A NOTE ON OPTIMAL SUBSET SELECTION PROCEDURES

by

Shanti S. Gupta, Purdue University

and

Deng-Yuan Huang, Academia Sinica, Taipei, Taiwan

Department of Statistics
Division of Mathematical Sciences
Mimeograph Series #470

Revised, June 1978

This research was supported by Office of Naval Research Contract N00014-75-C-0455 at Purdue University. Reproduction in whole or in part is permitted for any purpose of the United States Government.
1. Introduction. A common problem faced by an experimenter is one of comparing several categories or populations. The classical approach to this problem has been to use a test for homogeneity, i.e., to test whether all categories (populations) are identical or not. This approach is not always realistic and it is often inadequate. The inadequacy lies in the fact that only two decisions, accept or reject the hypothesis, are available to the experimenter. The experimenter is then faced with the problem of what to do next, especially if the hypothesis is rejected. These difficulties and inadequacies may be alleviated by formulating the problem as multiple decision problems aimed at selection or ranking (ordering) of the k populations. This has led to the rapid development of selection and ranking theory during the last two decades. Many reasonable rules have been proposed. Some desirable properties of these rules have been studied. However, very little work has been done to study the optimality of selection procedures, especially in the subset selection approach. In this paper, we are interested in deriving some methods to construct optimal subset selection procedures. Some classical selection rules are constructed as special cases.

Let \( \pi_1, \pi_2, \ldots, \pi_k \) represent \( k(\geq 2) \) populations (treatments) and let \( X_{11}, \ldots, X_{1n_1}, \ldots, X_{kn_k} \) be \( n_1, \ldots, n_k \) independent random observations from \( \pi_i \). The quality of

*This research was supported by Office of Naval Research Contract N00014-75-C-0455 at Purdue University. Reproduction in whole or in part is permitted for any purpose of the United States Government.
the ith population \( \pi_i \) is characterized by a real-valued parameter \( \theta_i \), usually unknown. Let \( \Omega = \{ \Omega^o = (\theta_1, \ldots, \theta_k) \} \) denote the whole parameter space.

Let \( \tau_{ij} = \tau_{ij}(0) \) be a measure of separation between \( \pi_i \) and \( \pi_j \). We assume that there exists a monotonically non-increasing function \( h \) such that \( \tau_{ij} = h(\tau_{ij}) \).

Let \( \Omega_i = \{ \theta | \tau_{ij} \geq \tau_0, j \neq i \}, 1 \leq i \leq k, \) and \( \Omega_0 = \Omega_i \), where \( \Omega_0 = \bigcup_{i=1}^{k} \Omega_i \). Thus \( \Omega_i \) is a partition of \( \Omega \). For this problem, we assume \( \tau_0 \) and \( \tau_{ij} \) as known with \( \tau_0 > \tau_{ij} \) for all \( i \). Let \( \tau_i = \min_{j \neq i} \tau_{ij}, 1 \leq i \leq k \). We define \( \tau^* = \max_{1 \leq i \leq k} \tau_i \). The population associated with \( \tau^* \) will be called the best population. It should be pointed out that if \( \theta \in \Omega_i \), then \( \tau_{ij} \geq \tau_j \) for all \( j \), since for some \( j, j \neq i \), \( \tau_{ij} = h(\tau_{ij}) \leq h(\tau_0) \geq h(\tau_{ij}) = \tau_{ij} \). Thus if \( \theta \in \Omega_i \), then \( \pi_i \) is the best population. A selection of a subset containing the best population is called a correct selection (CS). (Note that in case of ties any one of the best populations corresponding to \( \tau^* \) is "tagged" as the best population.) To illustrate the above notation, we assume that the observations are drawn from \( \pi_i \) which has a normal distribution with unknown mean \( \theta_i \) \( (i = 1, \ldots, k) \) and known variance \( \sigma^2 \). We can define \( \tau_{ij} = \theta_i - \theta_j \); then it can be seen that \( \tau_i = \theta_i - \theta[k] \) if \( \theta_i < \theta[k] \) and \( \tau_i = \theta_i - \theta[k-1] \) if \( \theta_i = \theta[k] \), where \( \theta[1] \leq \ldots \leq \theta[k] \). In this case, \( \tau_{ij} = 0 \) for all \( i \). Thus the population with the largest mean, \( \theta[k] \), is the best. If instead \( \tau_{ij} = \theta_j - \theta_i \), then the population with the smallest mean \( \theta[1] \) would be the best. In the above example, we have \( h(t) = -t \) which is a decreasing function.

