Statistical Selection Procedures in Multivariate Models*

by

Shanti S. Gupta
Purdue University
and
S. Panchapakesan
Southern Illinois University
Technical Report#86-52

Department of Statistics
Purdue University

November 1986

* Research supported by the Office of Naval Research Contract N00014-84-C-0167 at Purdue University. Reproduction in whole or part is permitted for any purpose of the United States Government.

Statistical Selection Procedures in Multivariate Models*

by

Shanti S. Gupta
Purdue University
and
S. Panchapakesan
Southern Illinois University

Abstract

Selection and ranking problems have been studied over the last thirty years, generally under one of two formulations: Bechhofer's indifference -zone approach and Gupta's subset selection approach. This paper deals with subset selection. Subset selection procedures in multivariate models are briefly reviewed. These include: (1) Procedures for selecting the best component in a multivariate normal population in terms of the component means as well as the component variances (Section 2), (2) Procedures for selecting the best from several multivariate normal populations in terms of (i) the Mahalanobis distance, (ii) the generalized variance, and (iii) the multiple correlation coefficient (Section 3), (3) Procedures (fixed sample size as well as inverse sampling) for selecting the most (least) probable cell in a multinomial distribution (Section 4), (4) Procedures for selecting the best from several multinomial populations in terms of the Shannon entropy function (Section 5), and (5) Procedures for choosing the best subset of the predictor variables in a linear regression model (Section 6).

Key Words and Phrases: subset selection, multivariate normal populations, Mahalanobis distance, generalized variance, multiple correlation coefficient, multinomial, Shannon entropy, linear regression, important predictor variables.

* Research supported by the Office of Naval Research Contract N00014-84-C-0167 at Purdue University. Reproduction in whole or part is permitted for any purpose of the United States Government.

Shanti S. Gupta and S. Panchapakesan

STATISTICAL SELECTION PROCEDURES IN MULTIVARIATE MODELS

1. INTRODUCTION

Since statistical inference problems were first posed in the now-familiar "selection and ranking" framework over three decades ago, these problems have been studied from several points of view using various goals and formulations. However, selection from multivariate populations is an important topic that has not been adequately studied in the literature. Our interest here is to briefly review developments pertaining to selection from multivariate models. In doing so, we consider: (1) selection from a single multivariate normal population, (2) selection from several multivariate normal populations, (3) selection from a multinomial population, (4) selection from several multinomial populations, and (5) selection from a set of predictor variables in a regression model.

For ranking multivariate populations, usually a scalar function of the unknown parameters has been chosen in all the investigations. This permits a complete order of the populations. The choice of the ranking measure depends, of course, on the specific situations. The selection procedure in these cases depends on a suitably chosen statistic which has a univariate distribution.

Let us consider k independent populations $\pi_1, \ldots, \pi_k$, where π_i has the underlying distribution function F_{θ_i}, $i = 1, \ldots, k$. The θ_i are unknown real-valued parameters; these represent the values of a certain quality characteristic θ for the k populations. The populations are ranked according to their θ-values. To be specific, π_i is defined to be *better than* π_j if $\theta_i \geq \theta_j$. The ordered θ_i are denoted by $\theta_{[1]} \leq \cdots \leq \theta_{[k]}$. It is assumed that there is no prior knowledge regarding the correct pairing of the ordered and the unordered θ_i. Selection problems have been generally studied under one of two formulations, namely, (1) the *indifference-zone* and (2) the *subset selection* formulations.

Considering the basic problem of *selecting the best population* (i.e. the population associated with $\theta_{[k]}$), the indifference-zone formulation of Bechhofer (1954) requires that one of the k populations be

chosen as the best. A *correct selection* (CS) is said to occur when any population associated with $\theta_{[k]}$ is selected. Any *valid procedure* R must guarantee a specified minimum probability of a correct selection (PCS) whenever the best and the next best populations are sufficiently (to be specified) apart. Let $\delta(\theta_{[k]}, \theta_{[k-1]})$ denote an appropriately chosen measure of the separation between the best and the next best populations, and $P(CS|R)$ denote the PCS using the rule R. Further, let

$$\Omega_{\delta^*} = \{\underset{\sim}{\theta} | \underset{\sim}{\theta} = (\theta_1, \ldots, \theta_k), \delta(\theta_{[k]}, \theta_{[k-1]}) \geq \delta^* > 0\}. \tag{1}$$

Any valid rule R should satisfy

$$P(CS|R) \geq P^* \text{ whenever } \underset{\sim}{\theta} \in \Omega_{\delta^*}. \tag{2}$$

Both δ^* and $P^* \epsilon (1/k, 1)$ are specified by the experimenter in advance. Suppose R is based on samples of size n from each population. Then the problem is to determine the smallest n for which the requirement (2) is satisfied. It should be noted that there is no guarantee to be met when $\underset{\sim}{\theta}$ belongs to $\Omega^c_{\delta^*}$, the complement of Ω_{δ^*}. The region $\Omega^c_{\delta^*}$ is the "indifference-zone" lending its name to the formulation.

In the *subset selection* formulation studied extensively beginning with the pioneering work of Gupta (1956, 1965), the basic problem is to select a nonempty subset of the k populations so that the best population is included in the selected subset with a specified minimum PCS. The size of S, the selected subset, is not determined in advance but by data themselves. Selection of any subset that includes the best population results in a correct selection. Letting Ω denote the entire parameter space, any valid rule R should satisfy

$$P(CS|R) \geq P^* \quad \text{for all } \underset{\sim}{\theta} \in \Omega. \tag{3}$$

This requirement (3) is called the *basic probability requirement*, or the *P^*-condition.* Any configuration $\underset{\sim}{\theta}$ which yields the infimum of PCS over Ω is called a *least favorable configuration* (LFC).

The expected value of $|S|$, the size of S, is a reasonable measure of the performance of a valid rule and has been generally used. Some other possible measures (considered by a few authors) are $E(|S|)/P(CS|R)$ and $E(|S|) - P(CS|R)$, the latter being the expected number of non-best populations included in S.

There are many variations and generalizations of the basic formulation using either of the two approaches described above. There are also related problems such as selecting populations that are better than a standard or a control. A comprehensive survey of the develop-

ments encompassing all these aspects with an extensive bibliography is given by Gupta and Panchapakesan (1979). Recently, Gupta and Panchapakesan (1985) have provided a critical review of developments in the subset selection theory with historical perspectives. For a categorized bibliography, see Dudewicz and Koo (1982).

In the present paper, we are concerned with subset selection procedures for multivariate populations. In Section 2, we discuss selection of the best component in a multivariate normal population in terms of the means as well as the variances. Selection from several multivariate normal populations is discussed in Section 3 using different criteria such as the Mahalanobis distance, the generalized variance, and the multiple correlation coefficient. Section 4 deals with selecting the most probable and the least probable cells in a multinomial distribution. Selection from several multinomial populations is discussed in Section 5 using the Shannon entropy function for comparison of the populations. Finally, Section 6 describes subset selection procedures for choosing a best set of predictor variables in a linear regression model.

