of this paper are most economical, that is, no other rules can satisfy the basic probability requirement with a smaller fixed sample size.

2. FORMULATION OF THE PROBLEM

Let Π_i denote a p-variate non-singular normal (μ_i, Σ_i) distribution $(i = 1,\ldots,k)$ where μ_i's are unknown. Let the ordered values of $\theta_i = \mu_i' \Sigma_i^{-1} \mu_i$ be denoted by

$$0 \leq \theta_{[1]} \leq \theta_{[2]} \leq \cdots \leq \theta_{[k]} \quad .$$

We are interested in selecting t (< k) "best" distributions in an unordered manner; a "better" distribution is defined to be one with a larger θ-value. The selection of any t largest θ-values is regarded as a correct selection (CS).

Let $\lambda = \left(\theta_{[1]},\ldots,\theta_{[k]}\right)$ denote a point in the parameter space Ω which is partitioned into a "preference zone" Ω^* and its complement, the "indifference zone" $\overline{\Omega}^*$. For specified Ω^* and P^*, $1/\binom{k}{t} < P^* < 1$, we require a decision procedure R for which the probability of a correct selection $P\left\{CS|R\right\}$ satisfies the basic probability requirement

$$\inf_{\Omega^*} P\left\{CS|R\right\} \geq P^* \quad . \tag{1}$$

3. PROPOSED PROCEDURES AND THE PROBABILITY OF A CORRECT SELECTION

First we propose selection procedure R_1 for case 1 where $\Sigma_1,\ldots,\Sigma_k$ are all known.

Procedure R_1 .

Take a random sample of size n $(n > p)$ form each Π_i and compute $U_i = \bar{X}_i' \Sigma_i^{-1} \bar{X}_i$, where $\bar{X}_i$ is the ith sample vector mean $(i = 1,\ldots,k)$. Rank U_i's, breaking ties (if any) with suitable randomization, and select the Π_i's corresponding to t largest U_i's and assert that these are the t best distributions.

Now consider the preference zone Ω^* defined as $\Omega_1 \cap \Omega_2$ where

$$\Omega_1 = \left\{\lambda \epsilon \Omega: \quad \theta_{[k-t+1]} - \theta_{[k-t]} \geq \delta_1 \right\}, \tag{2}$$

$$\Omega_2 = \left\{\lambda \epsilon \Omega: \quad \theta_{[k-t+1]} \geq \delta_2 \theta_{[k-t]} \right\}, \tag{3}$$

and $\delta_1 > 0$ and $\delta_2 > 1$ are specified constants. For $\Omega^* = \Omega_1 \cap \Omega_2$ and R_1, it is shown in [1] that the probability of a correct selection is minimized on Ω^* by the vector λ^* whose components are given by

$$\theta_{[i]} = \begin{cases} \delta_1/(\delta_2-1), & i = 1,\ldots,k-t \\ \delta_1\delta_2/(\delta_2-1), & i = k-t+1,\ldots,k . \end{cases} \tag{4}$$

Moreover, with the distribution function $F_p(x,\theta)$ given by

$$F_p(x,\theta) = e^{-\theta/2} \sum_{r=0}^{\infty} (\theta/2)^r [r!]^{-1} \int_0^x 2^{-(p+2r)/2} [\Gamma((p+2r)/2)]^{-1}$$

$$\times\, e^{-u/2}\, u^{((p+2r)/2)-1}\, du ,$$

for $x > 0$, $\theta \geq 0$ and zero otherwise, the smallest common sample size n required for R_1 to satisfy (1) is obtained as the solution of the integral equation

$$t \int_0^\infty F_p^{k-t}(x, n\delta_1/(\delta_2-1)) \, [1-F_p(x, n\delta_1\delta_2/(\delta_2-1))]^{t-1} dF_p(x, n\delta_1\delta_2/\delta_2-1)) = P^* \tag{5}$$

Note that the left side of equation (5) represents the infimum of the probability of a correct selection over $\Omega^* = \Omega_1 \cap \Omega_2$ for the selection procedure R_1.

Next for case 2 where $\Sigma_1, \ldots, \Sigma_k$ are all unknown, we propose selection procedure R_2.

Procedure R_2.