Let the observed sample vector be denoted by \( X' = (X_1', \ldots, X_k') \) where \( X_i \) has components \( X_{i1}, \ldots, X_{in} \), \( i = 1, \ldots, k \). Let \( \delta = (\delta_1, \ldots, \delta_k) \) be a selection procedure where \( \delta_i(x) \) is the probability of selecting \( \pi_i \) \( (1 \leq i \leq k) \) based on the observed vector \( x \). As measures of goodness or
optimality of a selection rule, consider two quantities (cf. Lehmann [1])
\begin{align*}
R(\varnothing, \delta) \text{ and } S(\varnothing, \delta). \quad \text{We define } S(\varnothing, \delta) & = P_0(CS|\delta) \text{ and } R(\varnothing, \delta) = \sum_{i=1}^{k} R(i)(\varnothing, \delta_i),
\end{align*}
where \( R(i)(\varnothing, \delta_i) = P(\text{Selecting } \pi_i|\delta_i) \). Let \( S \) be the size of the selected subset. Thus \( R(\varnothing, \delta) = E(S|\delta) \). For a specified \( \gamma, (0 < \gamma < 1) \), we may restrict our attention to the class \( \mathcal{Q} \) of all \( \delta \) such that
\begin{align*}
(1) \quad S(\varnothing, \delta) & \geq \gamma \quad \text{for } \varnothing \in \mathcal{Q}.
\end{align*}
We are interested in constructing an optimal procedure \( \delta^0 \) in \( \mathcal{Q} \) which minimizes the supremum of \( R(\varnothing, \delta) \) over \( \mathcal{Q} \) for all \( \delta \in \mathcal{Q} \), i.e.,
\begin{align*}
(2) \quad \sup_{\varnothing \in \mathcal{Q}} R(\varnothing, \delta^0) & = \min \sup_{\delta \in \mathcal{Q}} R(\varnothing, \delta).
\end{align*}
We restrict attention to those selection procedures which depend upon the observations only through a sufficient and maximal invariant statistic \( Z_{ij} \) which are defined as follows:
\begin{align*}
Z_{ij} & = f(X_{i1}, \ldots, X_{in_i}; X_{j1}, \ldots, X_{jn_j}).
\end{align*}
This \( Z_{ij} \) is based on the \( n_i \) and \( n_j \) observations from \( \pi_i \) and \( \pi_j \) \((i,j = 1,2,\ldots,k)\), respectively. It is well known that the distribution of \( Z_{ij} \) depends only on \( T_{ij} \). For any \( i \), let the joint density of \( Z_{ij}, j \neq i \), be \( p_0(z_i) \). Let \( p_0(z_i) \) be denoted by \( p_0(z_1) \) when \( T_{11} = \ldots = T_{ik} = T_{ii} = \text{constant} \) and by \( p_i(z_1) \) when \( T_{11} = \ldots = T_{ik} = T_0, 1 \leq i \leq k \). In a given problem the function \( f \) is so chosen as to indicate the measure of separation between the populations in a reasonable way. In case of the above normal means example, a choice of \( Z_{ij} \) might be \( \bar{X}_i - \bar{X}_j \), where \( \bar{X}_i = \frac{1}{n_i} \sum_{k=1}^{n_i} X_{ik} \) and \( \bar{X}_j = \frac{1}{n_j} \sum_{k=1}^{n_j} X_{jk} \). Let \( \nu \) be a \( \sigma \)-finite measure on \( \mathbb{R}^{k-1} \).
Oosterhoff [2] defines a monotone likelihood ratio for a random vector 
\( \mathbf{X} \) with \( m \) components as follows: Let \( \theta \) be an \( m \)-dimensional vector of parameters. A partial ordering of points in \( \mathbb{R}^m \) is defined by \( x_1 \prec x_2 \), \( x'_1 = (x'_{11}, \ldots, x'_{1m}) \), 
\[ i = 1, 2 \text{, meaning that } x_{1j} \leq x_{2j} \text{ for } j = 1, 2, \ldots, k, \text{ and the inequality is } \]
strict for at least one component. The density \( f_{\theta}(x) \) has monotone likelihood ratio (MLR) if for all \( \theta_1 \prec \theta_2 \), \( f_{\theta_2}(x)/f_{\theta_1}(x) \) is nondecreasing in \( x \).