2. SELECTION FROM A SINGLE MULTIVARIATE NORMAL POPULATION

Consider a p-variate normal population $N_p(\underset{\sim}{\mu}, \Sigma)$ with mean vector $\underset{\sim}{\mu}' = (\mu_1, \ldots, \mu_p)$ and covariance matrix $\Sigma = (\sigma_{ij})$, which is assumed to be positive definite. In this section, we consider ranking the p components according to their means μ_i, and according to their variances σ_{ii}.

2.1. Selection in Terms of the Means

Let $\underset{\sim}{X}' = (X_1, \ldots, X_p)$ be the sample mean based on n independent (vector) observations from the population. We first consider the case of known Σ and assume, without loss of generality, that $\sigma_{ii} = 1$ for $i = 1, \ldots, p$. For selecting the component associated with $\mu_{[p]}$, the largest μ_i, Gnanadesikan (1966) considered the procedure

$$R_1: \text{ Select the } i\text{th component if and only if } X_i \geq X_{[p]} - \frac{d_1}{\sqrt{n}} \quad (4)$$

where $X_{[1]} \leq \ldots \leq X_{[p]}$ denote the ordered X_i, and $d_1 = d_1(n, p, \Sigma) > 0$ is the smallest number such that the P^*-condition is

satisfied. It is easily shown that

$$\inf_{\Omega} P(CS|R_1) = \Pr\{Y_p \geq Y_j - d_1, \ j = 1, \ldots, p-1\}, \quad (5)$$

where $Y_i = \sqrt{n}(X_{(i)} - \mu_{[i]})$, $X_{(i)}$ is the component sample mean associated with $\mu_{[i]}$, and $\Omega = \{\underset{\sim}{\mu} : \ -\infty < \mu_i < \infty, \ i = 1, \ldots, p\}$. For evaluating d_1 for which the right-hand side of (5) equals P^*, we need to know $A = (a_{ij})$, the covariance matrix of $\underset{\sim}{Y}' = (Y_1, \ldots, Y_p)$. Even though Σ is known, we do not know the correspondence between the σ_{ij} and the a_{ij} except when $p = 2$. For $p = 2$, the right-hand side of (5) equals $\Phi[d_1/\sqrt{2(1-\sigma_{12})}]$, where $\Phi(\cdot)$ is the cdf of a standard normal random variable; this gives

$$d_1 = d_1(n, 2, \Sigma) = \sqrt{2(1-\sigma_{12})}\Phi^{-1}(P^*). \quad (6)$$

For $p > 2$, Gnanadesikan (1966) obtain two different lower bounds for the infimum of PCS. Letting $d_{01} = \min\{d_1/\sqrt{2(1-a_{pj})}, \ j = 1, \ldots, p-1\}$, one gets

$$\inf_{\Omega} P(CS|R_1) \geq \Pr\{Z_j \leq d_{01}, \ j = 1, \ldots, p-1\} \quad (7)$$

where $\underset{\sim}{Z}' = (Z_1, \ldots, Z_{p-1})$ has $N_{p-1}(\underset{\sim}{0}, B)$ distribution and B has a known structure with elements being 0, or $[2(1-a_{jp})]^{-\frac{1}{2}}$, or $-[2(1-a_{jp})]^{-\frac{1}{2}}$, $j = 1, \ldots, p-1$. One lower bound for the right-hand side of (7) obtained by Gnanadesikan (1966) is $\Phi^{p-1}(d_{01})$ based on an inequality due to Slepian (1962). The other lower bound is $(2-p) + (p-1)\Phi(d_{01})$ obtained by using a Bonferroni inequality. For $p = 2$, the two bounds coincide. While d_{01}, using either lower bound, is a conservative value for d_1, the computations of Gnanadesikan (1966) show that d_{01} in the former case (Slepian inequality) is closer to the exact value. However, the difference between the two approximate values decreases as P^* increases and is very small for $P^* \geq .90$.

The determination of the constant d becomes easier when $\sigma_{ij} = \rho > 0$, $i \neq j$. In this case, we get

$$\inf_{\Omega} P(CS|R_1) = \int_{-\infty}^{\infty} \Phi^{p-1}(x + \frac{d}{\sqrt{1-\rho}}) d\Phi(x) \quad (8)$$

and $H = d/\sqrt{2(1-\rho)}$ are tabulated by Gupta (1963a) and by Gupta, Nagel and Panchapakesan (1973) who have also considered the selec-

tion problem in this special case.

When the covariance matrix Σ is unknown, let us assume that $\sigma_{ii} = \sigma^2$ for $i = 1, \ldots, p$, and let s_ν^2 denote an estimator of σ^2 on ν degrees of freedom, statistically independent of the X_i. In this case, Gnanadesikan (1966) proposed the procedure

$$R_2: \quad \text{Select the } i\text{th component if and only if } X_i \geq X_{[p]} - \frac{d_2 s_\nu}{\sqrt{n}} \tag{9}$$

where $d_2 = d_2(\nu, p, P^*) > 0$ is the smallest number for which the P^*-condition is satisfied. For this procedure,

$$\begin{aligned} \inf_{\Omega} P(CS|R_2) &\geq \Pr\{t_i \leq d_{01},\ i = 1, \ldots, p-1\} \\ &\geq 1 - \sum_{i=1}^{p-1} \Pr\{t_i \geq d_{01}\} \end{aligned} \tag{10}$$

where $t_i = Z_i/s_\nu$, $\underset{\sim}{Z}' = (Z_1, \ldots, Z_{p-1})$ has the same distribution as in the known Σ case, $\nu s_\nu^2/\sigma^2$ has a chi-square distribution with ν degrees of freedom, d_{01} is defined as before, and $\Omega = \{(\underset{\sim}{\mu}, \Sigma)\}$. Equating the last member of the inequalities in (10) to P^*, an approximate value of d_{01} is given by

$$(2-p) + (p-1)G_\nu(d_0) = P^* \tag{11}$$

where $G_\nu(\cdot)$ is the cdf of a Student's t variable with ν degrees of freedom. In the special case of $\sigma_{ij} = \rho\sigma^2$, $\rho > 0$, d_{01} can be evaluated as an equicoordinate percentage point of a multivariate t distribution. The d_{01} values are tabulated by Gupta and Sobel (1957), Krishnaiah and Armitage (1966), and Gupta, Panchapakesan and Sohn (1985).

