The Annals of Statistics
1976, Vol. 4, No. 3, 614-620

SELECTION OF LARGEST MULTIPLE CORRELATION COEFFICIENTS: EXACT SAMPLE SIZE CASE[1]

BY KHURSHEED ALAM, M. HASEEB RIZVI AND HERBERT SOLOMON

Clemson University and Stanford University

Consider k ($\geqq 2$) independent p-variate ($p \geqq 2$) normal distributions $N(\mu_i, \Sigma_i)$, $i = 1, 2, \cdots, k$, where the mean vectors μ_i and the covariance matrices Σ_i are all unknown. Let θ_i denote for the ith distribution the squared population multiple correlation coefficient between the first variate and the set of $(p-1)$ variates remaining. A procedure based on the natural ordering of the k sample squared multiple correlation coefficients, each computed from a random sample of size n ($\geqq p+2$), is considered for the problem of selection of the t ($< k$) largest θ_i's. Given $(1-\theta_{[k-t]}) \geqq \delta(1-\theta_{[k-t+1]})$ and $\theta_{[k-t+1]} \geqq \gamma\theta_{[k-t]}$, where $\theta_{[i]}$ denotes the ith smallest θ and $\delta > 1$ and $\gamma > 1$ are preassigned constants, it is shown that the probability of a correct selection is minimized for $\theta_{[i]} = (\delta-1)/(\delta\gamma-1)$, $i = 1, \cdots, k-t$ and $\theta_{[i]} = \gamma(\delta-1)/(\delta\gamma-1)$, $i = k-t+1, \cdots, k$. For a given P^* (< 1), the exact common sample size n is then determined so that the infimum of the probability of a correct selection is not smaller than P^*. For $p = 2$, the problem reduces to selecting t largest correlation coefficients from the k bivariate normal distributions.

1. Introduction and formulation of the problem. In a recent article, Rizvi and Solomon [5] consider the problem of selection of t largest from among k multiple correlation coefficients, each arising from one of k independent p-variate normal distributions with unknown mean vectors and unknown covariance matrices. The problem there is formulated as a ranking problem with a particular choice of an indifference zone in the product parameter space; the main result concerns the minimization of the asymptotic probability of a correct selection for large common sample sizes when the natural selection procedure based on sample multiple correlation coefficients is used for ranking. The present article offers an exact (not asymptotic) solution to the above problem for a slightly different specification of the indifference zone.

Consider k ($\geqq 2$) independent p-variate ($p \geqq 2$) normal distributions $N(\mu_i, \Sigma_i)$, $i = 1, \cdots, k$, where the mean vectors μ_i and the covariance matrices Σ_i are all unknown. For the ith distribution, let θ_i be the squared multiple correlation coefficient between the first variate and the set of $(p-1)$ variates remaining. Let $0 \leqq \theta_{[1]} \leqq \theta_{[2]} \leqq \cdots \leqq \theta_{[k]} < 1$ denote the ordered values of the components of $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_k) \in \Omega$. For $1 \leqq t < k$, the problem of interest is the selection

Received March 1975.

[1] Work supported in part by contract N00014-67-A-0112-0085 (NR-042-267) at Stanford University and by contract N00014-75-C-0451 at Clemson University.

AMS 1970 *subject classifications.* Primary 62F07; Secondary 62H99.

Key words and phrases. Ranking and selection, indifference zone, least favorable configuration, multiple and simple correlation coefficients, exact sample size.

