ON SOME DECISION-THEORETIC CONTRIBTUIONS TO THE PROBLEM OF SUBSET SELECTION*

by

Jason C. Hsu
Purdue University

Department of Statistics
Division of Mathematical Sciences
Mimeograph Series #491

June 1977

*This research was supported in part by the Office of Naval Research contract N00014-75-C-0455 at Purdue University.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

Ranking and selection procedures, subset selection procedures in particular, provide in a realistic manner attractive ways of handling problems that are commonly treated by the 2-action procedure of a global F-test, and the many many-action procedure of a typical multiple range test. Consider the common one-way layout situation in analysis of variance. Usually the experimenter wants to know more than just whether all the treatment effects are equal, but he may not want to make inferences concerning all pair-wise differences of means, or all linear contrasts of means. One of the more frequently occurring situations for which this is so is where the experimenter simply wishes to know which of the treatments gives the best product. In this situation, formulating the problem as a selection problem is appropriate. Subset selection procedures are often thought of as screening procedures. If the data indicate several treatments are better than the remaining treatments but no treatment is clearly the best, then perhaps the experimenter ought to retain all of the better treatments for future consideration. Of course the concept of subset selection has a much wider scope of application than just one-way layouts. We shall give two examples to show how situations for which the subset selection formulation is appropriate arise in practice.

The first example is adapted from an experiment conducted here at Purdue. Suppose we wish to determine how people perceive various colors. For instance, would most people perceive red as being hot? gray as being cold? Results of this type of experiments have been applied in practice. For example, certain fast food chain paints all its restaurants in certain colors because studies have shown people tend to leave the premise more quickly if the premise is painted in those colors. So suppose n experimental subjects are chosen and there are k available colors. Corresponding to each adjective of interest each subject chooses one of the k colors that he or she preceives to fit the adjective most closely. Then for this experiment the underlying distribution is the multinomial distribution with the number of observations equal to n and the number of cells equal to k. For the type of application mentioned one wants to select the color or colors corresponding to the cell with the highest frequency of occurance. Hence the subset selection formulation is appropriate.

The second example arose in the field of Bionucleonics. This was an actual experiment that the author came across in consulting. In manufacturing radioactive trace elements carrier particles were to be made from a petroleum base by the use of heavy pressure. When injected into the patient's body, particles that are too large get absorbed by the wrong organs and those too small go out of the system. There were four possible pressure settings. To each pressure setting there corresponds a particle size distribution. The object here was to select the pressure setting such that for the same total amount of radiation the amount of radiation attached to particles in the desirable size range is the largest. Hence the subset selection formulation is appropriate. Notice here the parameter of interest can be an extremely complex function of the theoretical particle size distribution.

Heuristically proposed 'subset selection' procedures of Gupta (1956, 1965) have been in existence for some time. For related work and thinking along subset selection lines, reference should be made to Paulson (1945) and Seal (1955). However, unlike the F-test and the multiple range tests, the use of these procedures in practice have been virtually nil. This, the author believes, can be attributed to two main reasons:

1) No computer packages exist to facilitate the use of these subset selection procedures. None of the commonly used statistical packages (e.g., SPSS, BMD) includes subset selection procedures as part of the package.

2) Research concerning the performance of these heuristically proposed procedures is inadequate. Potential users can not, in general, be guaranteed any optimality properties of these procedures.

It is generally recognized that for multivariate problems uniformly best procedures usually do not exist. In fact in most of the situations of practical interest, there do not even exist uniformly best unbiased procedures. Hence it is reasonable to look for procedures that do well on the average, averaged over the parameter space by some prior. This approach has been taken in the first part of the thesis. The essentially complete class of Bayes procedures and their limits is investigated. The concept of Total Monotone Likelihood Ratio is introduced as the multivariate analog of univariate monotone likelihood ratio. Then a multivariate analog of the classical univariate result of Karlin and Rubin (1956) that monotone procedures form an essentially complete class, is proved for a loss function which seems natural to the subset selection problem by proving that Bayes procedures are monotone.

Bayes procedures typically require numerical integrations to implement and this makes them unsuitable for practical use. Besides, the use of Bayes procedures is by no means universally accepted. So if there is available an easy to implement procedure whose performance is close to that of the Bayes procedure, then this procedure ought to be used.

This possibility is explored in Chapter 3 for the case of normal populations problem and normal exchangeable priors. As it turns out, Gupta's procedure is good compared to the Bayes procedure throughout the range of the normal prior while Seal's procedure is good only when the normal prior is concentrated, that is, when the normal prior is very informative. As of yet we do not know how these procedures perform when the priors are not normal, in particular when the priors have longer tails than the normal distribution. But from what we know Gupta's procedure seems to be the logical choice when the observations arise from normal distributions.

There are heuristically proposed procedures for many other distributions in the exponential family of distributions. Little is known concerning the performance of these procedures. They really have to be investigated case by case. But in the case where the parameter of interest is a location parameter and the underlying distribution is not entirely known there are known good robust estimators of the parameter. Under mild regularity conditions they are asumptotically normal. From the results of Chapter 3, one would expect Gupta's procedure based on these robust estimators to be asymptotically good. In Chapter 4 of the thesis robust and nonparametric versions of Gupta's procedure are proposed and their performance studied. One procedure in particular, the procedure based on simultaneous confidence bounds derived from rank tests, is a truly nonparametric subset selection procedure. It controls the infimum of the probability of a correct selection for any sample size. Because it is based essentially on the Hodges-Lehmann estimator it also has good asymptotic performance.

CHAPTER 1

SOME DECISION-THEORETIC PRELIMINARIES AND KNOWN RESULTS

In this chapter we give some decision-theoretic preliminaries and list some known results particularly those applicable to finite action problems. Although we shall make use of only one of the results in this thesis, namely Bayes procedures and their limits form an essentially complete class, it seems desirable to have the important results listed in an orderly fashion for the benefit of future workers in the field. We want to emphasize that these results pertain to all finite action problems. Hence they are applicable to the classification problem, the identification problem, the complete ranking problem, the treatments versus control problem, the selection problem using the indifference zone approach, and the selection problem using the subset selection approach. We follow throughout the development in Brown (1974).

We begin by describing in a mathematically precise fashion the formulation of the statistical decision problem.

Definition 1.1. The sample space S is a measurable space with σ-field $\mathcal{B}_S$.

Definition 1.2. A parameter space Φ is a measurable space with σ-field $\mathcal{B}_\Phi$.

Notation. We denote by $P(\mathcal{B})$ the set of all probability measures on the σ-field $\mathcal{B}$.

Definition 1.3. A parametrized set of possible distribution is a $\mathcal{B}_\Phi$ measurable map from Φ to $P(\mathcal{B}_S)$. We denote the value of this map at a pair $\phi\varepsilon\Phi$, $S\varepsilon\mathcal{B}_S$ by $F(S|\phi)$. Note that to say $F(\cdot|\cdot)$ is $\mathcal{B}_\Phi$ measurable means:

(1) For each $\phi\varepsilon\Phi$, $F(\cdot|\phi)$ is a probability distribution on $\mathcal{B}_S$.

(2) For each $S\varepsilon\mathcal{B}_S$, $F(S|\cdot)$ is a measurable map of $(\Phi,\mathcal{B}_\Phi)$ into $(R,\mathcal{B}(R))$. ($\mathcal{B}(R)$ denotes the Baire σ-field on R, the reals.)

The set of distributions $\{F(\cdot|\phi):\phi\varepsilon\Phi\}$ is called the set of possible distributions.

Remark 1.1. It is possible to parametrize any set of distributions. Suppose $F\subseteq P(\mathcal{B})$ is a set of probability distributions. One can set $\Phi = F$ and define $\mathcal{B}_\Phi$ to be the σ-field consisting of all subsets of Φ. This definition of $\mathcal{B}_\Phi$ guarantees that $F(\cdot|\cdot)$ is measurable.

Notation. If the family of distributions $\{F(\cdot|\phi):\phi\varepsilon\Phi\}$ is dominated by some σ-finite measure μ, then $f(\cdot|\phi)$ denotes a version of the density $dF/d\mu$, that is,

$$F(S|\phi) = \int_S f(s|\phi)d\mu(s).$$

Note that if $L_1(S,\mathcal{B}_S,\mu)$ is seperable then f may be chosen to be a measurable function from $(S\times\Phi,\mathcal{B}_S\times\mathcal{B}_\Phi)$ to $(R,\mathcal{B}(R))$.

Definition 1.4. The action space is a measurable space A with σ-field $\mathcal{B}_A$.

Definition 1.5. A decision procedure δ is a $\mathcal{B}_S$ measurable map from S to $P(\mathcal{B}_A)$. We shall denote the value of δ at $s\varepsilon S, B\varepsilon\mathcal{B}_A$ by $\delta(s,B)$. Note that δ measurable means

(1) For each $s\varepsilon S, \delta(s,\cdot)\varepsilon P(\mathcal{B}_A)$.

(2) For each $B\varepsilon\mathcal{B}_A$, $\delta(\cdot,B)$ is a measurable function from $(S,\mathcal{B}_S)$ into $(R,\mathcal{B}(R))$.

Notation. $\mathcal{D}$ denotes the set of all decision procedures.

Definition 1.6. The set of available decision procedures, denotes by $\mathcal{D}_o$, is a subset of $\mathcal{D}$.

Remark 1.2. Some examples of $\mathcal{D}_o$ are the class of invariant procedures, the class of monotone procedures etc.

Definition 1.7. The loss function L is a measurable function from $(\Phi x A, \mathcal{B}_\Phi x \mathcal{B}_A)$ to $([0,\infty], \mathcal{B}([0,\infty]))$.

Definition 1.8. The risk function of a procedure δ is the function $R(\cdot,\delta):\Phi\to[0,\infty]$ defined by

$$R(\phi,\delta) = \int\int L(\phi,a)\delta(x,da)F(dx|\phi).$$

Definition 1.9. Let $T = \{t:\Phi\to[0,\infty]\}$ have the weak (Tychonoff) topology defined by $t_\alpha\to t$ if and only if $t_\alpha(\phi)\to t(\phi)$ for all $\phi\varepsilon\Phi$. For $V\subseteq T$ let $\tilde{V} = \{t: t\varepsilon T, \exists t'\varepsilon V \ni t' \leq t\}$ where $t' \leq t$ means $t'(\phi) \leq t(\phi)$ for all $\phi\varepsilon\Phi$.

Note: T is compact Hausdorff.

Notation. For $\mathcal{D}_o \subseteq \mathcal{D}$ let $\Gamma(\mathcal{D}_o)$ denote the set of all risk functions corresponding to $\mathcal{D}_o$.

Definition 1.10. For any non-negative measure P on $(\Phi,\mathcal{B}_\Phi)$ and $\delta\varepsilon\mathcal{D}$, define the integrated risk $B(P,\delta)$ by

$$B(P,\delta) = \int R(\phi,\delta)P(d\phi).$$

Definition 1.11. For any non-negative measure P on $(\Phi, \mathcal{B}_\Phi)$, $\delta^* \varepsilon \mathcal{D}_o$ is said to be a Laplace procedure for P relative to $\mathcal{D}_o$ if and only if

$$B(P,\delta^*) = \inf_{\delta\varepsilon\mathcal{D}_o} B(P,\delta).$$

If P is a probability measure, that is, $P \varepsilon \mathcal{P}(\mathcal{B}_\phi)$, then δ^* is called a Bayes procedure.

Definition 1.12. Given $F(\cdot|\cdot)$ and $P(\cdot)$ as above define Π' to be the measure generated by F and P on the product space $(S x\Phi, \mathcal{B}_S x \mathcal{B}_\Phi)$. Thus Π' is the measure generated by the relation

$$\Pi'(S x\Lambda) = \int_\Lambda F(S|\phi)dP(\phi)$$

for $S\varepsilon\mathcal{B}_S$, $\Lambda\varepsilon\mathcal{B}_\phi$. Let Π denote the projection of Π' on $(S,\mathcal{B}_S)$; i.e. $\Pi(S) = \Pi'(S x\Phi)$.

Notation. If it exist we denote the $\mathcal{B}_S$ measurable conditional measure on $\mathcal{B}_\Phi$ given S relative to Π' by $P(\cdot|\cdot)$. That is, $P(\cdot|\cdot)$ is $\mathcal{B}_S$ measurable and $P(\cdot|\cdot)$ satisfies

$$\int_S P(\Lambda|s)\Pi(ds) = \Pi'(S x\Lambda) \quad \text{for all} \quad S\varepsilon\mathcal{B}_S,\ \Lambda\varepsilon\mathcal{B}_\phi.$$

If $P(\cdot)$ is a probability distribution then $P(\cdot|\cdot)$ is called the posterior distribution on Φ given $(S,\mathcal{B}_S)$.

Definition 1.13. When $P(\cdot|\cdot)$ exists define for $a\varepsilon A$, $s\varepsilon S$

$$B(a|s) = \int L(\phi,a)P(d\phi|s).$$

$B(a|s)$ may be described as the posterior risk incurred from taking action a.

Theorem 1.1. (Brown, 1974). Assume $P(\cdot|\cdot)$ exists. Let

$$(1.1) \qquad A^*(s) = \{a : a \varepsilon A, B(a|s) = \inf_{a \varepsilon A} B(a|s)\}.$$

Suppose $A^*(s)$ is non-empty a.e. Π and there exists a procedure $\delta^* \varepsilon \mathcal{D}$ such that $\delta^*(s, A^*(s)) = 1$ a.e. Π.
Then δ^* is a Laplace procedure for P. If $B(P) < \infty$ then any other Laplace procedure δ must also satisfy $\delta(s, A^*(s)) = 1$ a.e. Π.

Remark 1.3. If A is finite then $A^*(s)$ is non-empty a.e. Π.

Corollary 1.1. (Brown, 1974). Suppose the set $A^*(s)$ as defined in (1.1) consists of a single point of A a.e. Π. Suppose that $\mathcal{B}_A$ contains all single points and that there is a measurable function $d: S \to A$ such that $d(s) = A^*(s)$ a.e. Π. Then the non-randomized procedure δ^* defined by

$$\delta^*(s, \cdot) = \varepsilon_{d(s)}(\cdot)$$

where $\varepsilon_{d(s)}(\cdot)$ denotes the probability measure which gives probability 1 to the point $d(s)$ is a Laplace procedure for P.

Suppose in addition $F(\cdot|\phi)$ is absolutely continuous with respect to Π for every $\phi \varepsilon \Phi$ and there is a $\delta \varepsilon \mathcal{D}$ such that $B(P, \delta) < \infty$. Then δ^* defined above is the unique Laplace procedure.

The following theorem is well known.

Theorem 1.2. (Brown, 1974). If $\delta^* \varepsilon \mathcal{D}$ is the unique Laplace procedure for some P, then δ^* is admissable.

A wide variety of statistical results are based on the compactness of $\widetilde{\Gamma(\mathcal{D}_o)}$. We shall list three important ones.