2.2. Selection in Terms of the Variances

We now define the best component as the one associated with the smallest σ_{ii}. A natural procedure is analogous to that of Gupta and Sobel (1962a) in the uncorrelated case. This procedure is

$$R_3: \quad \text{Select the } i\text{th component if } s_{ii} \leq \frac{1}{c} \min_{1 \leq j \leq p} s_{jj} \tag{12}$$

where $c = c(p, n, P^*) \in (0, 1)$ is the largest number for which the P^*-condition is satisfied, and $S = (s_{ij})$ is the sample covariance matrix

based on n independent (vector) observations from the population. This procedure has been considered by Frischtak (1973), who has shown that, for $p = 2$, the infimum of PCS is attained when $\sigma_{11} = \sigma_{22}$ and $\sigma_{12} = 0$. Thus c can be obtained from the tables of Gupta and Sobel (1962b).

For $p \geq 3$, Frischtak (1973) obtained only an asymptotic ($n \to \infty$) solution, using the asymptotic normality of $\log(s^2_{(1)}/s^2_{(j)})$, $j = 2, \ldots, p$, after suitable normalization; here $s^2_{(i)}$ is the s_{ii} associated with the ith smallest σ_{ii}. The asymptotic solution c is given by

$$\Pr\{Y_j \leq \sqrt{\frac{n-1}{2}} \log c,\ j = 2, \ldots, p\} = P^* \tag{13}$$

where the Y_j are standard normal random variables with equal correlation 0.5, and can be obtained from the tables of Gupta (1963a) and Gupta, Nagel and Panchapakesan (1973).

3. SELECTION FROM SEVERAL MULTIVARIATE NORMAL POPULATIONS

Let $\pi_1, \ldots, \pi_k$ be k p-variate normal populations, $N_p(\underset{\sim}{\mu}_i, \Sigma_i)$, $i = 1, \ldots, k$, where the $\underset{\sim}{\mu}_i$ are the mean vectors and the Σ_i are positive definite covariance matrices. For defining the best population, several measures have been used such as the generalized variance, Mahalanobis distance, and the multiple correlation coefficient. Also, comparison with a control has been studied using as criteria linear combinations of the elements of the mean vector and those of the covariance matrix. We now discuss these briefly.

3.1. Selection in Terms of Mahalanobis Distance

Let $\lambda_i = \underset{\sim}{\mu}_i' \Sigma_i^{-1} \underset{\sim}{\mu}_i$, the Mahalanobis distance of π_i from the origin. We first assume that the Σ_i are *known*. Let $\underset{\sim}{X}_{ij}$, $j = 1, \ldots, n$, denote n (vector) observations from π_i, $i = 1, \ldots, k$. Define $Y_{ij} = \underset{\sim}{X}_{ij}' \Sigma_i^{-1} \underset{\sim}{X}_{ij}$ and $Y_i = \sum_{j=1}^{n} Y_{ij}$. For selecting a subset containing the population associated with $\lambda_{[k]}$, Gupta (1966) proposed the procedure

$$R_4: \quad \text{Select } \pi_i \text{ if and only if } Y_i \geq c_4 Y_{[k]} \tag{14}$$

where $0 < c_4 = c_4(k, p, n, P^*) < 1$ is to be chosen suitably to meet

the P^*-condition. It has been shown [Gupta (1966) and Gupta and Studden (1970)] that the infimum of PCS occurs when $\lambda_1 = \cdots = \lambda_k = 0$. Thus the constant c_4 is given by

$$\int_0^\infty G_\nu^{k-1}(\frac{x}{c_4})dG_\nu(x) = P^* \tag{15}$$

where $G_\nu(x)$ is the cdf of a standardized (i.e unit scale parameter) gamma variable with $\nu = np/2$ degrees of freedom. The values of c are tabulated by Gupta (1963b) and Armitage and Krishnaiah (1964).

An analogous procedure can be defined for selecting the population with the smallest λ_i. In this case, the appropriate constant can be obtained from the tables of Gupta and Sobel (1962b) and Krishnaiah and Armitage (1964).

It should be noted that the procedure R_4 is based on the statistics $Y_i = \sum_{j=1}^{n} \underset{\sim}{X}_{ij}' \Sigma_i^{-1} \underset{\sim}{X}_{ij}$ rather than $Z_i = \overline{\underset{\sim}{X}}_i' \Sigma_i^{-1} \overline{\underset{\sim}{X}}_i$, where $\overline{\underset{\sim}{X}}_i$ denote the sample mean vector from π_i. If we use Z_i instead of Y_i in R_4, the infimum of PCS and hence the constant c_4 do not depend on n. This makes the procedure unsatisfactory. One can, of course, use a different type of procedure. For example, we can define R' : Select π_i if and only if $Z_i \geq Z_{[k]} - d,\ d > 0$. Such a procedure has not been investigated.

When the Σ_i are *unknown and not necessarily equal*, Gupta and Studden (1970) proposed and studied the rule

$$R_5 : \quad \text{Select } \pi_i \text{ if and only if } T_i \geq c_5 T_{[k]} \tag{16}$$

where $T_i = \overline{\underset{\sim}{X}}_i' S_i^{-1} \overline{\underset{\sim}{X}}_i$, S_i is the usual sample covariance matrix with $(n-1)$ as the divisor, and $0 < c_5 = c_5(k, n, p, P^*) < 1$ is chosen suitably to satisfy the P^*-condition. It has been shown by Gupta and Studden (1970) that

$$\inf_{\Omega} P(CS|R_5) = \int_0^\infty F_{p,n-p}^{k-1}(\frac{x}{c_5})dF_{p,n-p}(x) \tag{17}$$

where $F_{p,n-p}(x)$ is the cdf of a central F-variable with p and $n-p$ degrees of freedom. The values of c_5 for which the right-hand side of (17) equals P^* have been tabulated by Gupta and Panchapakesan (1969) for various values of k, P^*, p, and n.

Gupta and Studden (1970) also studied the problem of selecting

the population associated with the smallest λ_i. Their rule is

$$R_5': \quad \text{Select } \pi_i \text{ if and only if } T_i \leq \frac{1}{c_5'} T_{[1]} \tag{18}$$

where $0 < c_5' = c_5'(k, n, p, P^*) < 1$ is to be chosen suitably. In this case,

$$\inf_{\Omega} P(CS|R_5') = \int_0^\infty [1 - F_{p,n-p}(c_5' x)]^{k-1} dF_{p,n-p}(x). \tag{19}$$

The constant c_5' for which the right-hand side of (19) equals P^* has been tabulated by Gupta and Panchapakesan (1969) for several combinations of k, P^*, p, and n.

When $\Sigma_1 = \cdots = \Sigma_k = \Sigma$ and Σ is *unknown*, one would define a procedure with $T_i = \overline{\underset{\sim}{X}}_i' S^{-1} \overline{\underset{\sim}{X}}_i$ in R_5, where S is the usual pooled estimator of Σ. This procedure was proposed by Gupta and Studden (1970) and studied later by Chattopadhyay (1981). He has discussed evaluation of the constant in an approximate sense, i.e. the infimum of PCS is approximately P^* but can be on either side of it.