of t distributions with $\theta_i \geqq \theta_{[k-t+1]}$ by some procedure R on the basis of random samples of common size n from each of the k distributions. Denote by CS the correct selection of all distributions with $\theta_{[j]}$, $j = k - t + 1, \cdots, k$ and let $P(\boldsymbol{\theta})$ denote $\Pr\{\text{CS} \mid R\}$. For some preassigned P^*, $1/\binom{k}{t} < P < 1$, it is specifically required to determine the smallest n for which $P(\boldsymbol{\theta})$ is no smaller than P^*, but since such a condition cannot be met for any n without restricting the values of $\boldsymbol{\theta}$, the formal requirement for a selection procedure $R = R(n)$ is that it satisfy the condition

$$\inf_{\Omega^*} P(\boldsymbol{\theta}) \geqq P^* , \tag{1.1}$$

where Ω is partitioned into a preference zone Ω^* and an indifference zone $\bar{\Omega}^*$. Thus the determination of n depends on the particular choice of Ω^* or its complement $\bar{\Omega}^*$. We let $\Omega^* = \Omega_1 \cap \Omega_2$, where

$$\Omega_1 = \{\boldsymbol{\theta} \in \Omega : 1 - \theta_{[k-t]} \geqq \delta(1 - \theta_{[k-t+1]})\} , \tag{1.2}$$

$$\Omega_2 = \{\boldsymbol{\theta} \in \Omega : \theta_{[k-t+1]} \geqq \gamma\theta_{[k-t]}\} , \tag{1.3}$$

and $\delta > 1$ and $\gamma > 1$ are specified constants.

The preference zone specified in [5] is $\Omega_0 = \Omega_1' \cap \Omega_2$, where

$$\Omega_1' = \{\boldsymbol{\theta} \in \Omega : \theta_{[k-t+1]} - \theta_{[k-t]} \geqq \delta_1, 0 < \delta_1 < 1\} . \tag{1.4}$$

A diagram in the Cartesian $(\theta_{[k-t]}, \theta_{[k-t+1]})$-plane readily brings out the difference between Ω^* and Ω. It is observed that for a given value of γ, $\Omega^* \subset \Omega_0$ for $\delta_1 = (\gamma - 1)(\delta - 1)/(\gamma\delta - 1)$ and $\Omega_0 \subset \Omega^*$ for $\delta = 1/(1 - \delta_1)$.

The natural selection procedure R proposed here is the same as in [5]. To be explicit, take a random sample of size n ($n \geqq p + 2$) from each of the k distributions and compute the sample squared multiple correlation coefficient Y_i, $i = 1, \cdots, k$. Rank the Y_i's, breaking ties, if any, with appropriate randomization and select the distributions corresponding to the t largest Y_i's.

Section 2 shows that $P(\boldsymbol{\theta})$ is minimized over Ω^* at a boundary point of Ω^* for all $n \geqq p + 2$. Section 3 contains some remarks about properties of the given procedure and the relationship of this paper with some other work in the literature.

2. Main result on minimization of $P(\boldsymbol{\theta})$. As preliminaries, a few results concerning the distribution of a typical sample squared multiple correlation coefficient Y based on a random sample of size n ($\geqq p + 2$) and having population squared multiple correlation coefficient θ are given. Let

$$F(a, b; c; x) = \sum_{r=0}^{\infty} \frac{(a)_r (b)_r}{(c)_r} \frac{x^r}{x!}$$

denote the hypergeometric function, where $(a)_0 = 1$ and

$$(a)_r = a(a + 1) \cdots (a + r - 1) , \qquad r = 1, 2, \cdots .$$

The probability density function (pdf) of Y is given by

$$g_y(a, c, \theta) = (1 - \theta)^a B_y(c, a - c) F(a, a; c; \theta y) , \qquad 0 < y < 1 , \tag{2.1}$$

where

$$B_y(a, b) = \frac{\Gamma(a + b)}{\Gamma(a)\Gamma(b)} y^{a-1}(1 - y)^{b-1}, \tag{2.2}$$

$$a = (n - 1)/2, \quad c = (p - 1)/2.$$

The formula for the derivative of a hypergeometric function, namely

$$\frac{d^m}{dx^m} F(a, b; c; x) = \frac{(a)_m (b)_m}{(c)_m} F(a + m, b + m; c + m; x) \tag{2.3}$$

yields

$$\frac{d}{d\theta} g_y(a, c, \theta) = \frac{a}{1 - \theta} [g_y(a + 1, c + 1, \theta) - g_y(a, c, \theta)], \tag{2.4}$$

$$\frac{d}{d\theta} G_y(a, c, \theta) = \frac{a}{1 - \theta} [G_y(a + 1, c + 1, \theta) - G_y(a, c, \theta)], \tag{2.5}$$

where $G_y(a, c, \theta)$ denotes the cumulative distribution function (cdf) of Y; (2.4) and (2.5) will be used later for proving the theorem of this section.