Theorem 1.3. (Brown, 1974). Suppose $\widetilde{\Gamma(\mathcal{D}_o)}$ is compact in T. Then

(1) There exists a minimal complete class relative to $\mathcal{D}_o$.

(2) There exists an admissable minimax procedure relative to $\mathcal{D}_o$.

In order to state the next result, which is the only result that will be used in this thesis, it is necessary to describe what is meant by the limit of a net of decision procedures. This is most easily done when the family of distributions $\{F(\cdot|\phi):\phi\varepsilon\Phi\}$ is dominated by some σ-finite measure μ, and A has an appropriate topology on it for which A is compact and $\mathcal{B}_A = \mathcal{B}(A)$, the Baire σ-field on A. So under these assumptions we define convergence.

Definition 1.14. Under the assumptions stated above, a net $\{\delta_\alpha\}$ is said to converge to δ in the weak topology on $\mathcal{D}$ if and only if for every $f\varepsilon L_1(S,\mathcal{B}_S,\mu)$ and $\ell\varepsilon C(A)$ ($C(A)$ is the class of real-valued continuous function on A)

$$\iint f(s)\ell(a)\delta_\alpha(s,da)\mu(ds) \to \iint f(s)\ell(a)\delta(s,da)\mu(ds).$$

Definition 1.15. Any non-negative measure on (Φ,B_Φ) is called a prior. A measure P on (Φ,B_Φ) is called simple if P is a discrete measure concentrated on a finite set of Φ.

We now state the third result.

Theorem 1.4. (Brown, 1974). Suppose the family of distributions $\{F(\cdot|\phi):\phi\varepsilon\Phi\}$ is dominated by some σ-finite measure μ, A is compact second countable and $L(\phi,\cdot)$ is lower semi-continuous for each $\phi\varepsilon\Phi$. Let $\mathcal{D}_o$ be (weakly) compact and $\widetilde{\Gamma(\mathcal{D}_o)}$ closed and convex in T. Then the (weak) closure of the set of Bayes procedures for simple priors relative to the set $\mathcal{D}_o$ is an essentially complete class in $\mathcal{D}_o$.

The following theorem gives a sufficient condition for $\widetilde{\Gamma(\mathcal{D}_o)}$ to be compact.

Theorem 1.5. (Brown, 1974). Suppose the family of distributions $\{F(\cdot|\phi):\phi\varepsilon\Phi\}$ is dominated by some σ-finite measure μ, and A has an appropriate topology on it for which A is compact second countable Hausdorff, $\mathcal{B}_A = \mathcal{B}(A)$, and $L(\phi,\cdot)$ is lower semi-continuous for every $\phi\varepsilon\Phi$. Then $\mathcal{D}_o$ closed in the (weak) topology on $\mathcal{D}$ implies that $\widetilde{\Gamma(\mathcal{D}_o)}$ is compact and hence closed in T.

Remark 1.4. If the hypothesis of Theorem 1.5 is satisfied, then for Theorem 1.3 to apply one has to check that $\mathcal{D}_o$ is weakly closed. For Theorem 1.4 to apply, one checks in addition that $\widetilde{\Gamma(\mathcal{D}_o)}$ is convex.

Remark 1.5. Suppose the family of distributions $\{F(\cdot|\phi):\phi\varepsilon\Phi\}$ is dominated by some σ-finite measure μ and A is finite. Then by giving A the discrete topology the hypothesis of Theorem 1.5 is satisfied for any loss function L.

CHAPTER 2

DECISION-THEORETIC RESULTS FOR THE SUBSET SELECTION PROBLEM

This chapter deals with some decision-theoretic results for subset selection problems. In Sections 1 and 2 we show that essentially nothing is lost if we restrict our attention to Bayes procedures only. In particular in Section 1 it is shown that relative to the class of all subset selection procedures, Bayes procedures together with their limits form an essentially complete class. In Section 2 it is shown that relative to the class of permutationally invariant procedures, Bayes procedures for exchangeable priors, together with their limits, form an essentially complete class.

In Section 3 the concepts of Total Monotone Likelihood Ratio (TMLR) and Total Stochastic Monotone Property (TSMP) are introduced as multivariate generalizations of the concepts of (univariate) monotone likelihood ratio and (univariate) stochastic ordering. The related concept of Property M, first introduced in Eaton (1967), is also described. Implications of each of the concepts and relationships between the three are studied in detail. Examples of families of distributions having the various properties are also given. These concepts are used in all of the succeeding sections to obtain results concerning the form of Bayes procedures.

In Section 4 the form of Bayes procedures when the densities have property M and the loss function is monotone is investigated.

In Section 5 we describe a loss function which seems natural for the subset selection problem. For this loss function sufficient conditions for procedures to be Bayes are given. For the same loss function a sufficient condition for the uniqueness of Bayes procedures is given in Section 6.

In Section 7 the main theorem of this chapter is proved. It is shown that under certain conditions the class of monotone procedures forms an essentially complete class.

2.1. Decision-Theoretic Results for the General Subset Selection Problem

The sample space S is a measurable space with an associated σ-field B_S.

The parameter space Φ is a measurable space with an associated σ-field B_Φ.

We assume the set of possible distributions $\{F(\cdot|\phi): \phi\varepsilon\Phi\}$ is dominated by some σ-finite measure μ.

The action space A is the set of all non-empty subsets of $\{1,2,\ldots,p\}$ together with the power set of A as its associated σ-algebra B_A. The action $a \subseteq \{1,2,\ldots,p\}$ is to be interpreted as the action of selecting the populations $\{\Pi_i, i\varepsilon a\}$.

A subset selection procedure δ is a measurable function from (S, B_S) to $P(B_A)$. The class of all subset selection procedures is donated by $\mathcal{D}$.

The loss function L is a measurable function from $(\Phi \times A, \mathcal{B}_\Phi \times \mathcal{B}_A)$ to $([0,\infty], \mathcal{B}([0,\infty]))$.

By Remark 1.5 and Theorem 1.5 of Chapter 1, $\widetilde{\Gamma(\mathcal{D})}$ is compact. Now $\widetilde{\Gamma(\mathcal{D})}$ is always convex. Hence Theorem 1.3 and Theorem 1.4 apply. However, we shall only make use of Theorem 1.4 which is restated as

Theorem 2.1.1. Relative to $\mathcal{D}$, the (weak) closure in the topology on $\mathcal{D}$ of the procedures that are Bayes relative to $\mathcal{D}$ forms an essentially complete class.

We shall assume, throughout the thesis, that the parameter space Φ is a subset of the Euclidean space R^{p+r}, and that for $\phi \varepsilon \Phi$, the first p components of ϕ are the parameters of interest, and the last r components of ϕ are nuisance parameters. When we write $(\theta,\psi) \varepsilon \Phi$, θ will always be the p-dimensional vector of parameters of interest and ψ will always be the r-dimensional vector of nuisance parameters. The projection of Φ onto the first p coordinates will be denoted by Θ and the projection of Φ onto the last r coordinates will be denoted by Ψ. Note that $\Theta \times \Psi$ does not necessarily equal Φ. We shall assume that $\mathcal{B}_\Phi$ is the σ-field inherited from the Borel σ-field on R^{p+r}.

We shall assume, through the entire thesis, that the sample space S is a subset of the Euclidean space R^{p+q}. When we write $(x,y) \varepsilon S$, x shall always be a p-dimensional vector and y shall always be a q-dimensional vector. Roughly speaking, x will be the part of the observation that gives information concerning the relative ordering of the θ's while y will be the remaining part of the observation. The projection of S onto the first p coordinates will be denoted by X and the projection of S onto the last q coordinates will be denoted by Y. Again, $X \times Y$ need not equal S. We shall assume that $\mathcal{B}_S$ is the σ-field inherited from the Borel σ-field on R^{p+q}.

To fix ideas, consider the following example. Suppose n observations are taken from each of the p independent normal populations with unknown means and unknown, possibly unequal, variances. Say the observations are $Z_{i\alpha}$, $\alpha = 1, \ldots, n$, $i = 1, \ldots, p$, $\{Z_{i\alpha}, \alpha = 1, \ldots, n\}$ iid $N(\mu_i, \sigma_i^2)$. By sufficiency we can reduce the observations to $(\bar{Z}_1, \ldots, \bar{Z}_p, s_1^2, \ldots, s_p^2)$. Suppose we want to select in terms of the means. Then $\theta = (\mu_1, \ldots, \mu_p)$, $\psi = (\sigma_1^2, \ldots, \sigma_p^2)$, $x = (\bar{Z}_1, \ldots, \bar{Z}_p)$ and $y = (s_1^2, \ldots, s_p^2)$.

2.2. Decision-Theoretic Results for the (Permutationally) Invariant Subset Selection Problem

Notation. Let S_p be the group of permutations on $\{1, 2, \ldots, p\}$. (The symmetric group of order p). The element of S_p which interchanges i and j, leaving all other members of $\{1, 2, \ldots, p\}$ fixed, is denoted by (i, j). For $(x, y) \varepsilon R^{p+q}$ and $\pi \varepsilon S_p$, define $\pi(x, y)$ by $\pi(x, y) = (\pi x, y)$ where πx is defined by $(\pi x)_i = x_{\pi^{-1} i}$. Similarly, for $(\theta, \psi) \varepsilon R^{p+r}$ and $\pi \varepsilon S_p$, $\pi(\theta, \psi)$ is defined by $\pi(\theta, \psi) = (\pi\theta, \psi)$ where $(\pi\theta)_i = \theta_{\pi^{-1} i}$.

For any set $S \subseteq R^{p+q}$, πS will denote the image of S under π. Similarly for any set $\Lambda \subseteq R^{p+r}$, $\pi\Lambda$ denotes the image of Λ under π.

We assume that the sample space S is a Borel subset of R^{p+r} invariant under S_p, that is, for any $\pi \varepsilon S_p$, $\pi S = S$. We assume that B_S is the σ-field inherited from the Borel σ-field on R^{p+q}.

We assume the parameter space Φ is a Borel subset of R^{p+r} invariant under S_p, that is, for any $\pi \varepsilon S_p$, $\pi\Phi = \Phi$. We assume B_Φ is the σ-field inherited from the Borel σ-field on R^{p+r}.

It is assumed that the family of possible distributions $\{F(\cdot,\cdot|\theta,\psi):(\theta,\psi)\varepsilon\Phi\}$ is dominated by some σ-finite measure μ. We further assume the following invariance properties for the densities $\{f(\cdot,\cdot|\theta,\psi):(\theta,\psi)\varepsilon\Phi\}$ and the measure μ:

$$f(\pi x,y|\pi\theta,\psi) = f(x,y|\theta,\psi),$$

$$d\mu(\pi x,y) = d\mu(x,y).$$

The space of possible actions A is the set of all non-empty subsets of $\{1,2,\ldots p\}$ together with the power set of A as the associated σ-algebra B_A.

A decision procedure δ is a measurable function from (S,B_S) to $P(B_A)$. The class of all decision procedures is denoted by $\mathcal{D}$. For $\delta\varepsilon\mathcal{D}$ and $\pi\varepsilon S_p$, define $\pi\delta$ by $\pi\delta(\cdot,a) = \delta(\cdot,\pi^{-1}a)$. A procedure δ is said to be (permutationally) invariant if and only if $\delta(\pi s,\pi a) = \delta(s,a) \; \forall \, s\varepsilon S$, $\pi\varepsilon S_p$, $a\varepsilon A$. Denote the class of invariant procedures by $\mathcal{D}_I$.

The loss function L is a measurable function from $(S \times A, \mathcal{B}_S \times \mathcal{B}_A)$ to $([0,\infty], \mathcal{B}([0,\infty]))$. We assume L satisfies the following invariance assumption:

$$L(\pi\phi, \pi a) = L(\phi, a) \ \forall\ \phi, a \text{ and } \pi.$$

Definition 2.2.1. A non-negative measure P on $(\Phi, \mathcal{B}_\Phi)$ is said to be exchangeable if and only if for any $\pi \varepsilon S_p$, $\Lambda \varepsilon \mathcal{B}_\Phi$, $P(\pi\Lambda) = P(\Lambda)$.

Clearly in the above setup the decision problem is invariant under the group S_p. The following Hunt-Stein type theorem gives support for considering only procedures in $\mathcal{D}_I$ if the prime consideration is the supremum of the risk.

Theorem 2.2.1. Given any $\delta \varepsilon \mathcal{D}$, $\exists\ \delta_I \varepsilon \mathcal{D}_I \ni \sup_{\phi\varepsilon\Phi} R(\phi, \delta_I) \le \sup_{\phi\varepsilon\Phi} R(\phi, \delta)$.

Proof. Define δ_I by $\delta_I(s,a) = (1/p!) \sum_{\pi\varepsilon S_p} \delta(\pi s, \pi a)$. Then δ_I is invariant. But $\sup_{\phi\varepsilon\Phi} R(\phi, \delta_I) = \sup_{\phi\varepsilon\Phi} (1/p!) \sum_{\pi\varepsilon S_p} R(\pi\phi, \delta) \le (1/p!) \sum_{\pi\varepsilon S_p} \sup_{\phi\varepsilon\Phi} R(\pi\phi, \delta) = (1/p!) \sum_{\pi\varepsilon S_p} \sup_{\phi\varepsilon\Phi} R(\phi, \delta) = \sup_{\phi\varepsilon\Phi} R(\phi, \delta)$.

If one does restrict attention to (permutationally) invariant procedures only, then the following can be proved.

Lemma 2.2.1. $\mathcal{D}_I$ is closed in $\mathcal{D}$ in the weak togology on $\mathcal{D}$.

Proof. Suppose a net $\{\delta_\alpha\}$ in $\mathcal{D}_I$ converges to δ. For fixed $a \varepsilon A$ and $\pi \varepsilon S_p$, let $A_i = \{s : \delta(\pi s, \pi a) - \delta(s,a) > 1/i\}$. Suppose $\mu(A_i) > 0$. Then there exists $B_i \subseteq A_i$ such that $0 < \mu(B_i) < \infty$. Now I_{B_i} and $I_{\pi B_i} \varepsilon L_1(S, \mathcal{B}_S, \mu)$. Both $I_{\{a\}}$ and $I_{\{\pi a\}} \varepsilon C(A)$. Hence $\int_{\pi B_i} \delta_\alpha(s, \pi a) d\mu(s) \to \int_{\pi B_i} \delta(s, \pi a) d\mu(s)$ and $\int_{B_i} \delta_\alpha(s,a) d\mu(s)$

$\int_{B_i} \delta(s,a)d\mu(s)$. But $\int_{\pi B_i} \delta_\alpha(s,\pi a)d\mu(s) = \int_{B_i} \delta_\alpha(s,a)d\mu(s)$ for all α. Therefore $\int_{B_i} [\delta(\pi s,\pi a)-\delta(s,a)]d\mu(s) = 0$. But the left side is $\geq (1/i)\mu(B_i)$. Hence $\mu(B_i) = 0$. Contradiction. So $\mu(A_i) = 0$. $\mu\{s:\delta(\pi s,\pi a)-\delta(s,a)>0\} = \lim_{i\to\infty} \mu(A_i)=0$. Similarly $\mu\{s:\delta(\pi s,\pi a)-\delta(s,a)<0\} = 0$. Therefore $\mu\{s:\delta(\pi s,\pi a) \neq \delta(s,a)\} = 0$. a and π are arbitrary. Hence $\delta\varepsilon\mathcal{D}_I$.