3.2. Selection in Terms of the Generalized Variance

It is meaningful to rank multivariate normal populations according to the amounts of dispersion in them. A frequently used measure of dispersion is the generalized variance which is the determinant of the covariance matrix. Let $\theta_i = |\Sigma_i|$, $i = 1, \ldots, k$. We define the best population as the one associated with the smallest θ_i. Let S_i be the sample covariance matrix based on a sample of size n from π_i, $i = 1, \ldots, k$. Gnanadesikan and Gupta (1970) proposed the rule

$$R_6: \quad \text{Select } \pi_i \text{ if and only if } W_i \leq \frac{1}{c_6} W_{[1]} \tag{20}$$

where $W_i = |S_i|$, and $0 < c_6 = c_6(k, n, p, P^*) < 1$ is to be chosen suitably to satisfy the P^*-condition. It has been shown that

$$\inf_{\Omega} P(CS|R_6) = \Pr\{Y_1 \leq \frac{1}{c_6} Y_j, \ j = 2, \ldots, k\} \tag{21}$$

where $Y_1, \ldots, Y_k$ are independent and identically distributed, each being the product of p independent factors, the rth factor having a chi-square distribution with $(n - r)$ degrees of freedom. An exact

solution for c_6 is obtained in the case of $p = 2$, using the fact that $2(n-1)^{p/2}(W_i/\theta_i)^{\frac{1}{2}}$ is then distributed as a chi-square variable with $2(n-2)$ degrees of freedom. The constant c_6 in this case can be obtained from the tables of Gupta and Sobel (1962b) and Krishnaiah and Armitage (1964).

When $p > 2$, one can use Hoel's approximation of the distribution of $Y_i^{1/p}$ by a gamma distribution with scale parameter θ^{-1} and shape parameter m, where $2m = p(n-p)$ and $2\theta = p[1-(2n)^{-1}(p-1)(p-2)]^{1/p}$. Another approximation is that of $p^{-1}\log Y_i$ using the normal approximation of $\log\chi^2$. Gnanadesikan and Gupta (1970) have studied these approximations.

Some alternative procedures have been proposed by Regier (1976). These procedures are R_6' : Select π_i if and only if $W_i \leq a(\prod_{j=1}^{k} W_j)^{1/k}$ and R_6'' : Select π_i if and only if $W_i \leq b\sum_{j=1}^{k} W_j/k$. Again, the evaluation of the constants a and b are based on normal approximation to $\log\chi^2$ and the asymptotic distribution of the sample variance, respectively. Regier (1976) has given some numerical comparisons of the three procedures.

3.3. Selection in Terms of Multiple Correlation Coefficient

We now assume that the $\underset{\sim}{\mu}_i$ and Σ_i are unknown. Let ρ_i denote the multiple correlation coefficient between the first variable and the rest in π_i. It is a measure of dependence between the two partitioned sets. Gupta and Panchapakesan (1969) investigated the problem of selecting a subset containing the population associated with $\rho_{[k]}(\rho_{[1]})$. Let R_i denote the multiple correlation coefficient between the first variable and the rest from the sample $\underset{\sim}{X}_{ij}$, $j = 1,\ldots,n$. Two cases arise: (1) the *conditional case* in which the variables 2 to p are fixed, and (2) the *unconditional case* in which all variables are random. Let $R_i^{*^2} = R_i^2/(1-R_i^2), i = 1,\ldots,k$. Gupta and Panchapakesan (1969) proposed the rule

$$R_7 : \text{ Select } \pi_i \text{ if and only if } R_i^{*^2} \geq c_7 R_{[k]}^{*^2} \tag{22}$$

for selecting the population associated with $\rho_{[k]}$, and the rule

$$R_7' : \text{ Select } \pi_i \text{ if and only if } R_i^{*^2} \leq \frac{1}{c_7'} R_{[1]}^{*^2} \tag{23}$$

for selecting the population associated with $\rho_{[1]}$, where $0 < c_7 = c_7(k, p, n-p, P^*) < 1$ and $0 < c_7' = c_7'(k, p, n-p, P^*) < 1$ are chosen suitably to meet the P^*-condition. The procedures proposed are the same for the conditional as well as the unconditional case. When $\rho_i \neq 0$, the distribution of R_i^{*2} is different in these two cases. However, the infimum of PCS occurs in either case when $\rho_1 = \cdots = \rho_k = 0$. The distribution of R_i^{*2} is the same in either case when $\rho_i = 0$. Thus, in either case, the constants c_7 and c_7' are given by

$$\int_0^\infty F_{2q,2m}^{k-1}\left(\frac{x}{c_7}\right) dF_{2q,2m}(x) = P^* \tag{24}$$

and

$$\int_0^\infty [1 - F_{2q,2m}(c_7' x)]^{k-1} dF_{2q,2m}(x) = P^* \tag{25}$$

where $q = (p-1)/2$, $m = (n-p)/2$, and $F_{2q,2m}(x)$ is the cdf of an F-variable with $2q$ and $2m$ degrees of freedom. The values of c_7 for selected values of k, P^*, m, and q are tabulated by Gupta and Panchapakesan (1969). The values of c_7' can be obtained from the same tables because $c_7'(p, q, m, P^*) = c_7(p, m, q, P^*)$.

3.4. Selection in Terms of Other Measures

Suppose the p variables under consideration are partitioned into two sets consisting of q_1 and $q_2 (q_1 + q_2 = p)$ variables. Let the corresponding partition of Σ_i be denoted by

$$\Sigma_i = \begin{pmatrix} \Sigma_{11}^{(i)} & \Sigma_{12}^{(i)} \\ \Sigma_{21}^{(i)} & \Sigma_{22}^{(i)} \end{pmatrix}, i = 1, \ldots, k.$$

Selection in terms of the conditional generalized variance of the q_2-set given the q_1-set has been considered by Gupta and Panchapakesan (1969). Frischtak (1973) discussed selection in terms $\gamma_i^2 = \frac{|\Sigma_i|}{|\Sigma_{11}^{(i)}||\Sigma_{22}^{(i)}|}$ but has obtained only an asymptotic solution.

For the problem of selecting populations that are better than a control, Krishnaiah (1967) used linear combinations of the elements of the covariance matrices for making comparisons. Krishnaiah and Rizvi (1966) used several linear combinations of the elements of the mean vectors for comparison and studied procedures to select a subset

containing good populations (defined through comparison with the control). For more details, reference can also be made to Gupta and Panchapakesan (1979).