A well-known lemma, to be used repeatedly in the sequel, is stated below without proof; two other lemmas are proved.

LEMMA 2.1. *Let* $H(z) = (\sum_{i=0}^{\infty} b_i z^i)/(\sum_{i=0}^{\infty} a_i z^i)$, *where the constants* a_i, b_i *are nonnegative, and* $\sum a_i z^i$ *and* $\sum b_i z^i$ *converge for all* $z > 0$. *If the sequence* $\{b_i/a_i\}$ *is monotone then* $H(z)$ *is a monotone function of* z *in the same direction.*

LEMMA 2.2. *The distribution of* Y *is stochastically increasing in* θ.

PROOF. Let $y' > y > 0$. Then $F(a, a; c; \theta y')/F(a, a; c; \theta y)$ is nonincreasing in θ by Lemma 2.1. The pdf of Y, therefore, has a monotone likelihood ratio in y for θ, from which it follows that the cdf $G_y(a, c, \theta)$ of Y is nonincreasing in θ for each y.

LEMMA 2.3. *Function*

$$u(\theta) = [G_y(a, c, \theta) - G_y(a + 1, c + 1, \theta)]/g_y(a, c, \theta) \tag{2.6}$$

is nonincreasing in θ, *and*

$$v(\theta) = \theta u(\theta)/(1 - \theta) \tag{2.7}$$

is nondecreasing in θ.

PROOF. Write θy as $y - (1 - \theta)y$ in $F(a, a; c; \theta y)$, expand it in a Taylor series and use (2.3) to obtain

$$F(a, a; c; \theta y) = \sum_{r=0}^{\infty} (-1)^r \frac{(1 - \theta)^r y^r}{r!} \frac{(a)_r (a)_r}{(c)_r} F(a + r, a + r; c + r; y). \tag{2.8}$$

From (2.1), (2.8) and the formula (see Erdélyi [2], 2.8(26))

$$cx^{c-1}(1 - x)^{b-c-1} F(a, b; c; x) = \frac{d}{dx} [x^c (1 - x)^{b-c} F(a + 1, b; c + 1; x)], \tag{2.9}$$

one obtains after simplification

$$cG_y(a, c, \theta) = y(1 - y)B_y(c, a - c)(1 - \theta)^a \sum_{r=0}^{\infty} \frac{(-1)^r(1 - \theta)^r y^r}{r!} \frac{(a)_r(a)_r}{(c + 1)_r} \times F(a + r + 1, a + r; c + r + 1; y)$$

and

$$acG_y(a + 1, c + 1, \theta) = y(1 - y)B_y(c, a - c)(1 - \theta)^a \sum_{r=1}^{\infty} \frac{(-1)^{r+1}(1 - \theta)^r y^r}{(r - 1)!} \times \frac{(a)_r(a)_r}{(c + 1)_r} F(a + r + 1, a + r; c + r + 1; y) .$$

Hence,

$$\begin{aligned} c(G_y(a, c, \theta) - G_y(a + 1, c + 1, \theta)) &= y(1 - y)B_y(c, a - c)(1 - \theta)^a \sum_{r=0}^{\infty} \frac{(-1)^r(1 - \theta)^r y^r}{r!} \\ &\quad \times \frac{(a)_r(a + 1)_r}{(c + 1)_r} F(a + r + 1, a + r; c + r + 1; y) \\ &= y(1 - y)B_y(c, a - c)(1 - \theta)^a F(a + 1, a; c + 1; \theta y) . \end{aligned} \tag{2.10}$$

In view of (2.1) and (2.10), $u(\theta)$ given by (2.6) becomes

$$u(\theta) = \frac{y(1 - y)}{c} \cdot \frac{F(a + 1, a; c + 1; \theta y)}{F(a, a; c; \theta y)} . \tag{2.11}$$

Apply Lemma 2.1 to the right side of (2.11) to conclude that $u(\theta)$ is nonincreasing in θ.