Now we state and prove a theorem which provides a solution to the restricted minimax problem as stated in (1) and (2) (cf. Lehmann [1]).

**Theorem:** Suppose that the density \( p_\theta(z) \) has the MLR property. If \( R(\theta, \delta^0) \)
is maximized at \( \tau_{ij} = \tau_{ii} = \text{constant} \), for all \( i,j \), where \( \delta^0 \) is given by
\[
\delta_i^0(z_i) = \begin{cases} 
1 & \text{if } p_i(z_i) > c p_0(z_i), \\
\lambda_i & \text{if } c p_i(z_i) = c p_0(z_i), \\
0 & \text{if } c p_i(z_i) < c p_0(z_i), 
\end{cases}
\]
such that \( c(>0) \) and \( \lambda_i \) are determined by \( \int \delta_i^0 p_i(z_i) \) = \( \gamma \), \( 1 \leq i \leq k \). Then 
\( \delta^0 = (\delta_1^0, \ldots, \delta_k^0) \) minimizes \( \sup R(\theta, \delta) \) subject to \( \inf S(\theta, \delta) \geq \gamma \).

**Proof.** For any \( \delta \in \mathcal{S} \),
\( \theta \in \mathcal{B} \) implies \( \theta \in \mathcal{B}_i \) for some \( i \), thus
\[
S(\theta, \delta) = \int \delta_i(z_i) p_\theta(z_i) d\nu(z_i) \geq \min_{1 \leq i \leq k} \inf_{\theta \in \mathcal{B}_i} \int \delta_i(z_i) p_\theta(z_i) d\nu(z_i).
\]
We have
\[
\inf_{\theta \in \mathcal{B}_i} S(\theta, \delta) = \min_{1 \leq i \leq k} \inf_{\theta \in \mathcal{B}_i} \int \delta_i(z_i) p_\theta(z_i) d\nu(z_i).
\]
Hence for any \( \delta \in \mathcal{S} \), \( \inf_{\theta \in \mathcal{B}_i} \int \delta_i(z_i) p_\theta(z_i) d\nu(z_i) \geq \gamma \), \( 1 \leq i \leq k \), and by the assumption that \( \int \delta_i^0 p_i = \gamma \), it follows that
\[ f(\delta_i - \delta_0)(p_i - cp_0) \leq 0 \]

which implies
\[ f\delta_i p_0 \leq f\delta_i p_0. \]

By the assumption, \( \delta_0(z_i) \) is nondecreasing in \( z_i \), we have
\[ \inf_{\theta \in \Omega} S(\theta, \delta_0) = \min_{1 \leq i \leq k} f\delta_i p_i = \gamma. \]

If \( R(\theta, \delta) \) is maximized at \( \tau_{ij} = \tau_{ij} = \text{constant}, \) for all \( i, j \), then
\[ \sup_{\theta \in \Omega} R(\delta, \delta) \geq \sum_{i=1}^{k} f\delta_i p_0 \geq \sum_{i=1}^{k} f\delta_i p_0 = \sup_{\theta \in \Omega} R(\theta, \delta), \]

which completes the proof.

Example: Let \( \pi_1, \pi_2, \ldots, \pi_k \) be \( k \) independent normal populations with means \( \theta_1, \ldots, \theta_k \) and common variance \( \sigma^2 = 1 \). Our interest is to select a nonempty subset of the \( k \) populations containing the best. The ordered \( \theta_i \)'s are denoted by \( \theta[1] \leq \ldots \leq \theta[k] \). It is assumed that there is no prior knowledge of the correct pairing of the ordered and the unordered \( \theta_i \)'s. Our goal is to select a nonempty subset of the \( k \) populations so as to include the population associated with \( \theta[k] \).