4. SELECTION FROM A MULTINOMIAL POPULATION

Let $p_1, \ldots, p_k$ denote the unknown cell probabilities of a k-cell multinomial distribution. The ordered cell probabilities are denoted by $p_{[1]} \leq \cdots \leq p_{[k]}$. Gupta and Nagel (1967) proposed and studied procedures for selecting the most (least) probable cell based on a single sample of size n. Let $X_1, \ldots, X_k$ denote the cell counts. Their procedure for selecting the most probable cell is

$$R_8: \text{ Select the } i\text{th cell if and only if } X_i \geq X_{[k]} - D \tag{26}$$

and the procedure for selecting the least probable cell is

$$R_8': \text{ Select the } i\text{th cell if and only if } X_i \leq X_{[1]} + C \tag{27}$$

where $D = D(k, n, P^*)$ and $C = C(k, n, P^*)$ are the smallest nonnegative integers for which the P^*-condition is satisfied in each case.

An intereting point about R_8 and R_8' is that, unlike similar analogous rules for normal means, normal variances, etc., the analyses in the maximum and minimum cases do not run parallel. The LFC for either procedure is completely known only when $k = 2$. In this case, it is given by $p_1 = p_2 = \frac{1}{2}$. For $k > 2$, the LFC (in terms of the ordered p_i) is of the type $(0, \ldots, 0, s, p, \ldots, p)$, $s \leq p$, in the case of R_8 and is of the type $(p, \ldots, p, q)$, $p \leq q$, in the case of R_8'. An alternative to R_8 is the inverse sampling selection rule of Panchapakesan (1971, 1973). Observations are made one at a time until the cell count reaches a predetermined integer M in one of the cells. At termination, let $X_1, \ldots, X_k$ be the cell counts (one of them is M). The selection rule is

$$R_9: \text{ Select the } i\text{th cell if and only if } X_i \geq M - D \tag{28}$$

where $D (0 \leq D \leq M)$ is the smallest nonnegative integer for which the P^*-condition is satisfied. For R_9, the infimum of PCS occurs when all the cell probabilities are equal.

Again, for selecting the most probable cell, Gupta and Huang (1975) proposed the rule

$$R_{10}: \text{ Select the } i\text{th cell if and only if } X_i + 1 \geq cX_{[k]} \tag{29}$$

where $c = c(k, N, P^*) \epsilon (0, 1)$ is the largest number for which the P^*-condition is met. The motivation for the rule R_{10} comes from their conditional selection rules for Poisson populations. A conservative value of c can be obtained from their results for Poisson populations.

Recently, Chen (1985) considered an inverse sampling selection rule for selecting a subset containing the least probable cell. For his procedure R_{11} the observations are made one at a time until *either* (1) the count in any cell reaches r, *or* (2) $(k-1)$ cells reach count of at least $r'(1 \leq r' \leq r+1)$. If (1) occurs before (2), the rule R_{11} selects the cells with counts $X_i < r'$. If (2) occurs before (1), then R_{11} selects the cell with count $X_i < r'$. The constants r and r' are to be chosen so as to satisfy the P^*-condition. It has been shown by Chen (1985) that the infimum of $P(CS|R_{11})$ occurs when all the cell probabilities are equal.

Minimax subset selection rules have been investigated by Berger (1979) and Berger and Gupta (1980). For selecting the least probable cell, Berger (1980) investigated a minimax subset selection rule taking as loss the size of the selected subset or the number of non-best cells selected. In another paper, Berger (1982) investigated minimax and admissible subset selection rules for the least probable cell taking as the loss the number of non-best cells selected. His rule, however, satisfies the P^*-condition only if P^* is sufficiently large. For the corresponding procedure for the most probable cell, the P^*-condition has been verified only in certain special cases.

The importance of multinomial selection rules is accented by the fact that they provide distribution-free procedures. Suppose that $\pi_1, \ldots, \pi_k$ have continuous distributions F_{θ_i}, $i = 1, \ldots, k$. We assume that $\{F_\theta\}$ is a stochastically increasing family in θ. Let p_i denote the probability that in a set of k observations, one from each distribution, the observation from π_i is the largest, $i = 1, \ldots, k$. Selecting the stochastically largest (smallest) population is then equivalent to selecting the population associated with the largest (smallest) p_i. If we take observations a vector at a time and note which population yielded the largest observation, the problem can be converted to the multinomial cell problem.

5. SELECTION FROM SEVERAL MULTINOMIAL POPULATIONS

Let $\pi_1, \ldots, \pi_k$ be k multinomial populations each with m cells and let the unknown cell probabilities of π_i be $p_{i1}, \ldots, p_{im}$, $i = 1, \ldots, k$. Let $H_i \equiv H(p_{i1}, \ldots, p_{im}) = -\sum_{j=1}^{m} p_{ij} \log p_{ij}$, the Shannon entropy func-

tion associated with π_i. The function is a measure of the uncertainty with regard to the nature of the outcomes from π_i. We want to select the population associated with the largest H_i. For $m = 2$, the problem reduces to that of selecting the binomial population associated with the largest $\psi(\theta_i) = -\theta_i \log \theta_i - (1-\theta_i)\log(1-\theta_i)$, where θ_i is the success probability. In this case, Gupta and Huang (1976) proposed the rule

$$R_{12}: \quad \text{Select } \pi_i \text{ if and only if } \psi(\frac{X_i}{n}) \geq \max_{1 \leq j \leq k} \psi(\frac{X_j}{n}) - d_{12} \qquad (30)$$

where X_i is the number of successes in n trials associated with π_i, and $d_{12} = d_{12}(k, n, P^*)$ is the smallest nonnegative constant such that $0 < d \leq \psi([n/2]/n)$ for which the P^*-condition is satisfied. Here $[n/2]$ denotes the largest integer $\leq n/2$. The infimum of $P(CS|R_{12})$ takes place when $\theta_1 = \cdots = \theta_k = \theta$. However, the common value θ for which the infimum takes place is not known. Gupta and Huang (1976) have obtained a conservative value of d using the approach of Gupta, Huang and Huang (1975), who used this approach to obtain a conservative value for the constant defining the procedure of Gupta and Sobel (1960) for selecting the binomial population with the largest success probability. For more details on this, see Gupta and Panchapakesan (1979, 1985).

To discuss the selection procedure of Gupta and Wong (1977) in the case of $m > 2$, let $\underset{\sim}{a} = (a_1, \ldots, a_m)$ and $A_r = \sum_{i=r}^{m} a_{[i]}$, where $a_{[1]} \leq \cdots \leq a_{[m]}$ are the ordered components. Vector $\underset{\sim}{a} = (a_1, \ldots, a_m)$ is said to *majorize* vector $\underset{\sim}{b} = (b_1, \ldots, b_m)$ of the same dimension (written $\underset{\sim}{a} \succ \underset{\sim}{b}$) if $A_r \geq B_r$ for $r = 2, \ldots, m$, and $A_1 = B_1$. Further, a function f is said to be *Schur-concave* if $f(\underset{\sim}{x}) \leq f(\underset{\sim}{x}')$ whenever $\underset{\sim}{x} \succ \underset{\sim}{x}'$.