Now if $\delta_1, \delta_2 \,\varepsilon\mathcal{D}_I$, then $\alpha\delta_1 + (1-\alpha)\delta_2\varepsilon\mathcal{D}_I$ for any $\alpha\varepsilon[0,1]$. Hence $\widetilde{\Gamma(\mathcal{D}_I)}$ is convex. Combining this and the previous lemma we see that Theorem 1.3 and Theorem 1.4 apply. However, we shall only make use of Theorem 1.4 in this thesis which we restate as:

Theorem 2.2.2. Relative to the class of (permutationally) invariant procedures $\mathcal{D}_I$, the (weak) closure in the topology on $\mathcal{D}_I$ of the Bayes procedures relative to $\mathcal{D}_I$ forms an essentially complete class.

Given a prior it is often easier to find its Bayes procedure(s) relative to $\mathcal{D}$ than to find its Bayes procedure(s) relative to $\mathcal{D}_I$. The following theorem gives the needed connection.

Theorem 2.2.3. The class of procedures that are Bayes relative to $\mathcal{D}_I$ is contained in the class of procedures that are Bayes relative to $\mathcal{D}$ for exchangeable priors.

Proof. Suppose $\delta\varepsilon\mathcal{D}_I$ is Bayes relative to $\mathcal{D}_I$ for some prior probability measure P on $(\Phi, \mathcal{B}_\Phi)$. Then it is easy to see that δ is Bayes relative to $\mathcal{D}_I$ for the prior P_o defined by $P_o(\Lambda) = (1/p!) \sum_{\pi\varepsilon S_p} P(\pi\Lambda)$. For P_o relatively to $\mathcal{D}$ a

Bayes procedure exists hence an invariant Bayes procedure exist. Call it δ'. But $B(\delta,P_o) \leq B(\delta',P_o)$. Hence δ is Bayes relative to $\mathcal{D}$ for P_o.

2.3. Some Orderings on Families of Distributions

In univariate statistical inference, the concept of monotone likelihood ratio plays a central role. Therefore it is reasonable to think that the concept of multivariate monotone likelihood ratio should be important in multivariate statistical inference. Unfortunately there has never been a unified theory of multivariate monotone likelihood ratio. In studying different problems different definitions of multivariate monotone likelihood ratio were proposed. In Pratt (1956), a definition of monotone likelihood ratio on contours was given. Karlin and Truax (1960), and later Hall and Kudo (1968), used essentially the same definition in studying slippage tests. In studying the complete ranking problem, Bahadur and Goodman (1952) and Lehmann (1966) made certain independence and permutational invariance assumptions and used univariate MLR. It was later found by Eaton (1969) that for the complete ranking problem a weaker condition which he called Property M suffices. As it turns out, none of the above concepts is really adequate for the problem at hand, namely the subset selection problem. We have therefore chosen to give our own definition of multivariate monotone likelihood ratio and also the corresponding definition of multivariate stochastic ordering. It is clear that these new concepts will be useful in studying other multiple comparison problems. However, it is not yet clear whether they will be useful in other, more general, multivariate inference problems.

Definition 2.3.1. For any fixed $B \subseteq K = \{1,2,\ldots,p\}$ (B may be the empty set or the whole set), define a partial ordering 'less than or equal to in B' ($\overset{B}{\leq}$) on X as follows: For $x = (x_1, x_2, \ldots, x_p)$, $x' = (x'_1, x'_2, \ldots, x'_p)$, $x \overset{B}{\leq} x'$ if and only if

$$x_i \begin{smallmatrix} \leq \\ \geq \end{smallmatrix} x_i' \quad \text{as } i \begin{smallmatrix} \in \\ \notin \end{smallmatrix} B.$$

In words, $x \overset{B}{\leq} x'$ if and only if x is less than or equal to x' in those coordinates that are in B and greater than or equal to x' in those coordinates that are not in B.

The above definition induces a partial ordering on S as follows:

Definition 2.3.2. For (x,y), $(x',y') \in S$, $(x,y) \overset{B}{\leq} (x',y')$ if and only if $x \overset{B}{\leq} x'$ and $y = y'$.

Definition 2.3.3. A set $A \subseteq S$ is said to be nondecreasing in B ($\nearrow$B) if and only if $s \in A$, $s \overset{B}{\leq} s' \Rightarrow s' \in A$.

Definition 2.3.4. A function $h: S \to R$ is said to be nondecreasing in B($\nearrow$B) if and only if $s \overset{B}{\leq} s' \Rightarrow h(s) \leq h(s')$. Thus a set is nondecreasing in B if and if its indicator function as a function is nondecreasing in B.

In exactly the same way we define $\overset{B}{\leq}$ and $\nearrow$B on Θ, Φ and for functions from Φ to R.

Definition 2.3.5. The distribution $F(\cdot|\phi_1)$ is said to be stochastically smaller in B than $F(\cdot|\phi_2)$ if and only if for any measurable set $A \subseteq S$ that is nondecreasing in B, $F(A|\phi_1) \leq F(A|\phi_2)$.

Definition 2.3.6. The family of distributions $\{F(\cdot|\phi): \phi\varepsilon\Phi\}$ is said to be stochastically nondecreasing in B(SM↗B) if and only if for any measurable set $A \subseteq S$ that is nondecreasing in B, $F(A|\phi)$ as a function of ϕ is nondecreasing in B.

Remark 2.3.1. Lehmann (1955) in studying ordered families of distributions defined and investigated families of distributions that are stochastically nondecreasing in $K = \{1,2,\ldots,p\}$.

Definition 2.3.7. The family of distributions $\{F(\cdot|\phi): \phi\varepsilon\Phi\}$ is said to have Total Stochastic Monotone Property (TSMP) if and only if it is stochastically nondecreasing in B for every $B \subseteq K = \{1,2,\ldots,p\}$.

Definition 2.3.8. The family of densities $\{f(\cdot|\phi): \phi\varepsilon\Phi\}$ is said to have nondecreasing likelihood ratio in B (MLR↗B) if and only if

$$s \overset{B}{\leq} s',\ \phi \overset{B}{\leq} \phi' \Rightarrow f(s'|\phi)f(s|\phi') \leq f(s|\phi)f(s'|\phi').$$

Remark 2.3.2. In investigating multivariate one-sided tests, Oosterhoff (1969) defined and used nondecreasing likelihood ratio in $K = \{1,2,\ldots,p\}$.

Definition 2.3.9. The family of densities $\{f(\cdot|\phi): \phi\varepsilon\Phi\}$ is said to have Total Monotone Likelihood Ratio (T) if and only if it has nondecreasing likelihood ratio in B for every $B \subseteq K = \{1,2,\ldots,p\}$.

Definition 2.3.10. (Eaton, 1967). Under the symmetric setup of Section 2.2, the family of densities $\{f(\cdot|\phi):\phi\varepsilon\Phi\}$ is said to have Property M if and only if for $(x,y)\varepsilon S$, (θ,ψ) $\varepsilon\Phi$, $x = (x_1,x_2,\ldots,x_p)$, $\theta = (\theta_1,\theta_2,\ldots,\theta_p)$,

$$x_i \le x_j,\ \theta_i \le \theta_j \Rightarrow f(x,y|\theta,\psi) \ge f(x,y|(i,j)\theta,\psi)$$

In exactly the same way we define TSMP, TMLR and, in the case of exchangeable prior, Property M for the posterior distributions and densities $\{F(\cdot|s):s\varepsilon S\}$ and $\{f(\cdot|s):s\varepsilon S\}$.

Definition 2.3.11. For any subset selection procedure δ let $\delta_i(s) = \sum_{i\varepsilon a} \delta(s,a)$, i.e. $\delta_i(s)$ is the probability of selecting i having observed s. A subset selection procedure δ is said to be monotone if and only if for each i $\delta_i(s)$ is essentially $\nearrow_{\{i\}}$, that is, there do not exist S_1, $S_2 \subseteq S$, $\mu(S_1)$, $\mu(S_2) > 0$, $s_1 \le s_2$ for all $s_1\varepsilon S_1$, $s_2\varepsilon S_2$ such that $\text{ess}\sup_{s\varepsilon S_1} \delta_i(s) > \text{ess}\inf_{s\varepsilon S_2} \delta_i(s)$.

Remark 2.3.3. What we call 'monotone' procedures traditionally have been called 'just' procedures in the literature. See Nagel (1970) and Gupta and Nagel (1971). Following Gupta and Huang (1976), we have changed the terminology to 'monotone' since we have in mind the analog of the classical univariate result of Karlin and Rubin (1956) on the class of monotone procedures.

Theorem 2.3.1. TMLR $\Rightarrow$ TSMP.

Proof. Suppose $\{f(\cdot|\phi):\phi\varepsilon\Phi\}$ has TMLR.

Let $B \subseteq K = \{1,2,\ldots,p\}$ be fixed. For $\phi \overset{B}{\leq} \phi'$ let S^- and S^+ be the sets in S for which $f(s|\phi') < f(s|\phi)$ and $f(s|\phi') > f(s|\phi)$ respectively.

Suppose h, a measurable function from $(S,\mathcal{B}_S)$ to $(R,\mathcal{B}(R))$, is nondecreasing in B. Let $a = \sup_{S^-} h(s)$ and $b = \inf_{S^+} h(s)$. Then $b-a \geq 0$ by MLR↗B and

$$E(h|\phi') - E(h|\phi)$$

$$\geq a[F(S^-|\phi') - F(S^-|\phi)] + b[F(S^+|\phi') - F(S^+|\phi)]$$

$$= (b-a)\ [F(S^+|\phi') - F(S^+|\phi)] \geq 0.$$

This is true for all B. Hence TMLR ⇒ TSMP.

Theorem 2.3.2. Under the symmetric setup of Section 2.2, TMLR ⇒ Property M.

Proof. Suppose $\{f(\cdot,\cdot|\theta,\psi):(\theta,\psi)\varepsilon\Phi\}$ has TMLR. Suppose $(x,y)\varepsilon S$, $(\theta,\psi)\varepsilon\Phi$, $x = (x_1,x_2,\ldots,x_p)$, $\theta = (\theta_1,\theta_2,\ldots,\theta_p)$. Clearly $x_i \leq x_j$, $\theta_i \leq \theta_j \Rightarrow x \overset{\{i\}}{\leq} (i,j)x$, $\theta \overset{\{i\}}{\leq} (i,j)\theta$. By MLR↗{i} we have $f(x,y|\phi,\psi)f((i,j)x,y|(i,j)\theta,\psi) \geq f((i,j)x,y|\theta,\psi)f(x,y|(i,j)\theta,\psi)$ or

$$f^2(x,y|\theta,\psi) \geq f^2(x,y|(i,j)\theta,\psi).$$

This completes the proof.

Examples of Families of Densities Having TMLR

We first make the easy observation that if $f(x|\theta) = \prod_{i=1}^{p} f_i(x_i|\theta_i)$ and $X = \prod_{i=1}^{p} X_i$, $\Theta = \prod_{i=1}^{p} \Theta_i$, then TMLR is equivalent to univariate MLR, that is, $\{f(\cdot|\theta):\theta\varepsilon\Theta\}$ has TMLR if and only if for each i, $\{f_i(\cdot|\theta_i):\theta_i\varepsilon\Theta_i\}$ has univariate MLR. In addition, we have the following

Theorem 2.3.3. Any family of distributions whose densities are of the form

$$(2.3.1) \qquad C(\theta,\psi)\ \exp\left[\sum_{i=1}^{p} Q_i(\theta_i,\psi)x_i\right] g(\psi,y)h(x,y)$$

where each Q_i for fixed ψ is nondecreasing in θ_i has TMLR.

Proof. Suppose (x,y), $(x',y)\varepsilon S$; (θ,ψ), $(\theta',\psi)\varepsilon\Phi$; $x \overset{B}{\leq} x'$ and $\theta \overset{B}{\leq} \theta'$. We need to show that

$$\sum_{i=1}^{p} Q_i(\theta_i,\psi)x_i + \sum_{i=1}^{p} Q_i(\theta_i',\ \psi)x_i' \geq \sum_{i=1}^{p} Q_i(\theta_i,\psi)x_i' + \sum_{i=1}^{p} Q_i(\theta_i',\ \psi)x_i$$

or equivalently

$$\sum_{i=1}^{p} [Q_i(\theta_i',\ \psi)-Q_i(\theta_i,\ \psi)]\ (x_i' - x_i) \geq 0$$

But this follows from $x \overset{B}{\leq} x'$ and $\theta \overset{B}{\leq} \theta'$. B is arbitrary. Hence the densities have TMLR.

Example 1. The multinomial density

$$\binom{n}{x_1 \ldots x_p}\theta_i^{x_1}\ldots\theta_p^{x_p} = \binom{n}{x_1\ldots x_p}\ e^{\sum_{i=1}^{p} x_i \ln\theta_i}$$

is in the form of (2.3.1.).

Example 2. The Direchlet density

$$\frac{\Gamma\left(\sum_{i=1}^{p} x_i\right)}{\Gamma(x_i)\ldots\Gamma(x_p)}\ \theta_1^{x_1-1}\ \ldots\theta_p^{x_p-1}$$

is in the form of (2.3.1.).

Example 3. Consider the case of taking n observations each from p independent normal populations with unknown means and unknown but equal variances. The sufficient statistic in this case is $\bar{x}_1,\ldots,\bar{x}_p,S^2$, the sample means and the pooled estimate of the common variance. The joint density

$$C(\underset{\sim}{\mu},\sigma^2)e^{\sum_{i=1}^{p}\frac{\mu_i}{\sigma^2/n}\bar{x}_i}e^{\frac{-p(n-1)}{2\sigma^2}s^2}h(\bar{x}_1,\bar{x}_2,\ldots,\bar{x}_p,s^2)$$

is in the form of (2.3.1.).

Examples of Families of Distributions Having TSMP

We first note that if $F(x|\theta) = \prod_{i=1}^{p} F(x_i|\theta_i)$ then TSMP reduces to univariate stochastic ordering, that is, $\{F(\cdot|\theta):\theta\varepsilon\Theta\}$ has TSMP if and only if for each i, $\{F_i(\cdot|\theta_i):\theta_i\varepsilon\Theta_i\}$ is stochastically ordered where Θ_i is the projection of Θ onto the ith coordinate. In addition, we have

Theorem 2.3.4. Suppose $\mathcal{X} = \Theta = R^p$ and $F(x,y|\theta,\psi) = F_o(x-\theta,y|\psi)$, that is, $\mathcal{F} = \{F(\cdot,\cdot|\theta,\psi):(\theta,\psi)\varepsilon\Phi\}$ is a location family of distributions, then $\mathcal{F}$ has TSMP.