Take a random sample of size n $(n > p)$ from each Π_i and compute $V_i = (np)^{-1}(n-p)\overline{X}_i' S_i^{-1}\overline{X}_i$, where $\overline{X}_i$ and S_i are respectively the sample vector mean and sample covariance matrix (that is, maximum likelihood estimate of Σ_i) from Π_i, $i = 1, \ldots, k$. Rank V_i's, breaking ties (if any) with suitable randomization, and select the Π_i's corresponding to t largest V_i's and assert that these are the t best distributions.

For $\Omega^* = \Omega_1 \cap \Omega_2$, where Ω_1 is defined by (2) and Ω_2 by (3), and R_2, it is again shown in [1] that the probability of a correct selection is minimized over Ω^* by the vector λ^* whose components are given by (4). Furthermore, with the distribution function $G_{p,n-p}(x,\theta)$ given by

$$G_{p,n-p}(x,\theta) = e^{-\theta/2}\left[\Gamma((n-p)/2)\right]^{-1} \sum_{r=0}^{\infty} (\theta/2)^r \left[r!\right]^{-1} \int_0^x \Gamma((p/2) + ((n-p)/2) + r)$$
$$\times \left[\Gamma((p/2) + r)\right]^{-1} v^{(p/2)+r-1} (1 + v)^{(p/2)+((n-p)/2)+r} \, dv \ ,$$

for $x > 0$, $\theta \geq 0$ and zero otherwise, the smallest common sample size n required for R_2 to satisfy (1) is obtained as the solution of the integral equation

$$t \int_0^\infty G^{k-t}_{n,n-p}(x, n\delta_1/(\delta_2-1)) \, [1-G_{p,n-p}(x, n\delta_1\delta_2/(\delta_2-1))]^{t-1}$$

$$\times \, dG_{p,n-p}(x, n\delta_1\delta_2/(\delta_2-1)) = P^* \quad . \qquad (6)$$

Note that the left side of (6) represents the infimum of the probability of a correct selection over $\Omega^* = \Omega_1 \cap \Omega_2$ for the selection procedure R_2.

4. TABLES AND ILLUSTRATIONS

The left side of (5) and (6) are evaluated by appropriate quadrature and (5) or (6) are then solved for n. This has been done extensively by Milton and Rizvi [3]. Tables I and II are extracted from [3]. Table I gives values of $n\delta_1$ as solution of (5) for $P^* = .95$, $t = 1$, $k = 2(1)5$, $p = 1, 3, 5, 7, 9, 19, 29$ and $\delta_2 = 1.01, 1.05(.05)$ $1.25(.25)2.00(.50)3.00$. Table II gives values of (n, δ_1) as solution of (6) for $P^* = .95$, $t = 1$, $k = 2$, $p = 4, 10$ and $\delta_2 = 1.50, 2.00, 3.00$.

Suppose we wish to select the best of two automation systems that operate under 9-variate normal distributions with known covariance matrices Σ_1 and Σ_2. Moreover, suppose we wish to select $\theta_{[2]}$ (that is the best system) only if $\theta_{[2]} - \theta_{[1]} \geq 5.0$ as well as $\theta_{[2]} \geq 1.5\, \theta_{[1]}$, and require the selection procedure R_1 to have the probability of a correct selection not less than 0.95. Then from Table I we obtain $n\delta_1 = 55.15$ so that we need 12 observations from each of the two 9-variate normal distributions for carrying out procedure R_1.

Next, suppose we are interested in the selection of the best of two automation systems operating under 10-variate normal distributions with unknown covariance matrices Σ_1 and Σ_2. Furthermore, suppose we are interested in this selection only if $\theta_{[2]} - \theta_{[1]} \geq 5.0$ as well as $\theta_{[2]} \geq 1.5\ \theta_{[1]}$, and require the probability of a correct selection using R_2 to be at least 0.95. Then from Table II we obtain $n = 87.292$ so that we need 88 observations from each of the two 10-variate normal distributions for carrying out procedure R_2.

ACKNOWLEDGEMENT

This work was supported in part by National Aeronautics and Space Administration under Grant No. NGR 36-008-040, Supplement 2 at the Ohio State University.