In our selection problem, we assume that there is a population whose associated vector of cell probabilities is majorized by the associated vector of cell probabilities of any other population. Such a population will have the largest H_i because the entropy function is Schur-concave. Let $\varphi_i = \varphi(\frac{X_{i1}}{n}, \ldots, \frac{X_{im}}{n})$, where φ is a Schur-concave function, and $X_{i1}, \ldots, X_{im}$ are the cell counts based on n independent observations from π_i, $i = 1, \ldots, k$. Gupta and Wong (1977) proposed the rule

$$R_{13}: \quad \text{Select } \pi_i \text{ if and only if } \varphi_i \geq \max_{1 \leq j \leq k} \varphi_j - d_{13} \qquad (31)$$

where $d_{13} = d_{13}(k, m, n, P^*)$ is the smallest positive constant for

which the P^*-condition is satisfied. Gupta and Wong obtained a conservative value of d using the idea of conditioning as in the paper of Gupta and Huang (1976).

6. SELECTION OF VARIABLES IN LINEAR REGRESSION

In applying regression analysis in practical situations for prediction purposes, we are often faced with a large number of independent variables. In such situations, it may be sufficient to consider a subset of these predictor variables for "adequate" prediction. There arises then a problem of choosing a "good" subset of these variables. Hocking (1976) and Thompson (1978a,b) have reviewed several criteria and techniques that have been used in practice. However, these are ad hoc procedures and are not designed to control the probability of selecting the important variables. McCabe and Arvesen (1974), and Arvesen and McCabe (1975) were the first to formulate this problem in the framework of Gupta-type subset selection.

Consider the standard linear model

$$\underset{\sim}{Y} = X\underset{\sim}{\beta} + \underset{\sim}{\epsilon} \tag{32}$$

where X is an $N \times p$ known matrix of rank $p \leq N$, $\underset{\sim}{\beta}$ is a $p \times 1$ parameter vector, and $\underset{\sim}{\epsilon} \sim N(0, \sigma^2 I_N)$. This model with p independent variables is considered as the "true" model. Now, consider all reduced models that are formed by taking all possible subsets of size $t (< p)$ from the p independent variables. These models are described by

$$\underset{\sim}{Y} = X_i\underset{\sim}{\beta}_i + \underset{\sim}{\epsilon}_i,\ i = 1, \ldots, k = \binom{p}{t}, \tag{33}$$

where X_i is an $N \times t$ matrix (of rank t), $\underset{\sim}{\beta}_i$ is a $t \times 1$ parameter vector, and $\underset{\sim}{\epsilon}_i \sim N(0, \sigma_i^2 I_N)$. It should be noted that the models in (33) are considered for prediction purposes and must be compared under the true model assumptions. The expectations of residual mean squares in the corresponding ANOVA evaluated under the true model assumption are σ_i^2, $i = 1, \ldots, k$. For the goal of selecting the design X_i (or the corresponding set of independent variables) associated with $\sigma_{[i]}^2$, Arvesen and McCabe (1975) proposed the rule

$$R_{14}: \text{ Select the design } X_i \text{ if and only if } SS_i \leq \frac{1}{c_{14}} SS_{[1]} \tag{34}$$

where SS_i is the residual sum of squares in the ANOVA corresponding to the design X_i, and $0 < c_{14} = c_{14}(p, t, N, P^*) < 1$ is to be chosen to satisfy the P^*-condition. An exact evaluation of the constant c_{14} is difficult. Arvesen and McCabe showed that the PCS is asymptotically $(N \to \infty)$ minimized when $\underset{\sim}{\beta} = \underset{\sim}{0}$. The evaluation of c_{14} is not easy even under this asymptotic LFC. An algorithm has been given by McCabe and Arvesen (1974) for determining c_{14} under the asymptotic LFC for given P^* and X, using Monte Carlo methods.

In the above formulation, the size t is arbitrarily fixed. Huang and Panchapakesan (1982) considered a different formulation taking into consideration all possible reduced models. They considered the regression model with $\underset{\sim}{\beta}' = (\beta_0, \ldots, \beta_p)$, and $X = (\underset{\sim}{1}\underset{\sim}{x}_1 \ldots \underset{\sim}{x}_{p-1})$, where $\underset{\sim}{1}' = (1, \ldots, 1)$ and $\underset{\sim}{x}_i' = (x_{i1}, \ldots, x_{iN})$, $i = 1, \ldots, p-1$. For fixed $\alpha \epsilon \{0, 1, \ldots, p-1\}$, consider all the $\binom{p-1}{\alpha}$ subsets of the set of predictor variables $\{x_1, \ldots, x_{p-1}\}$ and the corresponding reduced models obtained from (32). Associated with these reduced models are the multiple correlation coefficients $R_{i\alpha}$, $i = 1, 2, \ldots, \binom{p-1}{\alpha}$. Let $\theta_{i,\alpha} = E(1 - R_{i\alpha}^2)$. Any reduced model with the associated parameter $\theta_{i,\alpha}$ is said to be *inferior* if $\theta_{1,p-1} \leq \delta^* \theta_{i,\alpha}$, where $\delta^* \epsilon (0, 1)$ is a specified constant. (The parameter $\theta_{1,p-1}$ is associated with the true model). Huang and Panchapakesan (1982) considered the problem of eliminating all inferior models. A *correct decision* (CD) is selection of any subset of the models such that all inferior models are excluded from the selected subset. They proposed and studied the procedure

$$R_{15}: \quad \text{Exclude a model if and only if } \hat{\theta}_{i,\alpha} \geq \frac{c_{15}}{\delta^*} \hat{\theta}_{1,p-1} \qquad (35)$$

where $\hat{\theta}_{i,\alpha} = 1 - R_{i,\alpha}^2$, and the constant $c_{15} = c_{15}(N, p, P^*) > \delta^*$ is determined such that the P^*-condition is satisfied.

The LFC for the rule R_{15} has been established only in the asymptotic $(N \to \infty)$ sense. For evaluating the constant under the asymptotic LFC$(\underset{\sim}{\beta} = \underset{\sim}{0})$, Huang and Panchapakesan (1982) used an algorithm similar to that of McCabe and Arvesen (1974).

Hsu and Huang (1982) considered the goal of selecting a subset of the models that contains all the *superior* models, namely, all models for which $\sigma_i^2 \leq \Delta \sigma^2$, where $\Delta > 1$ is a specified constant. For this problem, they investigated a sequential procedure.

Gupta, Huang and Chang (1984) studied the problem of eliminating inferior models, using the expected mean squares as the criterion for comparing any model with the true model. Their approach is different from those of the earlier papers in that they use simultaneous

tests of a family of hypotheses in constructing their procedure.