Let \( \bar{x}_i, 1 \leq i \leq k \), denote the sample means of independent samples of size \( n \) from these populations. Let the joint likelihood function of \( \bar{x}_i, i=1,2,\ldots, k \), be
\[ g_\theta(\bar{x}) = \prod_{j=1}^{k} g_{\theta_i}(\bar{x}_i), \]

where
\[ g_{\theta_i}(\bar{x}_i) = \frac{n}{\sqrt{2\pi}} e^{-\frac{n}{2} (\bar{x}_i - \theta_i)^2}, \quad 1 \leq i \leq k. \]

Let \( \tau_{ij} = \tau_{ij}(\theta) = \theta_i - \theta_j, 1 \leq j \leq k, j \neq i, \tau_{ii} = 0, \tau_0 = \Delta > 0 \) and
\[ \tau_{ij} = \bar{x}_i - \bar{x}_j, j \neq i. \]

Let \( \tau' = (z_{i1}, \ldots, z_{ik}) \) and \( \bar{z}_i = (\tau_{i1}, \ldots, \tau_{ik}) \).
\[
\frac{k-1}{2} p_0(z_i) = (2\pi)^{-\frac{k}{2}} |\Sigma|^{-1/2} \exp\left( -\frac{1}{2} (z_i - \overline{z}_i)' \Sigma^{-1} (z_i - \overline{z}_i) \right),
\]

where \( \Sigma_{(k-1)\times(k-1)} = \frac{2}{n} \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix} \) is the positive definite covariance matrix of \( Z_{ij} \)'s. We know that

\[
\frac{p_1(z_i)}{p_0(z_i)} = \exp\left( z_i' \Delta' \Sigma^{-1} z_i - \Delta' \Sigma^{-1} \Delta \right)
\]

is nondecreasing in \( Z_{ij} \), \( j \neq i \), where \( \Delta' = (\Delta_1, \ldots, \Delta_k) \). And

\[
\frac{p_1(z_i)}{p_0(z_i)} > c
\]

is equivalent to

\[
(3) \quad \tilde{x}_i > \frac{1}{k-1} \sum_{j \neq i} \tilde{x}_j + d.
\]

For any \( i \), let \( \mu_i = \sum_{j \neq i} (\theta_j - \theta_i) \), we have

\[
\mu_i = \mu - k\theta_i
\]

where \( \mu = \sum_{j=1}^k \theta_j \). Hence, \( \mu_1 = \mu_2 = \ldots = \mu_k \) if and only if \( \theta_1 = \theta_2 = \ldots = \theta_k \).

We order \( \mu_i \)'s as \( \mu[1] \leq \ldots \leq \mu[k] \). Since

\[
R(\theta, \delta) = \sum_{i=1}^k P(\tilde{x}_i > \frac{1}{k-1} \sum_{j \neq i} \tilde{x}_j + d)
\]

\[
= \sum_{i=1}^k \int_0^{\infty} \phi(\sqrt{k-1} (y - \frac{1}{k-1} \mu_i - d)) d\phi(y)
\]

\[
= f(\mu_1, \ldots, \mu_k)
\]

\[
\leq f(\mu[1], \ldots, \mu[1])
\]

\[
= f(\theta[k], \ldots, \theta[k])
\]
= f(0, ..., 0)
- \( R(\hat{0}, \delta) \),

hence, it follows that the supremum of \( R(\hat{0}, \delta^0) \) over \( \Omega \) occurs at \( 0_1 = \ldots = 0_k \),
i.e., \( \hat{\tau}_{ij} = \hat{\tau}_{ii} = 0 \). Thus the result of the theorem can be applied.

Note that the above procedure (3) is a rule of the type proposed by Seal [3] to select a subset containing the population associated with the largest \( \theta_i \)'s.

Acknowledgment

The authors wish to thank the referees for their valuable comments and suggestions which have improved and simplified the presentation.

References


In this paper, we are concerned with the construction of "optimal" subset selection procedures. Some classical selection procedures are considered as special cases.