Proof. Suppose $A\varepsilon\underset{B}{\mathcal{B}_{\mathcal{X}}}$ is nondecreasing in B, (θ,ψ), $(\theta',\psi)\varepsilon\Phi$ and $\theta \leq \theta'$. Then $A_\theta = \{(x,y):(x+\theta,y)\varepsilon A\} \subseteq A_{\theta'} = \{(x,y):(x+\theta',y)\varepsilon A\}$ and hence $F(A|\theta,\psi) = F_o(A_\theta|\psi) \leq F_o(A_{\theta'}|\psi) = F(A|\theta',\psi)$.

Examples of Families of Densities Having Property M

It is easy to see that under the symmetric setup of Section 2.2, if $f(x|\theta) = \prod_{i=1}^{p} f(x_i|\theta_i)$ then Property M is equivalent to univariate MLR. In addition the following theorem concerning elliptically contoured families of distributions was proved in Eaton (1967):

Theorem 2.3.5. (Eaton, 1967). Suppose $\{f(\cdot|\theta):\theta\varepsilon R^p\}$ are densities with respect to the Lebesque measure on R^p.

Suppose further

$$f(x|\theta) = C(\Lambda)g[(x-\theta)\ \Lambda(x-\theta)']$$

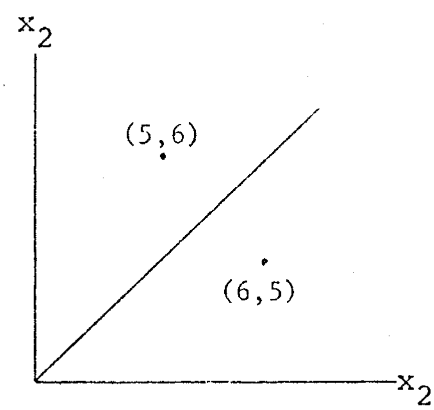
where Λ is a pxp positive definite matrix, g is strictly decreasing, and $C(\Lambda)$ is a positive constant. Then the following are equivalent:

(i) $\{f(\cdot|\theta):\theta\varepsilon R^p\}$ has Property M;

(ii) $\Lambda = C_1 I - C_2\, \underset{\sim}{1}'\underset{\sim}{1}$, $C_1 > 0$, $-\infty < C_2/C_1 < 1/p$.

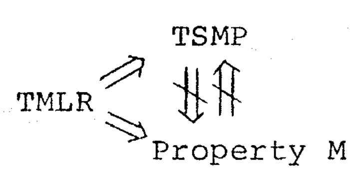
Clearly in the symmetric setup there are families of distributions having TSMP but not Property M since univariate stochastic ordering does not imply univariate MLR. On the other hand Property M does not imply TSMP either. The following example shows this.

θ \ x	(5,6)	(6,5)
(1,2)	0.9	0.1
(2,1)	0.1	0.9
(3,4)	0.6	0.4
(4,3)	0.4	0.6



This family of four distributions clearly has Property M. But it does not have TSMP. For instance, $P\{(5,6)|(1,2)\} = 0.9 > P\{(5,6)|(3,4)\} = 0.6$.

The following diagram summarizes the relationship between TMLR, TSMP and Property M.



The following is a generalization of a result in Lehmann (1955).

Theorem 2.3.6. Suppose $F(\cdot|\phi)$ is stochastically smaller than $F(\cdot|\phi')$ in B. Then for any measurable function h that is nondecreasing in B, if $E_\phi[h(X,Y)]$ and $E_{\phi'}[h(X,Y)]$ exist, then $E_\phi[h(X,Y)] \leq E_{\phi'}[h(X,Y)]$.

Proof. Let h^+ and h^- be the positive and negative parts of h respectively. We shall approximate h^+ by a sequence of simple functions. Let

$$h_n(x,y) = \begin{cases} (i-1)/2^n & \text{for } (x,y)\varepsilon S_i^{(n)} \\ n & \text{for } (x,y)\varepsilon S_N \end{cases}$$

where

$$S_i^{(n)} = \{(x,y):(i-1)/2^n \leq h^+(x,y) < i/2^n\},$$
$$i = 1,2,\ldots,n2^n,$$
$$S_N^{(n)} = \{(x,y):h^+(x,y) \geq n\}, \quad N = n2^n + 1.$$

Then $h_n = \sum_{i=2}^{N} 1/2^n \left(I_{S_i^{(n)}} + I_{S_{i+1}^{(n)}} + \ldots + I_{S_N^{(n)}}\right) =$

$$\sum_{i=2}^{N} 1/2^n \; I_{\bigcup_{j=i}^{N} S_j^{(n)}}$$

and $h_n \to h^+$. Now for each i, $\bigcup_{j=i}^{N} S_j^{(n)}$ is nondecreasing in B since h^+ is nondecreasing in B. Hence $E_\phi[h_n(X,Y)] \leq E_{\phi'}[h_n(X,Y)]$.

Using the Monotone Convergence Theorem, we have

$$E_\phi[h^+(X,Y)] = \lim_{n\to\infty} E_\phi[h_n(X,Y)] \leq \lim_{n\to\infty} E_{\phi'}[h_n(X,Y)] =$$
$$E_{\phi'}[h^+(X,Y)].$$

Similarly we can prove that $E_\phi[h^-(X,Y)] \leq E_{\phi'}[h^-(X,Y)]$. Thus if $E_\phi[h(X,Y)]$ and $E_{\phi'}[h(X,Y)]$ exist then $E_\phi[h(X,Y)] \leq E_{\phi'}[h(X,Y)]$.

The result that we will actually use is contained in the following corollary.

Corollary 2.3.1. Suppose the family of distributions $\{F(\cdot|\phi):\phi\varepsilon\Phi\}$ has TSMP. If $h:S \to [0,\infty]$ is measurable and nondecreasing in B, then $E_\phi[h(X,Y)]$ as a function of ϕ is nondecreasing in B.

Even though Property M does not imply TSMP, it does imply a sort of one-dimensional stochastic ordering which we shall prove through a series of lemmas. So for the remainder of this section, we assume the family of densities $\{f(\cdot,\cdot|\theta,\psi):(\theta,\psi)\varepsilon\Phi\}$ has Property M.

Lemma 2.3.1. For fixed $i,j(1\leq i,j\leq p)$ let $C_i^j = \{(x,y):(x,y)\varepsilon S,\ x_i \leq x_j\}$. Then for any set $C \subseteq C_i^j$ measurable with respect to $\mathcal{B}_S$ and any $(\theta,\psi)\varepsilon\Phi$,

$$F(C|\theta,\psi) \gtreqless F((i,j)C|\theta,\psi) \text{ as } \theta_i \lesseqgtr \theta_j.$$

Proof. Suppose $\theta_i \leq \theta_j$. Then

$$\begin{aligned} F(C|\theta,\psi) &= \int_C f(x,y|\theta,\psi)\,d\mu(x,y) \\ &\leq \int_C f(x,y|(i,j)\theta,\psi)\,d\mu(x,y) \\ &= \int_{(i,j)C} f(x,y|\theta,\psi)\,d\mu(x,y) \\ &= F((i,j)C|\theta,\psi). \end{aligned}$$

Similarly for $\theta_i \geq \theta_j$.

Lemma 2.3.2. Suppose $C \subseteq S$ is measurable with respect to $\mathcal{B}_S$ and nondecreasing in $\{i\}$. Then

$$F(C|\theta,\psi) \lesseqgtr F((i,j)C|\theta,\psi) \text{ as } \theta_i \lesseqgtr \theta_j.$$

Proof. Let $C_o = \{(x,y:(x,y)\varepsilon C,((i,j)x,y)\varepsilon C\}$, $C_i = \{(x,y):(x,y)\varepsilon C,((i,j)x,y)\not\varepsilon C\}$ and $C_j = \{(x,y):(x,y)\not\varepsilon C, ((i,j)x,y)\varepsilon C\} = (i,j)C_i$. Then

$$F(C|\theta,\psi) = F(C_o|\theta,\psi) + F(C_i|\theta,\psi),$$

$$F((i,j)C|\theta,\psi) = F(C_o|\theta,\psi) + F(C_j|\theta,\psi).$$

But $(x,y)\varepsilon C_j \Rightarrow x_i < x_j$ for $x_i \geq x_j$ and $((i,j)x,y)\varepsilon C$ would imply $(x,y)\varepsilon C$ since C is nondecreasing in $\{i\}$. Hence $C_j \subseteq C_i^j$. Therefore by the previous lemma

$$F(C|\theta,\psi) \lesseqgtr F((i,j)C|\theta,\psi) \text{ as } \theta_i \lesseqgtr \theta_j.$$

Theorem 2.3.7. Suppose $h: \mathcal{S} \to [0,\infty]$ measurable with respect to $\mathcal{B}_\mathcal{S}$ is nondecreasing in $\{i\}$. Then $E_{\theta,\psi}[h(X,Y)] \leq E_{\theta,\psi}[h((i,j)X,Y)]$ if $\theta_i \leq \theta_j$.

Proof. $\underset{\{i\}}{E}_{\theta,\psi}[h((i,j)X,Y)] = E_{(i,j)\theta,\psi}[h(X,Y)]$. But $\theta_i \leq \theta_j \Rightarrow \theta \leq (i,j)\theta$. Hence by Corollary 2.3.1. $E_{\theta,\psi}[h(X,Y)] \leq E_{(i,j)\theta,\psi}[h(X,Y)] = E_{\theta,\psi}[h((i,j)X,Y)]$.

2.4. Form of Bayes Procedures Under The Symmetric Setup When The Densities Have Property M and the Loss Function is Monotone

The results of this section were essentially obtained in Goel and Rubin (1975). We assume the setup of the problem is the symmetric setup of Section 2.2. To refresh our memory, we review briefly the setup.

The sample $\mathcal{S}$ is a Borel subset of R^{p+q} invariant under S_p, the symmetric group of order p. The associated σ-field $\mathcal{B}_\mathcal{S}$ is the σ-field inherited from the Borel σ-field on R^{p+q}.

The parameter space Φ is a Borel subset of R^{p+r} invariant under S_p. The associated σ-field $\mathcal{B}_\Phi$ is the σ-field inherited from the Borel σ-field on R^{p+r}.

The family of possible distributions $\{F(\cdot,\cdot|\theta,\psi): (\theta,\psi)\varepsilon\Phi\}$ is dominated by some σ-finite measure μ. The densities $\{f(\cdot,\cdot|\theta,\psi):(\theta,\psi)\varepsilon\Phi\}$ and the dominating measure μ satisfy the following invariance assumptions: For any $\pi\varepsilon S_p$

$$f(\pi x,y|\pi\theta,\psi) = f(x,y|\theta,\psi)$$

$$d\mu(\pi x,y) = d\mu(x,y).$$

The action space $\mathcal{A}$ is the set of all nonempty subsets of $K = \{1,2,\ldots,p\}$ together with the power set of $\mathcal{A}$ as the associated σ-algebra $\mathcal{B}_A$.

A decision procedure is a measurable function from $(\mathcal{S},\mathcal{B}_S)$ to $\mathcal{P}(\mathcal{B}_A)$. The class of all decision procedures is denoted by $\mathcal{D}$. A procedure $\delta\varepsilon\mathcal{D}$ is said to be (permutationally) invariant if and only if $\delta(\pi s,\pi a) = \delta(s,a)\forall s,a$. The class of all invariant procedures in $\mathcal{D}$ is denoted by $\mathcal{D}_I$.

The loss function L is a measurable function from $(\mathcal{S}x\mathcal{A}, \mathcal{B}_S x\mathcal{B}_A)$ to $([0,\infty],\mathcal{B}([0,\infty]))$ satisfying the invariance condition

$$L(\pi\phi,\pi a) = L(\phi,a).$$

All the assumptions made in the above setup are invariance assumptions. We now make in addition two ordering assumptions.

We assume that the densities $\{f(x,y|\theta,\psi):(\theta,\psi)\varepsilon\Phi\}$ have Property M as defined in the previous section. That is, for $(x,y)\varepsilon S$, $(\theta,\psi)\varepsilon\Phi$, $x = (x_1,x_2,\ldots,x_p)$, $\theta = (\theta_1,\theta_2,\ldots,\theta_p)$,

$$x_i \geq x_j,\ \theta_i \geq \theta_j \Rightarrow f(x,y|\theta,\psi) \geq f(x,y|(i,j)\theta,\psi).$$

We assume the loss function L in addition to being invariant has the following monotonicity property:

For $(\theta,\psi)\varepsilon\Phi$, $\theta = (\theta_1,\theta_2,\ldots,\theta_p)$, $\theta_i \geq \theta_j$, $i\varepsilon a$, $j\not\varepsilon a$ imply

$$L((\theta,\psi),a) \leq L((\theta,\psi),(i,j)a).$$

For each $a\varepsilon A$ let $B_a = \{(x,y):(x,y)\varepsilon S,\ x_i \geq x_j \text{ for all } i\varepsilon a,\ j\not\varepsilon a\}$.

For each P, (x,y) and a consider $r_a(x,y)$ defined by

$$r_a(x,y) = \int_\Phi L((\theta,\psi),a)f(x,y|\theta,\psi)dP(\theta,\psi).$$

Note that $r_a(x,y)$ is proportional to the posterior risk $B(a|x,y)$.

Lemma 2.4.1. (Eaton, 1967). Under the assumptions made in this section, for any $a\varepsilon A$, $i\varepsilon a$ and $j\not\varepsilon a$ imply $r_a(x,y) \leq r_{(i,j)a}(x,y)\ \forall(x,y)\varepsilon B_a$.

Proof. Let $\Phi_0 = \{(\theta,\psi):(\theta,\psi)\varepsilon\Phi,\ \theta_i = \theta_j\}$, $\Phi_1 = \{(\theta,\psi):(\theta,\psi)\varepsilon\Phi,\ \theta_i > \theta_j\}$, $\Phi_2 = \{(\theta,\psi):(\theta,\psi)\varepsilon\Phi,\ \theta_i < \theta_j\}$. Then

(2.4.1) $$r_{(i,j)a}(x,y) - r_a(x,y) = \sum_{m=0}^{2}\int_{\Phi_m}[L((\theta,\psi),(i,j)a) - L((\theta,\psi),a)]\, f(x,y|\theta,\psi)dP(\theta,\psi).$$

The invariance assumptions imply that $L((\theta,\psi),(i,j)a) = L((\theta,\psi),a)$ for $(\theta,\psi)\varepsilon\Phi_0$ and

$$\int_{\Phi_2} [L((\theta,\psi),(i,j)a) - L((\theta,\psi),a)]f(x,y|\theta,\psi)dP(\theta,\psi)$$

$$= \int_{\Phi_1} [L((\theta,\psi),a) - L((\theta,\psi),(i,j)a)]$$

$$f(x,y|(i,j)\theta,\psi)dP(\theta,\psi).$$

Thus we can write (2.4.1) as

(2.4.2) $r_{(i,j)a}(x,y) - r_a(x,y)$

$$= \int_{\Phi_1} [L((\theta,\psi),(i,j)a) - L((\theta,\psi),a)]$$

$$[f(x,y|\theta,\psi)-f(x,y|(i,j)\theta,\psi)]dP(\theta,\psi).$$

Now $(\theta,\psi)\varepsilon\Phi_1$, $(x,y)\varepsilon B_a$ imply $f(x,y|\theta,\psi)-f(x,y|(i,j)\theta,\psi) \geq 0$. Also $(\theta,\psi)\varepsilon\Phi_1$, $i\varepsilon a$, $j\not\varepsilon a$ imply $L((\theta,\psi),(i,j)a) - L((\theta,\psi),a) \geq 0$. Hence (2.4.2) ≥ 0 and (2.4.1) ≥ 0. This completes the proof.