Now, for any reduced model, it is known that SS_i/σ_0^2 has (under the full assumption model) a noncentral chi-square distribution with $\nu = N - p + 1$ degrees of freedom and a noncentrality parameter $\lambda_i = (X\beta)'Q_i(X\beta)/2\sigma_0^2$, where $Q_i = I_N - X_i(X_i'X_i)^{-1}X_i'$, and σ_0^2 is the error variance in the full model. Recently, Gupta and Huang (1986) have considered the problem of eliminating *inferior* models, namely, those for which $\lambda_i \geq \Delta > 0$, where Δ is specified in advance. For this problem, they have proposed and investigated a two-stage procedure.

7. CONCLUSION

As we have seen, multivariate selection problems have wider applications. However, in many cases, the existing procedures have not been fully examined in terms of their performances as well as the determination of the LFC. Even the multinomial problems have to be studied more satisfactorily. Also, the criterion employed for ranking multivariate populations usually induce a complete ordering in the space of distributions. However, in many practical problems, there is a need to consider a partial ordering. There has been practically no development in this direction. Also, there has been no work done for distributions other than multivariate normal populations. It will be interesting to consider reliability related models such as increasing failure rate distributions in two or more dimensions.

8. ACKNOWLEDGEMENT

This research was supported by the Office of Naval Research Contract N00014-84-C-0167 at Purdue University. Reproduction in whole or part is permitted for any purpose of the United States Government.

Shanti S. Gupta
Statistics Department
Purdue University
West Lafayette, IN 47907

S. Panchapakesan
Department of Mathematics
Southern Illinois University
Carbondale, IL 62901

REFERENCES

Armitage, J. V. and Krishnaiah, P. R. (1964). 'Tables for the studentized largest chi-square distribution and their applications'. *ARL 64–188*, Aerospace Research Laboratories, Wright-Patterson Air Force Base, Dayton, Ohio.

Arvesen, J. N. and McCabe, G. P., Jr. (1975). 'Subset selection problems of variances with applications to regression analysis'. *Journal of the American Statistical Association*, **70**, 166–170.

Bechhofer, R. E. (1954). 'A single-sample multiple decision procedure for ranking means of normal populations with known variances'. *Annals of Mathematical statistics*, **25**, 16–39.

Berger, R. L. (1979). 'Minimax subset selection for loss measured by subset size'. *Annals of Statistics*, **7**, 1333–1338.

Berger, R. L. (1980). 'Minimax subset selection for multinomial distribution'. *Journal of Statistical Planning and Inference*, **4**, 391–402.

Berger, R. L. (1982). 'A minimax and admissible subset selection rule for the least probable multinomial cell'. *Statistical Decision Theory and Related Topics – III*, Vol. 2 (S. S. Gupta and J. O. Berger, eds.), Academic Press, New York, 143–156.

Berger, R. L. and Gupta, S. S. (1980). 'Minimax subset selection rules with applications to unequal variance (unequal sample size) problems'. *Scandinavian Journal of Statistics*, **7**, 21–26.

Chattopadhyay, A. K. (1981). 'Selecting the normal population with largest (smallest) value of Mahalanobis distance from the origin'. *Communications in Statistics – Theory and Methods*, **A10**, 31–37.

Chen, P. (1985). 'Subset selection for the least probable multinomial cell'. *Annals of the Institute of Statistical Mathematics*, **37**, 303–314.

Dudewicz, E. J. and Koo, J. O. (1982). *The Complete Categorized Guide to Statistical Selection and Ranking Procedures.* Series in Mathematical and Management Sciences, Vol. 6, American Sciences Press, Inc., Columbus, Ohio.

Frischtak, R. M. (1973). *Statistical Multiple Decision Procedures for Some Multivariate Selection Problems.* Ph.D. Thesis (also Technical Report No. 187), Department of Operations Research, Cornell University, Ithaca, New York.

Gnanadesikan, M. (1966). *Some Selection and Ranking Procedures for Multivariate Normal Populations.* Ph.D. Thesis, Department of Statistics, Purdue University, West Lafayette, Indiana.

Gnanadesikan, M. and Gupta, S. S. (1970). 'A selection procedure for multivariate normal distributions in terms of the generalized variances'. *Technometrics*, **12**, 103–117.

Gupta, S. S. (1956). *On A Decision Rule for A Problem in Ranking Means.* Ph.D. Thesis (also Mimeograph Series No. 150), Institute

of Statistics, University of North Carolina, Chapel Hill, North Carolina.

Gupta, S. S. (1963a). 'Probability integrals of the multivariate normal and multivariate t'. *Annals of Mathematical Statistics*, **34**, 792–828.

Gupta, S. S. (1963b). 'On a selection and ranking procedure for gamma populations'. *Annals of the Institute of Statistical Mathematics*, **14**, 199–216.

Gupta, S. S. (1965). 'On some multiple decision (selection and ranking) rules. *Technometrics*, **7**, 225–245.

Gupta, S. S. (1966). 'On some selection and ranking procedures for multivariate normal populations using distance functions'. *Multivariate Analysis* (P. R. Krishnaiah, ed.), Academic Press, New York, 457–475.

Gupta, S. S. and Huang, D.-Y. (1975). 'On subset selection procedures for Poisson populations and some applications to the multinomial selection problems'. *Applied Statistics* (R. P. Gupta, ed.), North-Holland, Amsterdam, 97–109.

Gupta, S. S. and Huang, D.-Y. (1976). 'On subset selection procedures for the entropy function associated with the binomial populations'. *Sankhya*, **38A**, 153–173.

Gupta, S. S. and Huang, D.-Y. (1986). 'Selecting important independent variables in linear regression models'. *Technical Report No. 86-29*, Department of Statistics, Purdue University, West Lafayette, Indiana.

Gupta, S. S., Huang, D.-Y., and Chang, C.-L. (1984). 'Selection procedures for optimal subsets of regression variables'. *Design of Experiments: Ranking and Selection*, (T. J. Santner and A. C. Tamhane, eds.), Marcel Dekker, New York, 67–75.

Gupta, S. S., Huang, D.-Y., and Huang, W.-T. (1975). 'On ranking and selection procedures and tests of homogeneity for binomial populations'. *Essays in Probability and Statistics* (S. Ikeda, T. Hayakawa, H. Hudimoto, M. Okamoto, M. Siatoni and S. Yamamoto, eds.), Shinko Tsusho Co. Ltd., Tokyo, Japan, Chapter 33, 501–533.

Gupta, S. S. and Nagel, K. (1967). 'On selection and ranking procedures and order statistics from the multinomial distribution'. *Sankhya*, **29B**, 1–34.

Gupta, S. S., Nagel, K. and Panchapakesan, S. (1973). 'On the order statistics from equally correlated normal random variables'. *Biometrika*, **60**, 403–413.

Gupta, S. S. and Panchapakesan, S. (1969). 'Some selection and ranking procedures for multivariate normal populations'. *Multivariate Analysis - II* (P. R. Krishnaiah, ed.), Academic Press, New York, 475–505.