TABLE I

$n\delta_1$ VALUES AS SOLUTION OF (5) WHEN P* = .95 AND t = 1 FOR DETERMINING COMMON SAMPLE SIZE REQUIRED TO SELECT THE BEST SYSTEM IN THE CASE OF ALL KNOWN COVARIANCE MATRICES

k	δ_2	p = 1	p = 3	p = 5	p = 7	p = 9	p = 19	p = 29
2	1.01	2172.00	2172.00	2172.00	2172.00	2172.00	2172.00	2172.00
2	1.05	443.60	443.70	443.70	443.80	443.80	444.00	444.30
2	1.10	227.10	227.20	227.30	227.40	227.50	228.00	228.50
2	1.15	154.90	155.10	155.20	155.30	155.50	156.20	156.90
2	1.20	118.80	119.00	119.20	119.30	119.50	120.40	121.30
2	1.25	97.10	97.32	97.54	97.76	97.98	99.07	100.13
2	1.50	53.56	53.97	54.37	54.76	55.15	57.02	58.77
2	1.75	38.93	39.49	40.03	40.56	41.08	43.49	45.68
2	2.00	31.54	32.23	32.89	33.53	34.15	36.95	39.42
2	2.50	24.03	24.95	25.80	26.60	27.36	30.66	33.45
2	3.00	20.19	21.29	22.28	23.20	24.05	27.65	30.61
3	1.01	2948.00	2948.00	2948.00	2948.00	2948.00	2948.00	2948.00
3	1.05	602.10	602.20	602.20	602.30	602.30	602.60	602.80
3	1.10	308.30	308.40	308.50	308.60	308.70	309.10	309.60
3	1.15	210.30	210.40	210.60	210.70	210.80	211.50	212.20
2	1.20	161.20	161.40	161.60	161.80	162.00	162.90	163.80
3	1.25	131.80	132.00	132.20	132.50	132.70	133.80	134.80
3	1.50	72.70	73.11	73.51	73.90	74.29	76.19	78.00
3	1.75	52.84	53.40	53.94	54.47	55.00	57.48	59.78
3	2.00	42.81	43.50	44.16	44.81	45.44	48.36	51.00

TABLE I - (Continued)

k	δ_2	p = 1	p = 3	p = 5	p = 7	p = 9	p = 19	p = 29
3	2.50	32.62	33.53	34.39	35.21	35.99	39.52	42.56
3	3.00	27.41	28.50	29.50	30.45	31.34	35.24	38.52
4	1.01	3413.00	3413.00	3413.00	3414.00	3414.00	3414.00	3414.00
4	1.05	697.20	697.30	697.30	697.30	697.40	697.60	697.90
4	1.10	357.00	357.10	357.20	357.30	357.40	357.80	358.30
4	1.15	243.50	243.60	243.80	243.90	244.00	244.70	245.40
4	1.20	186.70	186.90	187.10	187.30	187.40	188.30	189.20
4	1.25	152.60	152.80	153.00	153.30	153.50	154.60	155.70
4	1.50	84.18	84.59	84.99	85.38	85.77	87.68	89.51
4	1.75	61.18	61.74	62.28	62.82	63.34	65.86	68.21
4	2.00	49.57	50.25	50.92	51.57	52.20	55.18	57.89
4	2.50	37.77	38.68	39.54	40.37	41.16	44.77	47.93
4	3.00	31.74	32.82	33.83	34.79	35.70	29.72	43.15
5	1.01	3746.00	3746.00	3746.00	3746.00	3746.00	3746.00	3746.00
5	1.05	765.20	765.30	765.30	765.40	765.40	765.70	765.90
5	1.10	391.80	391.90	392.00	392.10	392.20	392.70	393.10
5	1.15	267.20	267.40	267.50	267.70	267.80	268.50	269.20
5	1.20	204.90	205.10	205.30	205.50	205.60	206.50	207.40
5	1.25	167.50	167.70	167.90	168.20	168.40	169.50	170.50
5	1.75	67.15	67.71	68.25	68.79	69.31	71.84	74.22
5	1.50	92.40	92.80	93.20	93.59	93.99	95.90	97.74
5	2.00	54.50	55.09	55.75	56.41	57.04	60.05	62.79
5	2.50	41.46	42.36	43.22	44.05	44.86	48.52	51.75
5	3.00	34.84	35.91	36.93	37.89	38.81	42.91	46.43