For each $s\varepsilon S$, let $H_m(s) = \{a:a\varepsilon A, s\varepsilon B_a, |a| = m\}$ where $|a|$ denotes the number of elements in a and $H(s) = \bigcup_{m=1}^{p} H_m(s)$.

Theorem 2.4.1. Under the assumptions made in this section, for any non-negative measure P on $(\Phi,\mathcal{B}_\Phi)$ which is invariant under S_p, a sufficient condition for a procedure δ^* to be Laplace for P relative to $\mathcal{D}_I$ is

$$\delta^*(s,T(s)) = 1 \qquad \text{where}$$

$$T(s) = \{a:a\varepsilon H(s), B(a|s) = \min_{\alpha\varepsilon H(s)} B(\alpha|s)\}.$$

Proof. By Theorem 1.1 a sufficient condition for δ^* to be Laplace is $\delta^*(s,S(s)) = 1$ where $S(s) = \{a:a\varepsilon A, B(a|s) = \min_{\alpha\varepsilon A} B(\alpha|s)\}$. By the previous lemma $a\varepsilon H_m(s) \Rightarrow B(a|s) =$

$\min_{|\alpha|=m} B(\alpha|s)$. Hence $B(a|s) = \min_{\alpha\in H(s)} B(\alpha|s) \Rightarrow B(a|s) = \min_{a\in A} B(\alpha|s)$, that is, $T(s) \subseteq S(s)$.

2.5. An Intuitive Loss Function

In interpreting subset selection procedures as screening procedures, we want the selected subset to contain the true 'best', but we do not want the subset to be too large. It seems reasonable then any symmetric loss function should contain at least the following three components:

1. Incorrect Selection (ICS) $= 1 - \sum_{i\varepsilon a} I_{\{\theta_i = \max_{j=1,\ldots,p} \theta_j\}}$;

2. The size of the selected subset a, denoted by $|a|$;

3. Some measure of the average 'goodness' of the selected set, e.g. $\max_{j=1,\ldots,p} \theta_j - \sum_{i\varepsilon a} \theta_i/|a|$.

The quantity $\max_{j=1,\ldots,p} \theta_j - \max_{i\varepsilon a} \theta_i$ can be considered as a combination of 1. and 3.

Traditionally ICS and $|a|$ have received the most attention. Goel and Rubin (1975) considered $c_1(\max_{j=1,\ldots,p} \theta_j - \max_{i\varepsilon a} \theta_i) + c_2 |a|$ and using essentially Theorem 2.3.7 obtained results concerning the form of Bayes procedures.

Chernoff and Yahav (1977) considered $c_1(\max_{j=1,\ldots,p} \theta_j - \max_{i\varepsilon a} \theta_i) + c_2(\max_{j=1,\ldots,p} \theta_j - \sum_{i\varepsilon a} \theta_i/|a|)$ and performed Monte Carlo studies assuming normal populations and normal exchangeable priors. They found that in terms of Bayes risk Gupta type procedures are extremely good compared to Bayes procedures but could not offer any explanation as to why this is so. Bickel and Yahav (1977) considered c_1(ICS) + $c_2(\max_{j=1,\ldots,p} \theta_j - \sum_{i\varepsilon a} \theta_i/|a|)$ and studied the asymptotic behavior of Bayes procedures as $p \to \infty$ assuming normal populations.

As it turns out the form of the Bayes procedures is fairly sensitive to what is used for 3. As the inclusion of the components 1. and 2. seem to be in general agreement, it seems reasonable to consider loss functions of the form $L(\phi,a) = c_1(\text{ICS}) + c_2\ |a|$.

Some care must be taken in defining ICS when several of the θ's are tied for the maximum. Traditionally when this happens one of the θ's tied for the maximum is arbitrarily 'tagged' as the 'best'. This makes $P_{\theta,\psi}\{\text{ICS}\}$ a continuous function of (θ,ψ) in the usual case, namely, when the setup of the problem is symmetric, the procedure considered is permutationally invariant, and the family of possible distributions is in the exponential family of distributions. In this thesis we shall use a slightly different definition of ICS which is equivalent to the traditional definition (in terms of risk functions) in the usual case described above but not equivalent in general.

<u>Definition 2.5.1.</u> $\text{ICS}(\phi,a) = 1 - \sum_{i\varepsilon a} \sum_{m=1}^{p} (1/m) I_{\Phi_{im}}(\phi)$ where

$$\Phi_{im} = \{\phi:\phi\varepsilon\Phi,\ \theta_i = \max_{j=1,\ldots p} \theta_j,\ \sum_{\alpha=1}^{p} I_{\{\theta_\alpha = \max_{j=1,\ldots,p} \theta_j\}} = m\}.$$

Remark 2.5.1. In words, Φ_{im} is the subset of Φ where θ_i is tied with m-1 other θ's as the maximum.

To fix ideas, suppose three of the θ's are tied for the maximum, if 'a' selects two of the three, then $ICS(\phi,a) = 1/3$.

We shall use throughout the thesis the following shorthand notation.

Notation. $I_{\{\theta_i = \max\limits_{j=1,\ldots,p} \theta_j\}}(\phi) \equiv \sum\limits_{m=1}^{p} (1/m) I_{\phi_{im}}(\phi)$

Note that defined this way $I_{\{\theta_i=\max\limits_{j=1,\ldots,p} \theta_j\}}(\phi)$ is $\nearrow\{i\}$.

We now describe a Bayes procedure.

Theorem 2.5.1. If the loss function is $c_1(ICS) + c_2|a|$, then for any prior, the procedure

$$\delta_1(s,a^*) = 1$$

where

$$a^* = \{i: P\{\theta_i = \max_{j=1,\ldots,p} \theta_j | s\} \geq c_2/c_1\} \text{ if } a^* \neq \phi$$

and

$$\delta_1(s,\{i\}) = 1/N(s)$$

where

$$P\{\theta_i = \max_{j=1,\ldots,p} \theta_j | s\} = \max_{m=1,\ldots,p} P\{\theta_m = \max_{j=1,\ldots,p} \theta_j | s\},$$

$$N(s) = \sum_{i=1}^{p} I_{\{P\{\theta_i=\max\limits_{j=1,\ldots,p} \theta_j|s\} = \max\limits_{m=1,\ldots,p} P\{\theta_m = \max\limits_{j=1,\ldots,p} \theta_j|s\}\}}(s),$$

$$\text{if } a^* = \phi$$

is Bayes.

Proof. The posterior risk $B(a|s) = c_1[1-\sum_{i\varepsilon a} P\{\theta_i = \max_{j=1,\ldots,p} \theta_j|s\}] + c_2|a|$. The result follows from Theorem 1.1.

Corollary 2.5.1. If $c_2/c_1 \leq 1/p$ then δ_1 reduces to

$$\delta_2(s,a^*) = 1$$

where

$$a^* = \{i:P\{\theta_i = \max_{j=1,\ldots,p} \theta_j|s\} \geq c_2/c_1\}.$$

Proof.

$$\sum_{i=1}^{p} P\{\theta_i = \max_{j=1,\ldots,p} \theta_j|s\} = 1,$$

so

$$c_2/c_1 \leq 1/p \Rightarrow a^* \neq \phi.$$

Suppose the setup of the problem is symmetric, then δ_1 is permutationally invariant for any exchangeable prior. In describing δ_1 one can therefore assume for any $s = (x_1, x_2,\ldots,x_p,y_1,y_2,\ldots,y_q)$ that $x_1 \leq x_2 \leq \ldots \leq x_p$.

Corollary 2.5.2. Suppose the densities $\{f(\cdot,\cdot|\theta,\psi): (\theta,\psi)\varepsilon\Phi\}$ have Property M and the prior is exchangeable. For $s = (x_1,x_2,\ldots,x_p,y_1,y_2,\ldots,y_q)$ assume without loss of generality $x_1 \leq x_2 \leq \ldots \leq x_p$. Then δ_1 can be written as

$$\delta_3(s,a^*) = 1 \quad \text{where} \quad a^* = \{i:i>i^*\} \quad \text{if } i^* < p,$$

$$\delta_3(s,\{i\}) = 1/N(x)$$

for each i such that $x_i = x_p$, $N(x) = \sum_{j=1}^{p} I_{\{x_j=x_p\}}$ if $i^* = p$

where

$$i^* = \text{largest integer } i \text{ such that } P\{\theta_i = \max_{j=1,\ldots,p} \theta_j|s\} < c_2/c_1.$$

Proof. This follows from the fact that the posterior densities with respect to the prior have Property M and so by Theorem 2.3.7. $x_i \le x_j \Rightarrow P\{\theta_i = \max_{j=1,\ldots,p} \theta_j | s\} \le P\{\theta_j = \max_{j=1,\ldots,p} \theta_j | s\}$.

2.6. Uniqueness of Bayes Procedures

The following gives a sufficient condition for the uniqueness of Bayes procedures.

Theorem 2.6.1. Suppose that the setup of the problem is the symmetric setup of Section 2.2, $f(x,y|\theta,\psi) = C(\theta,\psi) \exp[\sum_{i=1}^{p} Q(\theta_i,\psi)x_i + \sum_{j=1}^{q} R_j(\psi)y_j]h(x,y)$, and the dominating measure μ is the Lebesgue measure on R^{p+q}. Suppose also the loss function is $c_1(\text{ICS}) + c_2|a|$ with $c_2/c_1 < 1/p$. Then for any exchangeable prior, δ_2 is the unique Bayes procedure.

Proof. For any $S \varepsilon \mathcal{B}_S$, $F(S|\phi)=0$ for some $\phi \varepsilon \Phi$ implies $F(S|\phi)=0$ for all $\phi \varepsilon \Phi$. Hence for any prior, $F(\cdot|\phi)$ is absolutely continuous with respect to Π for every $\phi \varepsilon \Phi$. Since the loss function is bounded, for any prior P and any $\delta \varepsilon \mathcal{D}$, $B(P,\delta) < \infty$. Hence for Corollary 1.1 to apply we need to show $A^*(s)$ as defined in (1.1) consists of a single point of A a.e. Π. Without loss of generality assume $h(s) > 0$ for all $s \varepsilon S$. The posterior density with respect to the prior has the form

$$C(\theta,\psi)\exp[\sum_{i=1}^{p} Q_i(\theta,\psi)x_i + \sum_{j=1}^{q} R_j(\theta,\psi)y_j]h'(x,y)$$

and S is in the natural parameter space of the posterior. Consider the set E of (real) solutions of $P\{\theta_i = \max_{j=1,\ldots,p} \theta_j | s\} = c_2/c_1$ in D where D is the interior of the natural parameter space. For fixed $(x_2,\ldots,x_p,y_1,\ldots,y_q)$, $P\{\theta_i = \max_{j=1,\ldots,p} \theta_j | s\}$ is analytic in x_1. Likewise for every other component of (x,y). This fact implies $\mu(E) = 0$ or $E = D$. We will show this for $p=2$, $q=0$. The proof for the general case follows by induction.

Let D_1 and E_1 be the projections of D and E onto the first axis respectively. For each $x_1 \varepsilon D_1$, let $D_{x_1} = \{x_2 : (x_1,x_2) \varepsilon D\}$ and $E_{x_1} = \{x_2 : (x_1,x_2) \varepsilon E\}$. If $\mu(E) > 0$, then there exist $\varepsilon > 0$ and $E_1' \subseteq E_1$ such that $\mu(E_1') > 0$ and $\mu(E_{x_1}) > \varepsilon$ for each $x_1 \varepsilon E_1'$. In particular for each $x_1 \varepsilon E_1'$, E_{x_1} is uncountable and hence is the whole of D_{x_1}. But then there exists an interval (a,b) such that for each $x_2 \varepsilon (a,b)$, the set $\{x_1 : (x_1,x_2) \varepsilon E\}$ is uncountable for otherwise E_1' would be countable. Hence $\{(x_1,x_2) : (x_1,x_2) \varepsilon D,\ x_2 \varepsilon (a,b)\} \subseteq E$. By analytic continuation we have $D = E$.

Suppose $D = E$. Then $P\{\theta_i = \max_{j=1,\ldots,p} \theta_j | s\} = c_2/c_1$ a.e. Π. By symmetry $P\{\theta_i = \max_{j=1,\ldots,p} \theta_j | s\} = c_2/c_1$ a.e. Π for each i, $i=1,\ldots,p$. But then $\sum_{i=1}^{p} p\{\theta_i = \max_{j=1,\ldots,p} \theta_j | s\} = p(c_2/c_1) < 1$ a.e. Π. Contradiction. Hence $\mu(E) = 0$ which implies $A^*(s)$ as defined in (1.1) consists of a single point of A a.e. Π.

The following figure might clarify the proof.

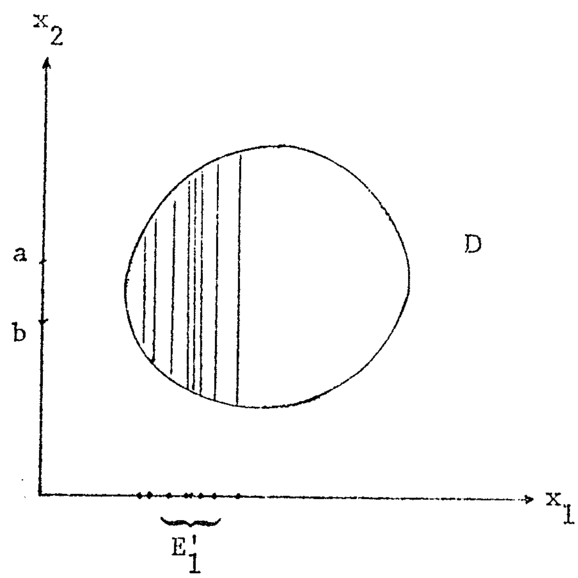


Figure 1. The Sets D and E.

2.7. Form Of Bayes Procedures When The Densities Have TMLR and the Loss Function is $c_1(\text{ICS}) + c_2|a|$

We have seen earlier that the concept of Property M is important in the study of Bayes procedures partially because the densities of (x,y) given (θ,ψ) having Property M implies for any exchangeable prior P the posterior densities of (θ,ψ) given (x,y) with respect to P have Property M and hence the posterior distributions have the one-dimensional stochastic monotone property stated in Theorem 2.3.7. Likewise, the concept of Total Monotone Likelihood Ratio (TMLR) is important because the densities of (x,y) given (θ,ψ) having TMLR implies for any prior P the posterior densities of (θ,ψ) given (x,y) with respect to P have TMLR and hence the posterior distributions have Total Stochastic Monotone Property

(TSMP). We state this latter fact formally.