Also,

$$\begin{aligned} \theta y F(a + 1, a; c + 1; \theta y) &= \sum_{r=1}^{\infty} \frac{(a + 1)_{r-1}(a)_{r-1}}{(c + 1)_{r-1}} \cdot \frac{(\theta y)^r}{(r - 1)!} \\ &= \frac{c}{a} \sum_{r=1}^{\infty} \frac{(a)_r(a)_{r-1}}{(c)_r} \frac{(\theta y)^r}{(r - 1)!} , \end{aligned} \tag{2.12}$$

and

$$\begin{aligned} (1 - \theta y)F(a, a; c; \theta y) &= 1 + \sum_{r=1}^{\infty} \left\{ \frac{(a)_r(a)_r}{r(c)_r} - \frac{(a)_{r-1}(a)_{r-1}}{(c)_{r-1}} \right\} \frac{(\theta y)^r}{(r - 1)!} \\ &= 1 + \sum_{r=1}^{\infty} \frac{(a)_r(a)_{r-1}(\theta y)^r}{(c)_r(r - 1)!} \cdot \left(\frac{a - 1}{r} + \frac{a - c}{a + r - 1} \right) . \end{aligned} \tag{2.13}$$

Comparing the coefficients of θ^r in the series on the right sides of (2.12) and (2.13), and applying Lemma 2.1, one finds that $v(\theta)$ given by (2.7) is nondecreasing in θ.

THEOREM. *The probability $P(\boldsymbol{\theta})$ of a correct selection for the selection procedure R is minimized over Ω^* for any $\boldsymbol{\theta}$ whose components satisfy*

$$\begin{aligned} \theta_{[i]} &= (\delta - 1)/(\delta\gamma - 1) , \qquad i = 1, \cdots, k - t \\ &= \gamma(\delta - 1)/(\delta\gamma - 1) , \qquad i = k - t + 1, \cdots, k , \end{aligned} \tag{2.14}$$

and

$$\begin{aligned}\inf_{\Omega^*} P(\boldsymbol{\theta}) = t \int_0^1 G_y^{k-1}(a, c, (\delta - 1)/(\delta\gamma - 1)) \\ \times [1 - G_y(a, c, \gamma(\delta - 1)/(\delta\gamma - 1))]^{t-1} \\ \times g_y(a, c, \gamma(\delta - 1)/(\delta\gamma - 1))\, dy\,,\end{aligned} \tag{2.15}$$

where $a = (n - 1)/2$ *and* $c = (p - 1)/2$.

PROOF. The Pr $\{CS \mid R\}$ is given by

$$\begin{aligned}P(\boldsymbol{\theta}) = \sum_{i=k-t+1}^{k} \int_0^1 \prod_{\alpha=1}^{k-t} G_y(a, c, \theta_{[\alpha]}) \prod_{\beta=k-t+1, \beta\neq i}^{k} \{1 - G_y(a, c, \theta_{[\beta]})\} \\ \times g_y(a, c, \theta_{[i]})\, dy\,.\end{aligned} \tag{2.16}$$

Given $\theta_{[k-t]} = \theta$ (say), Lemma 2.2 and Lemma 2.1 of Alam and Rizvi [1] imply that $P(\boldsymbol{\theta})$ is minimized over Ω_1 for

$$\begin{aligned}\theta_{[i]} &= \theta && \text{for} \quad i = 1, \cdots, k - t \\ &= (\theta + \delta - 1)/\delta && \text{for} \quad i = k - t + 1, \cdots, k\,.\end{aligned} \tag{2.17}$$