TABLE II

(n,δ_1) VALUES AS SOLUTION OF (6) WHEN $P^* = .95$, $k = 2$ AND $t = 1$ FOR DETERMINING COMMON SAMPLE SIZE REQUIRED TO SELECT THE BEST OF TWO SYSTEMS IN THE CASE OF UNKNOWN COVARIANCE MATRICES

δ_2	$n\delta_1$	p = 4		p = 10	
		n	δ_1	n	δ_1
	160.0	0.000	0.000	0.000	0.000
	170.0	102.917	1.652	0.000	0.000
	180.0	100.353	1.794	109.255	1.648
	190.0	98.184	1.935	106.836	1.778
	200.0	96.311	2.077	104.761	1.909
	220.0	93.217	2.360	101.357	2.171
	240.0	90.804	2.643	98.693	2.432
	260.0	88.857	2.926	96.549	2.693
	280.0	87.259	3.209	94.791	2.954
	300.0	85.918	3.492	93.314	3.215
1.50	320.0	84.778	3.775	92.070	3.476
	340.0	83.799	4.057	90.993	3.737
	360.0	82.945	4.340	90.055	3.998
	380.0	82.202	4.623	89.236	4.258
	400.0	81.535	4.906	88.518	4.519
	420.0	80.951	5.188	87.874	4.780
	440.0	80.416	5.472	87.292	5.041
	460.0	79.950	5.754	86.778	5.301
	480.0	79.519	6.036	86.301	5.562
	500.0	79.121	6.319	85.880	5.822

TABLE II - (Continued)

δ_2	$n\delta_1$	p = 4		p = 10	
		n	δ_1	n	δ_1
	40.0	0.000	0.000	0.000	0.000
	50.0	74.997	0.667	99.458	0.503
	60.0	57.674	1.040	73.598	0.815
	70.0	49.593	1.411	62.310	1.123
	80.0	44.928	1.781	55.964	1.430
	90.0	41.890	2.148	51.915	1.734
	100.0	39.747	2.516	49.094	2.037
	110.0	38.162	2.882	47.027	2.339
2.00	120.0	36.934	3.249	45.437	2.641
	130.0	35.965	3.615	44.180	2.943
	140.0	35.171	3.981	43.166	3.243
	150.0	34.514	4.346	42.323	3.544
	160.0	33.961	4.711	41.618	3.845
	170.0	33.485	5.077	41.013	4.145
	180.0	33.077	5.442	40.498	4.445
	190.0	32.720	5.807	40.042	4.745
	200.0	32.407	6.172	39.647	5.044

TABLE II - (Continued)

δ_2	$n\delta_1$	p = 4		p = 10	
		n	δ_1	n	δ_1
	20.0	0.000	0.000	0.000	0.000
	30.0	45.540	0.659	90.854	0.330
	40.0	28.147	1.421	45.171	0.886
	50.0	23.091	2.165	35.256	1.418
	60.0	20.679	2.901	30.919	1.941
	70.0	19.268	3.633	28.491	2.457
	80.0	18.344	4.361	26.939	2.970
3.00	90.0	17.690	5.088	25.860	3.480
	100.0	17.202	5.813	25.067	3.989
	110.0	16.826	6.537	24.460	4.497
	120.0	16.526	7.261	23.979	5.004
	130.0	16.281	7.985	23.590	5.511
	140.0	16.077	8.708	23.269	6.017
	150.0	15.905	9.431	22.999	6.522

REFERENCES

1. K. ALAM AND M. H. RIZVI, "Selection from multivariate normal populations," Annals Inst. Stat. Math. **18**, 307-318 (1966).

2. W. J. HALL, "The most-economical character of some Bechhofer and Sobel decision rules," Annals Math. Statist. **30**, 964-969 (1959).

3. R. C. MILTON AND M. H. RIZVI, "Integrals involving non-central chi-squared and non-central F distributions", to be published (1971).

4. M. H. RIZVI, "Some selection problems involving folded normal distribution," Technometrics **13**, (May 1971).