Lemma 2.7.1. Suppose the densities $\{f(x,y|\theta,\psi):(\theta,\psi)\varepsilon\Phi\}$ have TMLR. Then for any prior P the posterior densities of (θ,ψ) given (x,y) with respect to P have TMLR and hence the posterior distributions have TSMP.

Proof. The posterior densities with respect to P are $\{f(x,y|\theta,\psi)/\int_\Phi f(x,y|\theta,\psi)dP(\theta,\psi):(x,y)\varepsilon S\}$ a.e. Π. They have TMLR because of the symmetry in (x,y) and (θ,ψ) in the definition of TMLR. The rest follows from Corollary 2.3.1.

Theorem 2.7.1. Suppose the densities $\{f(x,y|\theta,\psi):(\theta,\psi)\varepsilon\Phi\}$ have TMLR and the loss function $L(\phi,a) = c_1(ICS) + c_2|a|$ is such that $c_2/c_1 \leq 1/p$. Then the nonrandomized Bayes Procedure δ_2 defined in Section 2.5: $\delta_2(s,a^*) = 1$ where $a^* = \{i:P\{\theta_i=\max_{j=1,\ldots,p}\theta_j|s\} \geq c_2/c_1\}$ is monotone.

Proof. $I_{\{\theta_i=\max_{j=1,\ldots,p}\theta_j\}}$ is $\nearrow\{i\}$. By the previous lemma and Corollary 2.3.1, $P\{\theta_i=\max_{j=1,\ldots,p}\theta_j|x,y\}$ is $\nearrow\{i\}$. Hence δ_2 is monotone.

However, if the setup of the problem is symmetric, then the condition that $c_2/c_1\leq 1/p$ is not needed.

Theorem 2.7.2. Suppose the setup of the problem is the symmetric setup of Section 2.2, the densities $\{f(x,y|\theta,\psi):(\theta,\psi)\varepsilon\Phi\}$ have TMLR and the loss function is $c_1(ICS) + c_2|a|$. Then for any exchangeable prior, the Bayes procedure δ_3 defined in Section 2.5,

$$\delta_3(s,a^*) = 1 \quad \text{where} \quad a^* = \{i:P\{\theta_i=\max_{j=1,\ldots,p}\theta_j|s\} \geq c_2/c_1\} \quad \text{if} \quad a^* \neq \phi,$$

$$\delta_3(s,\{i\}) = 1/N(x)$$

where

$$x_i = \max_{j=1,\ldots,p} x_j, \quad N(x) = \sum_{i=1}^{p} I_{\{x_i = \max_{j=1,\ldots,p} x_j\}}$$

$$\text{if } a^* = \phi$$

is monotone.

We have the following trivial

Corollary 2.7.1. If the hypothesis of Theorem 2.7.1. (Theorem 2.7.2) is satisfied, and if for a prior (exchangeable prior) there is a unique Bayes procedure, then that Bayes procedure is monotone.

We are now in a position to prove the following essentially complete class theorem.

Theorem 2.7.3. Suppose the setup of the problem is the symmetric setup of Section 2.2, $f(x,y|\theta,\psi) = C(\theta,\psi) \exp [\sum_{i=1}^{p} Q(\theta_i,\psi)x_i + \sum_{j=1}^{q} R_j(\psi)y_j]h(x,y)$ where, for fixed ψ, Q is nondecreasing in θ_i, and the dominating measure μ is the Lebesque measure on R^{p+q}. Suppose also the loss function is $c_1(\text{ICS}) + c_2|a|$ with $c_2/c_1 < 1/p$. Then relative to $\mathcal{D}_I$, the class of monotone invariant procedures forms an essentially complete class.

Proof. Since the setup of the problem is symmetric, Theorem 2.2.1. implies the closure of Bayes procedures relative to $\mathcal{D}$ for exchangeable priors forms an essentially complete class relative to $\mathcal{D}_I$. By Theorem 2.6.1, for each exchangeable prior, δ_2 is the unique Bayes procedure. δ_2 is of course invariant. By Theorem 2.3.3. $\{f(\cdot,\cdot|\theta,\psi):(\theta,\psi)\in\Phi\}$

has TMLR. So by Theorem 2.7.1, δ_2 is monotone. Thus it remains to prove that limits of monotone procedures are monotone.

Suppose a net $\{\delta^\alpha\}$ of monotone procedures converges to δ. Suppose that δ is not monotone$_{\{i\}}$, that is, there exists i, S_1, $S_2 \varepsilon S$, $\mu(S_1)$, $\mu(S_2) > 0$, $s_1 \leq s_2$ for all $s_1 \varepsilon S_1$, $s_2 \varepsilon S_2$, and $\underset{s\varepsilon S_1}{\text{ess sup}}\ \delta_i(s) > \underset{s'\varepsilon S_2}{\text{ess inf}}\ \delta_i(s')$. It is easy to check that this implies there exists $W_1 \subseteq S_1$, $W_2 \subseteq S_1$, $0 < \mu(W_1)$, $\mu(W_2) < \infty$ such that $\delta_i(s_1) > \delta_i(s_2) + \varepsilon$ some $\varepsilon > 0$ for all $s_1 \varepsilon W_1$, $s_2 \varepsilon W_2$. Now $[\int_{W_1} \delta_i^\alpha(s)d\mu(s)]\mu(W_2) \leq \underset{s\varepsilon W_1}{\text{ess sup}}\ \delta_i^\alpha(s)\mu(W_1)\mu(W_2) \leq \underset{s\varepsilon W_2}{\text{ess inf}}\ \delta_i^\alpha(s)\mu(W_1)\mu(W_2) \leq [\int_{W_2} \delta_i^\alpha(s)d\mu(s)]\mu(W_1)$. Since $\int_{W_1} \delta_i^\alpha(s)d\mu(s) \to \int_{W_1} \delta_i(s)d\mu(s)$ and $\int_{W_2} \delta_i^\alpha(s)d\mu(s) \to \int_{W_2} \delta_i(s)d\mu(s)$ we have $[\int_{W_1} \delta_i(s)d\mu(s)]\mu(W_2) \leq [\int_{W_2} \delta_i(s)d\mu(s)]\mu(W_1)$. But $\delta_i(s_1) > \delta_i(s_2) + \varepsilon$ for all $s_1 \varepsilon W_1$, $s_2 \varepsilon W_2 \Rightarrow \underset{s\varepsilon W_1}{\text{ess inf}}\ \delta_i(s) \geq \underset{s\varepsilon W_2}{\text{ess sup}}\ \delta_i(s) + \varepsilon$. Hence $[\int_{W_1} \delta_i(s)d\mu(s)]\mu(W_2) \geq \underset{s\varepsilon W_1}{\text{ess inf}}\ \delta_i(s)\mu(W_1)\mu(W_2) \geq [\underset{s\varepsilon W_2}{\text{ess sup}}\ \delta_i(s) + \varepsilon]\mu(W_1)\mu(W_2) \geq [\int_{W_2} \delta_i(s)d\mu(s)]\mu(W_1) + \varepsilon\mu(W_1)\mu(W_2)$. Contradiction. Therefore δ is monotone and the proof is complete.

Our essentially complete class theorem is not entirely satisfactory in two ways. First, permutational invariance was used. The theorem ought to be true even if the setup is not symmetric. Two, the theorem is an essentially complete class theorem rather than a complete class theorem. By using the strict versions of Total Monotone Likelihood Ratio and Total Stochastic Monotone Property it may be possible to give a constructive proof of a complete class theorem. In this connection see Brown, Cohen and Strawderman (1976).

CHAPTER 3

SIMULATION RESULTS ON SUBSET SELECTION PROCEDURES

In the last two chapters our search for procedures that perform well on the average led us to the investigation of Bayes procedures. But Bayes procedures typically require numerical integration to implement and sometimes this makes them unsuitable for practical use. Besides, the use of Bayes procedures is by no means universally accepted. So if there are easy to use classical procedures that have performance close to those of Bayes procedures, then these classical procedures ought to be used. This possibility is explored in this chapter for the case of normal populations with normal exchangeable priors. The classical procedures of Gupta type and of Seal type are compared with Bayes procedures in terms of integrated risks. Though the Monte Carlo studies were done for the case p=8 only, indications are for each loss function and prior pair, there is always a procedure of Gupta type that performs almost as well as the corresponding Bayes procedure, while this is true for Seal type procedures only when the normal prior is very informative compared to the observations. In this connection Chernoff and Yahav (1977) earlier made similar studies for the loss functions $c_1(\max_{j=1,\ldots p} \theta_j - \max_{i\varepsilon a} \theta_i) + c_2(\max_{j=1,\ldots,p} \theta_j - \sum_{i\varepsilon a} \theta_i/|a|)$. They also found that Bayes procedures can be approximated closely by Gupta type procedures. Though we do not yet know whether Gupta type procedures are robust against priors, we can recommend their use in the case of normal populations as procedures that have at least some optimality properties.

Notation. In this chapter as well as the next, we shall adopt the following convention. Let $\theta_{[1]} \leq \cdots \leq \theta_{[p]}$ be the ordered components of θ. If there is exactly one θ_i such that $\theta_i = \theta_{[p]}$ then we shall denote $\theta_i = \theta_{[p]}$ by $\max_{j=1,\ldots,p} \theta_j$. If more than one θ_i are tied for $\theta_{[p]}$, then eaactly one of these θ_i is tagged as $\max_{j=1,\ldots,p} \theta_j$. For any subset selection procedure R, let $P_\theta\{CS|R\}$ denote the probability of a correct selection under θ when procedure R is used. More precisely, if $\delta(s,a)$ is the probability assigned to the subset a of $\{1,\ldots,p\}$ by R when s is observed, then $P_\theta\{CS|R\}$ is the expected value of $\sum_{a\varepsilon A} I_{\{\max_{j=1,\ldots,p} \theta_j \varepsilon \{\theta_i, i\varepsilon a\}\}} \delta(s,a)$ under θ, where I denotes the indicator function. Also we shall denote by $E_\theta(S|R)$ the expected subset size under θ when the procedure R is used. If we let $P_\theta(i|R)$ be the probability of selecting i under θ when procedure R is used, that is, $P_\theta(i|R)$ is the expected value of $\sum_{a\varepsilon A} I_{\{i\varepsilon a\}} \delta(s,a)$, then it is easy to see that $E_\theta(S|R) = \sum_{i=1}^{p} P_\theta(i|R)$.

Consider the following model:

$(X|\mu) \sim N(\mu, I)$,

where $X=(X_1,\ldots,X_p)$, $\mu=(\mu_1,\ldots,\mu_p)$ are vectors in R^p and

I is the pxp identity matrix,

$\mu \sim N(m\underset{\sim}{1},\ rI + sU)$ where m, r, s are constants, $r > 0$, $-r/p < s < r$, and $U = \underset{\sim}{1}'\underset{\sim}{1}$ where $\underset{\sim}{1} = (1,\ldots,1)$.

The above model is equivalent to

$$(X,\mu) \sim N\left((m\underset{\sim}{1}, m\underset{\sim}{1}), \begin{pmatrix} (1+r)I + sU & rI + sU \\ rI + sU & rI + sU \end{pmatrix}\right).$$

Hence a posteriori

$(\mu|x) \sim N(\hat{\mu}, \Sigma_{22.1})$ where

$$\hat{\mu} = m\underset{\sim}{1} + (rI + sU)[(1+r)I + sU]^{-1}(x - m\underset{\sim}{1})$$

$$= (r/1+r)\ x + \text{a multiple of } \underset{\sim}{1}$$

and

$$\Sigma_{22.1} = (rI + sU) - (rI + sU)[(1+r)I + sU]^{-1}(rI + sU)$$

$$= (rI + sU) - (rI + sU)(1+r)^{-1}\left[I - \frac{\frac{s}{1+r}}{1+p\frac{s}{1+r}}U\right](rI+sU)$$

$$= rI - (r^2/1+r)I + \text{a multiple of } U$$

$$= (r/1+r)I + \text{a multiple of } U.$$

Bayes Procedure.

Recall that the Bayes procedure , denoted by R_B, is as follows: Select i if and only if $X_i = \max_{j=1,\ldots,p} X_j$ and/or $P\{\mu_i = \max_{j=1,\ldots,p} \mu_j | X\} \geq c_2/c_1$.

Gupta's Procedure.

The classical procedure studied in Gupta (1956,1965), denoted by R_G, is as follows: Select i if and only if $X_i \geq \max_{j \neq i} X_j - d_G(P^*)$ where $d_G(P^*)$ is just large enough such that $\inf_{\mu \in R^p} P_\mu\{CS|R_G\} \geq P^*$ where $P^*(1/p \leq P^* \leq 1)$ is predetermined.

Seal's Procedure.

One particular procedure in the class of procedure studied in Seal (1955), denoted by R_S, is as follows: Select i if and only if $X_i \geq \sum_{j\neq i} X_j/(p-1) - d_S(P^*)$ where $d_S(P^*)$ is just large enough such that $\inf_{\mu\in R^p} P_\mu\{CS|R_S\} \geq P^*$ where $P^*(1/p \leq P^* \leq 1)$ is pre-determined.

We shall try to show that both Gupta type procedures and Seal type procedures are intuitive first approximations to Bayes procedures. First we state a well known lemma.

Lemma 3.1. Suppose $\underset{1xp}{V} \sim N(\underset{1xp}{\underset{\sim}{b}}, \underset{pxp}{I} + \delta \underset{pxp}{U})$, $-1/p < \delta < 1$. Then there exist λ and a such that if

$$W = (W_0, W_1, \ldots, W_p) \sim N\left((0,\underset{\sim}{b}), \begin{bmatrix} 1 & \lambda & \cdots & \lambda \\ \lambda & & & \\ \vdots & & I_{pxp} & \\ \lambda & & & \end{bmatrix}\right), \text{ then } (W_1 - aW_0, \ldots,$$

$W_p - aW_0)$ has the same distribution as V.

As a consequence of this lemma, the expected value of any measurable translation invariant function of V can be computed under $V \sim N(\underset{\sim}{b}, I)$ rather than $V \sim N(\underset{\sim}{b}, I+\delta U)$. We prove this formally.

Theorem 3.1. Suppose $V \sim N(\underset{\sim}{b}, I+\delta U)$, $-1/p < \delta < 1$. If h is measurable and translation invariant, that is, $h(v+c\underset{\sim}{1}) = h(v)$ for all $c\in R$, then

$$\int_{R^p} h(v)\, d\Phi_{\underset{\sim}{b}, I+\delta U}(v) = \int_{R^p} h(v)\, d\Phi_{\underset{\sim}{b}, I}(v)$$

where $\Phi_{\underset{\sim}{a}, B}$ denotes the normal cdf with mean vector $\underset{\sim}{a}$ and variance-covariance matrix B.