Therefore,

$$\begin{aligned}P_1 &= \inf_{\theta \in \Omega_1} P(\boldsymbol{\theta}) \\ &= \inf_{0\leq\theta\leq 1} t \int_0^1 G_y^{k-t}(a, c, \theta)[1 - G_y(a, c, \lambda(\theta))]^{t-1} g_y(a, c, \lambda(\theta))\, dy\,,\end{aligned} \tag{2.18}$$

where $\lambda(\theta) = (\theta + \delta - 1)/\delta$. Similarly, $P(\boldsymbol{\theta})$ is minimized over Ω_2, given $\theta_{[k-t]} = \theta$, for

$$\begin{aligned}\theta_{[i]} &= \theta && \text{for} \quad i = 1, \cdots, k - t \\ &= \gamma\theta && \text{for} \quad i = k - t + 1, \cdots, k\,.\end{aligned} \tag{2.19}$$

Therefore,

$$\begin{aligned}P_2 &= \inf_{\theta \in \Omega_2} P(\boldsymbol{\theta}) \\ &= \inf_{0\leq\theta\leq 1/\gamma} t \int_0^1 G_y^{k-t}(a, c, \theta)[1 - G_y(a, c, \gamma\theta)]^{t-1} g_y(a, c, \gamma\theta)\, dy\,.\end{aligned} \tag{2.20}$$

Let $A(\theta)$ and $B(\theta)$ denote respectively the integrals on the right side of (2.18) and (2.20). Differentiating $A(\theta)$ with respect to θ and using (2.4) and (2.5) yields

$$\begin{aligned}\frac{d}{d\theta} A(\theta) = \frac{a(k - t)}{1 - \theta} \int_0^1 G_y^{k-t-1}(a, c, \theta)[1 - G_y(a, c, \lambda(\theta))]^{t-1} \\ \times [G_y(a + 1, c + 1, \theta) - G_y(a, c, \theta)] g_y(a, c, \lambda(\theta))\, dy \\ - \frac{a(t - 1)}{\delta(1 - \lambda(\theta))} \int_0^1 G_y^{k-t}(a, c, \theta)[1 - G_y(a, c, \lambda(\theta))]^{t-2} \\ \times [G_y(a + 1, c + 1, \lambda(\theta)) - G_y(a, c, \lambda(\theta))] g_y(a, c, \lambda(\theta))\, dy \\ + \frac{a}{\delta(1 - \lambda(\theta))} \int_0^1 G_y^{k-t}(a, c, \theta)[1 - G_y(a, c, \lambda(\theta))]^{t-1} \\ \times [g_y(a + 1, c + 1, \lambda(\theta)) - g_y(a, c, \lambda(\theta))]\, dy\,.\end{aligned} \tag{2.21}$$

Integrate by parts the third integral on the right side of (2.21) to obtain after

simplification

$$\frac{d}{d\theta}A(\theta) = a(k-t)\int_0^1 G_y^{k-t-1}(a, c, \theta)[1 - G_y(a, c, \lambda(\theta))]^{t-1}$$

$$\times \left\{\frac{[G_y(a+1, c+1, \theta) - G_y(a, c, \theta)]}{(1-\theta)g_y(a, c, \theta)}\right.$$

$$\left. - \frac{G_y[a+1, c+1, \lambda(\theta)] - G_y[a, c, \lambda(\theta)]}{\delta(1-\lambda(\theta))g_y(a, c, \lambda(\theta))}\right\} \tag{2.22}$$

$$\times g_y(a, c, \lambda(\theta))g_y(a, c, \theta)\, dy$$

$$= \frac{a(k-t)}{1-\theta}\int_0^1 G_y^{k-t-1}(a, c, \theta)[1 - G_y(a, c, \lambda(\theta))]^{t-1}[u(\lambda(\theta)) - u(\theta)]$$

$$\times g_y(a, c, \lambda(\theta))g_y(a, c, \theta)\, dy\,.$$

Since $u(\theta) \geqq u(\lambda(\theta))$ by Lemma 2.3, $A(\theta)$ is nonincreasing in θ. Similarly, using Lemma 2.3 again, $B(\theta)$ is nondecreasing in θ.