Gupta, S. S. and Panchapakesan, S. (1979). *Multiple Decision Procedures: Theory and Methodology of Selecting and Ranking Populations*. John Wiley & Sons, Inc., New York.

Gupta, S. S. and Panchapakesan, S. (1985). 'Subset selection procedures: review and assessment'. *American Journal of Mathematical and Management Sciences*, **5**, 235–311.

Gupta, S. S., Panchapakesan, S. and Sohn, J. K. (1985). 'On the distribution of the studentized maximum of equally correlated normal random variables'. *Communications in Statistics – Simulation and Computation*, **14**, 103–135.

Gupta, S. S. and Sobel, M. (1957). 'On a statistic which arises in selection and ranking problems'. *Annals of Mathematical Statistics*, **28**, 957–967.

Gupta, S. S. and Sobel, M. (1960). 'Selecting a subset containing the best of several binomial populations'. *Contributions to Probability and Statistics* (I. Olkin, S. G. Ghurye, W. Hoeffding, W. G. Madow and H. B. Mann, eds.), Stanford University Press, Stanford, California, Chapter 20, 224–248.

Gupta, S. S. and Sobel, M. (1962a). 'On selecting a subset containing the population with the smallest variance'. *Biometrika*, **49**, 495–507.

Gupta, S. S. and Sobel, M. (1962b). 'On the smallest of several correlated F-statistics'. *Biometrika*, **49**, 509–523.

Gupta, S. S. and Studden, W. J. (1970). 'On a ranking and selection procedure for multivariate populations'. *Essays in Probability and Statistics* (R. C. Bose, I. M. Chakravarti, P. C. Mahalanobis, C. R. Rao and K. J. C. Smith, eds.), University of North Carolina, Chapel Hill, North Carolina, Chapter 16, 327–338.

Gupta, S. S. and Wong, W.-Y. (1977). 'Subset selection procedures for finite schemes in information theory'. *Colloquia Mathematica Societatis János Bolyai, 16: Topics in Information Theory* (I. Csiszár and P. Elias, eds.), 279–291.

Hocking, R. R. (1976). 'The analysis and selection of variables in regression analysis'. *Biometrics*, **32**, 1–49.

Hsu, T.-A. and Huang, D.-Y. (1982). 'Some sequential selection procedures for good regression models'. *Communications in Statistics – Theory and Methods*, **A11**, 411–421.

Huang, D.-Y. and Panchapakesan, S. (1982). 'On eliminating inferior regression models'. *Communications in Statistics – Theory and Methods*, **A11**, 751–759.

Krishnaiah, P. R. (1967). 'Selection procedures based on covariance matrices of multivariate normal populations'. *Blanch Anniversary Volume*, Aerospace Reearch Laboratories, U. S. Air Force Base, Dayton, Ohio, 147–160.

Krishnaiah, P. R. and Armitage, J. V. (1964). 'Distribution of the stu-

dentized smallest chi-square, with tables and applications'. *ARL 64–218*, Aerospace Research Laboratories, Wright-Patterson Air Force Base, Dayton, Ohio.

Krishnaiah, P. R. and Armitage, J. V. (1966). 'Tables for multivariate t-distribution'. *Sankhya*, **28B**, 31–56.

Krishnaiah, P. R. and Rizvi, M. H. (1966). 'Some procedures for selection of multivariate normal populations better than a control'. *Multivariate Analysis* (P. R. Krishnaiah, ed.), Academic Press, New York, 477–490.

McCabe, G. P., Jr. and Arvesen, J. N. (1974). 'Subset selection procedures for regression variables'. *Journal of Statistical Computation and Simulation*, **3**, 137–146.

Panchapakesan, S. (1971). 'On a subset selection procedure for the most probable event in a multinomial distribution'. *Statistical Decision Theory and Related Topics* (S. S. Gupta and J. Yackel, eds.), Academic Press, New York, 275–298.

Panchapakesan, S. (1973). 'On a subset selection procedure for the best multinomial cell and related problems'. Abstract. *Bulletin of the Institute of Mathematical Statistics*, **2**, 112–113.

Regier, M. H. (1976). 'Simplified selection procedures for multivariate normal populations'. *Technometrics*, **18**, 483–489.

Slepian, D. (1962). 'On the one-sided barrier problem for Gaussian noise'. *Bell System Technical Journal*, **41**, 463–501.

Thompson, M. L. (1978a). 'Selection of variables in multiple regression: Part I. A review and evaluation'. *International Statistical Review*, **46**, 1–19.

Thompson, M. L. (1978b). 'Selection of variables in multiple regression: Part II. Chosen procedures, computations and examples'. *International Statistical Review*, **46**, 129–146.

BEFORE COMPLETING FORM

1. REPORT NUMBER	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
Technical Report #86-52		

4. TITLE (and Subtitle)
STATISTICAL SELECTION PROCEDURES IN MULTIVARIATE MODELS

5. TYPE OF REPORT & PERIOD COVERED
Technical

6. PERFORMING ORG. REPORT NUMBER
Technical Report #86-52

7. AUTHOR(s)
Shanti S. Gupta and S. Panchapakesan

8. CONTRACT OR GRANT NUMBER(s)
N00014-84-C-0167

9. PERFORMING ORGANIZATION NAME AND ADDRESS
Purdue University
Department of Statistics
West Lafayette, Indiana 47907

10. PROGRAM ELEMENT, PROJECT, TASK, AREA & WORK UNIT NUMBERS

11. CONTROLLING OFFICE NAME AND ADDRESS
Office of Naval Research
Washington, DC

12. REPORT DATE
November 1986

13. NUMBER OF PAGES
22

14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)

15. SECURITY CLASS. (of this report)
Unclassified

15a. DECLASSIFICATION, DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)
Approved for public release, distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)
Subset selection, multivariate normal populations, Mahalanobis distance, generalized variance, multiple correlation coefficient, multinomial, Shannon entropy, linear regression, important predictor variables.

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)
Selection and ranking problems have been studied over the last thirty years, generally under one of two formulations: Bechhofer's indifference zone approach and Gupta's subset selection approach. This paper deals with subset selection. Subset selection procedures in multivariate models are briefly reviewed. These include: (1) Procedures for selecting the best component in a multivariate normal population in terms of the component means as well as the component variances (Section 2), (2) Procedures for selecting the best from several multivariate normal populations in terms of (i) the Mahalanobis distance, (ii) the generalized

DD FORM 1 JAN 73 1473

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

variance, and (iii) the multiple correlation coefficient (Section 3), (3) Procedures (fixed sample size as well as inverse sampling) for selecting the most (least) probable cell in a multinomial distribution (Section 4), (4) Procedures for selecting the best from several multinomial populations in terms of the Shannon entropy function (Section 5), and (5) Procedures for choosing the best subset of the predictor variables in a linear regression model (Section 6).