Proof.

$$\int_{R^p} h(v)\, d\Phi_{\underset{\sim}{b}, I+\delta}(v)$$

$$= \int_{R^{p+1}} h(w_1 - w_0, \ldots, w_p - w_0)\, d\Phi_{(0,\underset{\sim}{b}), \begin{bmatrix} 1 & \lambda \ldots \lambda \\ \lambda & \\ \vdots & I_{pxp} \\ \lambda & \end{bmatrix}}(w)$$

$$= \int_{R^{p+1}} h(w_1, \ldots, w_p)\, d\Phi_{(0,\underset{\sim}{b}), \begin{bmatrix} 1 & \lambda \ldots \lambda \\ \lambda & I_{pxp} \\ \vdots & \\ \lambda & \end{bmatrix}}(w)$$

$$= \int_{R^p} h(v)\, d\Phi_{\underset{\sim}{b}, I}(v).$$

Let us examine the Bayes procedure more closely. We first note that the set $\{\mu_i = \max_{j=1,\ldots,p} \mu_j\}$ is both translation invariant and scale invariant in μ. So

$$P\{\mu_i = \max_{j=1,\ldots,p} \mu_j | x\}$$

$$= \int I_{\{\mu_i = \max_{j=1,\ldots,p} \mu_j\}} d\Phi_{(r/1+r)x + \text{multiple of } \underset{\sim}{1},\ (r/1+r)I + \text{multiple of } U}(\mu)$$

$$= \int I_{\{\mu_i = \max_{j=1,\ldots,p} \mu_j\}} d\Phi_{(r/1+r)x + \text{multiple of } \underset{\sim}{1},\ (r/1+r)I}(\mu) \quad \text{by Theorem 3.1}$$

$$= \int I_{\{\mu_i = \max_{j=1,\ldots,p} \mu_j\}} d\Phi_{(r/1+r)x,\ (r/1+r)I}(\mu) \quad \text{by translation invariance}$$

$$= \int I_{\{\mu_i = \max_{j=1,\ldots,p} \mu_j\}} d\Phi_{(r/1+r)^{1/2}x,\ I}(\mu) \quad \text{by scale invariance}$$

$$= \int I_{\{\mu_i - \mu_j + (r/1+r)^{1/2}(x_i - x_j) \geq 0,\ j \neq i\}} d\Phi_{\underset{\sim}{0}, I}(\mu) \tag{3.1}$$

Hence the Bayes procedure selects i if and only if the p-1 dimensional vector $(X_i-X_j, j\neq i)$ is reasonably large. As a first approximation we may replace $(X_i-X_j, j\neq i)$ by $(X_i-\max_{j\neq i} X_j)\,\underset{\sim}{1}$ or by $(X_i - \sum_{j\neq i} X_j/(p-1))\,\underset{\sim}{1}$. Now

$$(3.1)\qquad \begin{aligned} &\geq \int I_{\{\mu_i-\mu_j+(r/1+r)^{1/2}(x_i-\max_{j\neq 1} x_j),\, j\neq i\}}\, d\Phi_{\underset{\sim}{0},I}(\mu) \\ &\leq \int I_{\{\mu_i-\mu_j+(r/1+r)^{1/2}(x_i-\sum_{j\neq i} x_j/(p-1)),\, j\neq i\}}\, d\Phi_{\underset{\sim}{0},I}(\mu) \end{aligned}$$

where the first inequality is obvious and the second inequality follows from the result of Marshall and Olkin (1974). So more realistically we would approximate $(x_i-x_j, j\neq i)$ by $(x_i-\max_{j\neq i} x_j+c_G)\,\underset{\sim}{1}$ and $(x_i-\sum_{j\neq i} x_j/(p-1)-c_S)\,\underset{\sim}{1}$ where c_G and c_S are positive numbers. In any case, the first approximation leads to the procedure of selecting i if and only if $X_i-\max_{j\neq i} X_j$ is reasonably large which is Gupta's procedure. The second approximation leads to the procedure of selecting i if and only if $X_i - \sum_{j\neq i} X_j/(p-1)$ is reasonably large which is Seal's procedure.

Monte Carlo studies were performed to determine how good these approximations are in terms of integrated risks. The three types of procedures being considered--Bayes procedures, Gupta type, and Seal type, all are translation invariant procedures. The loss function $L(\mu,a)$ for fixed a is both translation and scale invariant in μ. In computing the integrated risks the following sequence of reductions shows that the integrated risk of any translation invariant procedure δ is independent of m and s.

$$B(\delta, N(m\underset{\sim}{1}, rI + sU))$$

$$= \iint L(\mu,\delta(x))\, d\Phi_{(r/1+r)x + \text{multiple of } \underset{\sim}{1},\ (r/1+r)I + \text{multiple of } U}(\mu)\ d\Phi_{m\underset{\sim}{1},(1+r)I + sU}(x)$$

$$= \iint L(\mu,\delta(x))\, d\Phi_{(r/1+r)x,(r/1+r)I}(\mu)\ d\Phi_{m\underset{\sim}{1},(1+r)I + sU}(x) \text{ by Theorem 3.1}$$

$$= \iint L(\mu,\delta(x))\, d\Phi_{(r/1+r)^{1/2}x,I}(\mu)\, d\Phi_{m\underset{\sim}{1},(1+r)I + sU}(x)$$

$$= \iint L(\mu,\delta(x))\, d\Phi_{(r/1+r)^{1/2}x,I}(\mu)\, d\Phi_{\underset{\sim}{0},(1+r)I}(x) \quad \text{by Theorem 3.1}$$

$$= \iint L(\mu,\delta((r/1+r)^{-1/2}z))\, d\Phi_{z,I}(\mu)\, d\Phi_{\underset{\sim}{0},rI}(z).$$

For each $(r, c_2/c_1)$ pair, r and c_2/c_1 indexing the priors and the loss functions respectively, the best Gupta type procedure and the best Seal type procedure are found by simulation and their integrated risks are compared with the Bayes risk by simulation. As it turns out, throughout the range of r and c_2/c_1 studied, for each $(r,c_2/c_1)$, there is always a Gupta type procedure that perform almost as well as the Bayes procedure, while this being true for Seal type procedures only when r is roughly less than or equal to 1, i.e. only when the prior is very informative compared to the observations. For each $(r,c_2/c_1)$, Table I gives the approximate value of the optimal d_G, the constant associated

Table I. Lists for Each Prior and Loss Function Pair The Approximate Value of d_G Corresponding to the Best Gupta Type Procedure

	c_1/c_2													
$r^{1/2}$	10	12	14	16	18	20	25	30	35	40	45	50	55	60
0.5	1.5	1.7	1.9	2.1	2.3	2.4	2.7	2.9	3.1	3.2	3.3	3.4	3.6	3.6
0.6	1.7	1.9	2.1	2.2	2.4	2.5	2.7	2.9	3.1	3.2	3.3	3.4	3.5	3.5
0.8	1.5	1.7	1.9	2.0	2.1	2.2	2.4	2.5	2.7	2.8	2.9	2.9	3.0	3.0
1.0	1.6	1.8	1.9	2.0	2.1	2.2	2.4	2.5	2.6	2.7	2.8	2.9	3.0	3.0
2.0	1.2	1.6	1.7	1.8	1.9	2.0	2.1	2.2	2.3	2.4	2.4	2.5	2.6	2.6
3.4	1.5	1.7	1.8	1.8	1.9	2.0	2.1	2.2	2.3	2.4	2.4	2.5	2.5	2.6
4.0	1.7	1.8	1.9	2.0	2.1	2.1	2.2	2.4	2.4	2.5	2.6	2.6	2.7	2.7
6.0	1.8	1.9	2.0	2.1	2.1	2.2	2.3	2.4	2.5	2.6	2.6	2.7	2.7	2.8
8.0	1.8	1.9	2.0	2.1	2.1	2.2	2.3	2.4	2.5	2.5	2.6	2.7	2.7	2.7
10.	1.8	1.9	2.0	2.1	2.1	2.2	2.2	2.4	2.5	2.5	2.6	2.7	2.7	2.7

with the best Gupta type procedure. Table II gives for each $(r, c_2/c_1)$ the simulated integrated risks using the Bayes procedure and the best Gupta type procedure. In this connection Chernoff and Yahav (1977) made similar studies for the loss function $c_1(\max_{j=1,\ldots,p} \theta_j - \max_{i\varepsilon a} \theta_i) + c_2(\max_{j=1,\ldots,p} \theta_j - \sum_{i\varepsilon a} \theta_i/|a|)$. They also found that Bayes procedures can be closely approximated by Gupta type procedures. Notice that the values in Table II depend moderately on c_2/c_1 but for fixed c_2/c_1 are relatively insensitive to r. For the loss function they studied, Chernoff and Yahav (1977) also found this phenomenon.

The question that has to be answered before Gupta type procedures can be recommended as 'the' procedures to use in all normal populations situations is how they perform under priors other than the normal, i.e. are they robust against priors? Simulation studies in these cases became more difficult and have not yet been done. But until they are done we can still recommend the use of Gupta type procedures. They have at least some optimality properties.

Table II. Lists for Each Prior and Loss Function Pair the Simulated Integrated Risks of the Bayes Procedure and the Best Gupta Type Procedure in That Order

$r^{1/2}$	c_1/c_2 = 10	12	14	16	18	20	25	30	35	40	45	50	55	60
0.5	.565	.508	.471	.427	.367	.352	.296	.248	.214	.192	.190	.155	.142	.129
	.578	.513	.482	.437	.376	.359	.301	.257	.219	.195	.176	.160	.149	.130
0.6	.549	.460	.429	.396	.360	.324	.291	.242	.212	.185	.170	.147	.139	.128
	.567	.465	.437	.405	.370	.327	.294	.242	.215	.188	.171	.148	.139	.128
0.8	.474	.430	.364	.366	.338	.315	.256	.216	.220	.173	.148	.141	.130	.119
	.480	.442	.368	.377	.350	.318	.263	.225	.220	.187	.149	.145	.137	.122
1.0	.424	.356	.333	.299	.285	.261	.220	.190	.176	.161	.139	.131	.114	.111
	.436	.365	.339	.305	.292	.290	.224	.198	.180	.164	.145	.139	.117	.112
2.0	.242	.221	.247	.187	.159	.146	.119	.125	.105	.099	.089	.082	.066	.069
	.250	.228	.255	.194	.162	.151	.124	.127	.108	.101	.091	.089	.067	.071
3.0	.209	.186	.165	.144	.132	.142	.097	.091	.071	.067	.068	.052	.057	.043
	.217	.189	.169	.146	.132	.146	.098	.094	.071	.069	.071	.053	.058	.044
4.0	.153	.183	.142	.128	.124	.085	.076	.086	.057	.048	.049	.051	.044	.040
	.156	.187	.142	.130	.127	.086	.077	.087	.057	.048	.049	.051	.046	.040
6.0	.132	.121	.107	.096	.112	.083	.072	.051	.046	.045	.045	.036	.036	.032
	.132	.124	.108	.096	.117	.083	.074	.051	.047	.046	.046	.036	.036	.033
8.0	.122	.107	.095	.108	.078	.082	.059	.057	.046	.041	.039	.033	.031	.025
	.124	.108	.095	.109	.078	.082	.060	.057	.046	.041	.040	.033	.032	.025
10.	.147	.093	.088	.073	.085	.065	.061	.044	.044	.037	.034	.032	.030	.020
	.147	.093	.088	.073	.086	.065	.061	.044	.046	.037	.034	.032	.031	.020

CHAPTER 4

SOME ROBUST AND NONPARAMETRIC SUBSET SELECTION PROCEDURES

The search for good procedures generally has to be carried out individually for each distribution. In the last chapter we found that in the case of normal populations with variances known, Gupta type procedures have some near-optimality properties. Now consider the following location model. Suppose $X_{i\alpha}, i=1,\ldots,p, \alpha=1,\ldots,n$ are random variables such that their joint distribution is $\prod_{i=1}^{p} \prod_{\alpha=1}^{n} F(x_{i\alpha}-\theta_i)$ where F is only partially known or totally unknown and $\theta=(\theta_1,\ldots,\theta_p)$ is an unknown vector in R^p. Suppose we want to select in terms of the θ_i. There is a large body of literature dealing with robust and nonparametric estimation of the location parameter. All the good known estimators are asymptotically normal under reasonable regularity conditions. Hence, intuitively, Gupta type procedures based on these estimators should have good performance. In this chapter we propose two procedures to be used in the ε-contaminated normal populations case. These two procedures asymptotically control the probability of a correct selection. We also propose a third procedure to be used in the case where F is absolutely continuous but otherwise unknown. This procedure controls the probability of a correct selection for any sample size n. Since it is based essentially on the Hodges-Lehmann estimator, it inherits the high asymptotic relative efficiency of that estimator. It should be pointed out that the problem of nonparametric subset selection is an old one. Some of the earlier references are Lehmann (1963c), Puri and Puri (1969, 1968), McDonald (1969) and Rizvi and Woodworth (1970). The approach here differs from the earlier ones in that direct estimators of the parameter are employed in constructing the procedures.

4.1. ε-Contaminated Normal Populations With Scale Known

Let $X_{i\alpha}$, $i=1,\ldots,p, \alpha=1,\ldots,n$ be random variables having the joint distribution

$$\prod_{i=1}^{p} \prod_{\alpha=1}^{n} F(x_{i\alpha}-\theta_i)$$

where $F = (1-\varepsilon)\Phi + \varepsilon H$, $\varepsilon (0 \leq \varepsilon < 1)$ is a known constant, Φ is the standard normal cdf, H is an unknown symmetric distribution, and $\theta = (\theta_1,\ldots,\theta_p)$ is an unknown vector belonging to R^p.

Let C be the class of all ε-contaminated distributions, i.e.

$C = \{F:(1-\varepsilon)\Phi + \varepsilon H$, H symmetric distribution function$\}$.

We shall propose two asymptotically equivalent procedures, one based on Huber's maximum likelihood estimator, the other on the trimmed mean. One way to introduce these estimators is as follows:

The distribution F_o in C having the smallest Fisher information has the density

$$f_o(x) = (2\pi)^{-1/2}(1-\varepsilon)e^{-\rho_o(x)}$$

where

$$\rho_o(x) = \begin{cases} 1/2\ x^2 & \text{for } |x| < k \\ k|x| - 1/2\ k^2 & \text{for } |x| \geq k \end{cases}$$

with k depending on ε through

$$\varepsilon/1-\varepsilon = 2\phi(k)/k - 2\Phi(-k),$$

where $\phi = \Phi'$ is the standard normal density. Let

$$\psi_o(x) = \rho_o'(x) = \begin{cases} x & \text{for } |x| < k \\ k \text{ sign } (x) & \text{for } |x| \geq k \end{cases}.$$

We shall denote the maximum likelihood estimator of θ_i with respect to the above F_o by M_{in}. It is the solution of

$$\sum_{\alpha=1}^{n} \psi_o(x_{i\alpha}-\theta_i) = 0.$$

The (asymptotically) best estimator of θ_i based on linear combinations of order statistics is the trimmed mean $\sum_{\beta=1}^{n} h(\beta/(n+1))x_{i(\beta)}/n$ where $x_{i(\beta)}$ is the βth order statistic from $x_{i\beta}$, $\beta=1,\ldots,n$,

$$h(t) = \begin{cases} \text{constant for } F_o(-k)<t<F_o(k) \\ 0 \quad \text{otherwise} \end{cases}$$

such that $\sum_{\beta=1}^{n} h(\beta/(n+1))/n = 1$. We shall denote the trimmed mean estimator of θ_i by L_{in}.