From the monotonic natures of $A(\theta)$ and $B(\theta)$ as demonstrated above, one concludes that $P(\boldsymbol{\theta})$ is minimized over $\Omega^* = \Omega_1 \cap \Omega_2$ by any value of $\boldsymbol{\theta}$ for which (2.14) holds. The infimum of $P(\boldsymbol{\theta})$ over Ω^* is thus given by (2.15).

3. Concluding remarks. The preference zone Ω^* covers a portion of the upper left corner of the unit square in the $(\theta_{[k-t]}, \theta_{[k-t+1]})$ plane, and is designed to exclude the points (0, 0) and (1, 1) in that plane. If the values of γ and δ are close to 1 then Ω^* includes most of the region $\theta_{[k-t+1]} \geqq \theta_{[k-t]}$. On the other hand, if the values of γ and δ are large, then Ω^* includes only a small neighborhood of the point (0, 1). Therefore, for large P^* and small n the preference zone might be quite restrictive.

The parameter space specified by (2.14) is called the least favorable configuration (LFC). The comparable LFC in [5] is given by

$$\theta_{[i]} = \delta_1/(\gamma - 1)\,, \qquad i = 1, \cdots, k-t$$
$$= \delta_1\gamma/(\gamma - 1)\,, \qquad i = k-t+1, \cdots, k\,.$$

The requirement in [5] for δ_1 to be $0(1/n)$ somewhat limits the scope of applicability of the results in [5]. The present paper, in a way, complements [5] and enlarges the scope for use of the procedure R. For motivation, applications and reference to some formulations as alternatives to the present treatment, the reader is referred to [5]. Important problems arising in applications where the k multiple correlation coefficients are not independent do not receive consideration here. Frischtak [3] and Ramberg [4] consider some aspects of this latter problem within the framework of an indifference zone selection formulation like the one used here. Frischtak [3] also considers the problem of selecting the smallest vector coefficient of alienation (a generalization of multiple correlation coefficient) between two sets of components among k independent multivariate normal distributions. However, he employs preference zones different from those of [5] and this paper. Also, he only uses the asymptotic theory. The

specific problem treated in the present paper thus has no direct bearing to his work and any meaningful comparisons are precluded.

REFERENCES

[1] ALAM, K. and RIZVI, H. H. (1966). Selection from multivariate normal populations. *Ann. Inst. Statist. Math.* **18** 307–318.

[2] ERDÉLYI, A. (1953). *Higher Transcendental Functions.* Bateman Manuscript Project **1**. McGraw-Hill, New York.

[3] FRISCHTAK, R. M. (1973). Statistical multiple decision procedures for some multivariate selection problems. Technical report no. 187, Department of Operations Research, Cornell Univ.

[4] RAMBERG, J. S. (1969). A multiple decision approach to the selection on the best set of predictor variates. Technical report no. 79, Department of Operations Research, Cornell Univ.

[5] RIZVI, M. H. and SOLOMON, H. (1973). Selection of largest multiple correlation coefficients: asymptotic case. *J. Amer. Statist. Assoc.* **68** 184–188 (Corrigendum, 1974, *J. Amer. Statist. Assoc.* **69** 288).

KHURSHEED ALAM
MATHEMATICS DEPARTMENT
CLEMSON UNIVERSITY
CLEMSON, SOUTH CAROLINA 29631

HERBERT SOLOMON
DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305

M. HASEEB RIZVI
SYSOREX, INC.
1801 PAGE MILL ROAD
PALO ALTO, CALIFORNIA 94304