The following results are well known. See Huber (1964), Brickel (1967), and Jackel (1971).

<u>Theorem 4.1.1</u>. Under the sole assumption that H is symmetric, $n^{1/2}(M_{in}-\theta_i)$ is asymptotically normal with mean 0.

Let $\sigma_{M,F}^2$ be the asymptotic variance of $n^{1/2}(M_{in}-\theta_i)$ under F, then $\sup_{F\varepsilon C} \sigma_{M,F}^2 = \sigma_{M,F_o}^2 = \sigma_o^2$ where $\sigma_o^2 = [(1-\varepsilon)E_\phi\psi_o^2 + \varepsilon k^2]/[(1-\varepsilon)E_\phi\psi_o']^2$. Under the additional assumption that H has a continuous derivative, $n^{1/2}(L_{in}-\theta_i)$ is asymptotically normal with mean 0. Let $\sigma_{L,F}^2$ be the asymptotic variance of $n^{1/2}(L_{in}-\theta_i)$ under F, then under the additional assumption on H, $\sup_{F\varepsilon C} \sigma_{L,F}^2 = \sigma_{L,F_o}^2 = \sigma_o^2$ also.

Remark 4.1.1. Tables of σ_o^2 are available. See for example Huber (1964).

Let $P^*(1/p < P^* < 1)$ be the desired minimum probability of a correct selection. Let d be the positive number such that

$$\int \Phi^{p-1}(x+d)\,d\Phi(x) = P^* .$$

We now describe the two proposed procedures.

Procedure $R_L(n)$ Based on the Trimmed Mean

Select i if and only if $L_{in} \geq \max_{j=1,\ldots,p} L_{jn} - d\sigma_o/n^{1/2}$

Procedure $R_M(n)$ Based on Huber's Maximum Likelihood Estimator

Select i if and only if $M_{in} \geq \max_{j=1,\ldots,p} M_{jn} - d\sigma_o/n^{1/2}$.

Theorem 4.1.2. Under the assumption that H is symmetric, $\sup_{F\varepsilon C} \liminf_{n\to\infty} \inf_{\theta\varepsilon R^p} P\{CS|R_M(n)\} = P^*$. Under the additional assumption that H has a continuous derivative,

$$\sup_{F\varepsilon C} \liminf_{n\to\infty} \inf_{\theta\varepsilon R^p} P\{CS|R_L(n)\} = P^*.$$

Proof. Let F_n be the common distribution of $L_{in}-\theta_i$, $i=1,\ldots,p$, under F. Suppose $\theta_i = \max_{j=1,\ldots,p} \theta_j$, then

$$P_{F,\theta}\{CS|R_M(n)\} = \int \prod_{j\neq i} F_n(x+d\sigma_o/n^{1/2}-\theta_j)dF_n(x-\theta_i)$$

which is nondecreasing in θ_i and nonincreasing in every other component of θ. Hence

$$\inf_{\theta\varepsilon R^p} P_{F,\theta}\{CS|R_M(n)\} = P_{F,\underset{\sim}{0}}\{CS|R_M(n)\}.$$

where $\underset{\sim}{0} = (0,\ldots,0)$. Now

$$\lim_{n\to\infty} P_{F,\underset{\sim}{0}}\{CS|R_M(n)\} = \int\Phi^{p-1}(x+d\sigma_o/\sigma_{M,F})d\Phi(x).$$

But $\sup_{F\varepsilon C} \sigma^2_{M,F} = \sigma_o^2$. Hence

$$\inf_{F\varepsilon C}\lim_{n\to\infty}\inf_{\theta\varepsilon R^p} P_{F,\theta}\{CS|R_M(n)\} = \int\Phi^{p-1}(x+d)d\Phi(x) = P^*.$$

The proof for $R_L(n)$ is exactly similar.

4.2. F Absolutely Continuous Unknown Case

In this section we consider the case where F is absolutely continuous but otherwise unknown. Let $X_{i\alpha}$, $\alpha=1,2,\ldots,n$, $i=1,2,\ldots,p$ be independent random variables such that their joint distribution is $\prod_{i=1}^{p}\prod_{\alpha=1}^{n} F(x_{i\alpha}-\theta_i)$ where F is asbolutely continuous but otherwise unknown, and $\theta = (\theta_1,\ldots,\theta_p)$ as before is an unknown vector in R^p.

Let us denote the rank sum of $\{X_{j1},X_{j2},\ldots,X_{jn}\}$ when they are compared with $\{X_{i1},X_{i2},\ldots,X_{in}\}$ by $R_i^{(j)}$.

Let

$$D_{(1)}^{(ji)} < D_{(2)}^{(ji)} < \ldots < D_{(n^2)}^{(ji)}$$

denote the n^2 ordered differences $X_{i\alpha}-X_{j\beta}$, α, $\beta=1,2,\ldots,n$.

Since F is absolutely continuous we can assume these differences to be distinct. Let $W_{ji}(\Delta)$ denote the number of pairs (α,β) for which

$$X_{j\beta} < X_{i\alpha} - \Delta.$$

In accordance with tradition, we write $W_{ji}(0)$ simply as W_{ji}.

The following theorem, stated as Theorem 2.4 in Lehman (1975), gives the relationship between $D_{(\ell)}^{(ji)}$ and $W_{ji}(\Delta)$.

Theorem 4.2.1. (Lehmann, 1975). Suppose the difference $X_{i\alpha} - X_{j\beta}$ are distinct. Then for any integer m between 1 and n^2 and any real number Δ,

$$(4.2.1) \qquad D_{(m)}^{(ji)} \leq \Delta \quad \text{if and only if} \quad W_{ji}(\Delta) \leq n^2 - m$$

and

$$(4.2.2) \qquad D_{(m)}^{(ji)} > \Delta \quad \text{if and only if} \quad W_{ji}(\Delta) \geq n^2 - m + 1.$$

Gupta's procedure $R_N(n)$ for the case where F is the normal distribution with unknown variance is as follows:

Select i if and only if $\overline{X}_i > \overline{X}_j - d_n S/n^{1/2}$ for all j, $j \neq i$

where $\overline{X}_i$, i=1,...,p are the sample means, S is the pooled estimate of the standard deviation and d_n is just large enough such that $\inf_{\theta \varepsilon R^p} P_\theta\{CS|R_N(n)\} \geq P^*$, P^* pre-determined. Notice that this procedure is equivalent to the following:

Select i if and only if the 100P*% simultaneous confidence intervals $\{\theta_i - \theta_j < \overline{X}_i - \overline{X}_j + d_n S/n^{1/2}$ for all j, $j \neq i\}$ cover $\underset{\sim}{0} = (0,\ldots,0)$.

Theorem 4.2.1. enables us to construct nonparametric simultaneous confidence intervals for $(\theta_i - \theta_j, j \neq i)$ as follows:

$$(4.2.3) \quad P_{F,\theta}\{\theta_i-\theta_j < D_{(a)}^{(ji)} \text{ for all } j,\ j\neq i\}$$

$$= P_{F,\theta_o}\{0 < D_{(a)}^{(ji)} \text{ for all } j,\ j\neq i\}$$

$$= P_{\theta_o}\{n^2-a < W_{ji} \text{ for all } j,\ j\neq i\}$$

$$= P_{\theta_o}\{n^2-a < R_i^{(j)} -1/2\ n(n+1) \text{ for all } j,\ j\neq i\}$$

where θ_o is such that $\theta_1 = \theta_2 = \ldots = \theta_p$. Hence (4.2.3) can be computed exactly by enumerations.

We are now in a position to propose procedure $R_R(n)$, the nonparametric analog of Gupta's procedure $R_N(n)$ based on the Hodges-Lehmann estimator.

<u>Procedure $R_R(n)$ Based on the Hodges-Lehmann Estimator</u>

Select i if and only if $0 < D_{(a_n)}^{(ji)}$ for all j, $j\neq i$

or, equivalently,

Select i if and only if $n^2 - a_n < R_i^{(j)} -1/2\ n(n+1)$

for all j, $j\neq i$

where a_n is the smallest integer such that

$$P_{\theta_o}\{n^2-a_n < R_i^{(j)}-1/2\ n(n+1) \text{ for all } j,\ j\neq i\} \geq P^*.$$

We now show $\inf_{\theta\varepsilon R^p} P_{F,\theta}\{CS|R_R(n)\} \geq P^*$ for <u>any</u> sample size n. Suppose without loss of generality $\theta_i = \max_{j=1,\ldots,p} \theta_j$. Then

$$\inf_{\theta\varepsilon R^p} P_{F,\theta}\{CS|R_R(n)\} = \inf_{\theta\varepsilon R^p} P_{F,\theta}\{0<D_{(a_n)}^{(ji)} \text{ for all } j,\ j\neq i\}$$

$$= \inf_{\theta\varepsilon R^p} P_{F,\theta_o}\{\theta_i-\theta_j<D_{(a_n)}^{(ji)} \text{ for all } j,\ j\neq i\}$$

$$= P_{F,\theta_o}\{0<D_{(a_n)}^{(ji)} \text{ for all } j,\ j\neq i\}$$

$$= P_{\theta_o}\{n^2-a_n<R_i^{(j)}-1/2\ n(n+1) \text{ for all } j,\ j\neq i\}$$

$$\geq P^*.$$

The asymptotic value of a_n is given by the following theorem.

Theorem 4.2.2.

$\lim_{n\to\infty} (n^2/2-a_n)/(n^2(2n+1)/12)^{1/2} = -d/\sqrt{2}$ where d as before is determined by

$$\int\Phi^{p-1}(x+d)\,d\Phi(x) = P^* .$$

Proof.

$$P_{\theta_o}\{n^2-a_n < W_{ji} \text{ for all } j,\ j\neq i\}$$

$$= P_{\theta_o}\{(n^2/2 - a_n)/(n^2(2n+1)/12)^{1/2} < (W_{ji} - n^2/2)/$$

$$(n^2(2n+1)/12)^{1/2} \text{ for all } j,\ j\neq i\}.$$

It is well known that the random vector $((W_{ji}-n^2/2)/(n^2(2n+1)/12)^{1/2},\ j\neq i)$ under θ_o is asymptotically distributed as $N(\underset{\sim}{0},1/2(\underset{\sim}{I}+\underset{\sim}{1}'\underset{\sim}{1}))$. Hence $\lim_{n\to\infty} (n^2/2-a_n)/(n^2(2n+1)/12)^{1/2} = -d/\sqrt{2}$.

The following generalization of Lemma 4 of Lehmann (1963b) enables us to compute the asymptotic relative efficiency of procedure $R_R(n)$.

Lemma 4.2.1. Under the condition $\int f^2(x)dx<\infty$, for fixed i, as $n\to\infty$ the random vector $n^{1/2}(D_{(a_n)}^{(ji)}-(\theta_i-\theta_j),j\neq i)$ has a multivariate normal distribution with mean vector $(d/(12)^{1/2}\int f^2(x)dx)\ \underset{\sim}{1}$ and variance-covariance matrix

$$\sigma_{ii} = 1/6[\int f^2(x)dx]^2$$

$$\sigma_{ij} = 1/12[\int f^2(x)dx]^2$$

Proof. Assume without loss of generality that $\theta_i-\theta_j=0$ for all j, $j\neq i$. For any constant vector $\underset{\sim}{v} = (v_j, j\neq i)$,

$$P\{v_j<n^{1/2}D_{(a_n)}^{(ji)}, j\neq i\} = P\{n^2-a_n<W_{ji}(n^{-1/2}v_j), j\neq i\}$$

by Theorem 4.2.1. By Lemma 1 of Lehmann (1963a), the random vector $n^{-3/2}(W_{ji}(n^{-1/2}v_j)-E[W_{ji}(n^{-1/2}v_j)], j\neq i)$ is asymptotically normal with mean $\underset{\sim}{0}$ and variances equal to 1/6 and covariances equal to 1/12. Now for any j, $j\neq i$

$$\lim_{n\to\infty} n^{-3/2}\{E[W_{ji}(n^{-1/2}v_j)]-n^2/2\}$$

$$= \lim_{n\to\infty} n^{-3/2}\, n^2[P\{X_j-X_i > n^{-1/2}v_j\}-1/2]$$

$$= \lim_{n\to\infty} n^{1/2} \int[F(x-n^{-1/2}v_j)-F(x)]dF(x)$$

$$= -v_j\int f^2(x)dx$$

where the last step in justified because $\int f^2(x)dx<\infty$. See Olshen (1967) and Mehra and Sarangi (1967). Hence the random vector $n^{-3/2}\ ((W_{ji}(n^{-1/2}v_j)-n^2/2), j\neq i)$ is asymptotically normal with mean $-\int f^2(x)dx\ \underset{\sim}{v}$ and variances 1/6 and covariances 1/12. By Theorem 4.2.2., we have

$$\lim_{n\to\infty} P\{v_j < n^{1/2}D_{(a_n)}^{(ji)}, j\neq i\}$$

$$= \lim_{n\to\infty} P\{n^2-n^2/2-(n^2(2n+1)/12)^{1/2}d/\sqrt{2}+o(n^{3/2}) <$$

$$W_{ji}(n^{-1/2}v_j), j\neq i\}$$

$$= \lim_{n\to\infty} P\{-d+[(12)^{1/2}\int f^2(x)dx]v_j+o(n^{3/2})/n^{3/2} <$$

$$(12)^{1/2}n^{-3/2}[W_{ji}(n^{-1/2}v_j)-n^2/2]$$

$$+ [(12)^{1/2}\int f^2(x)dx]v_j, j\neq i\}$$

By taking limit we get the desired result.

Asymptotic Relative Efficiency of Procedure $R_R(n)$

Consider Gupta's normal means procedure $R_N(n)$ described earlier:

Select i if and only if $\overline{X}_i - \overline{X}_j + d_n S/n^{1/2} > 0$

for all j, $j \neq i$

where $\overline{X}_i$, $i=1,\ldots,p$ are the sample means, S is the pooled estimate of the common standard deviation, and d_n is determined by

$$\int_0^\infty \int_{-\infty}^\infty \Phi^{p-1}(x+sd_n)\,d\Phi(x)\,dG_v(s)$$

where G_v is the cdf of χ_v/v with $v = p(n-1)$.
Denote by σ^2 the variance of F. We shall assume that $\sigma^2<\infty$. Under the additional assumption $\int f^2(x)dx<\infty$, we see from Lemma 4.2.1. that for any i, $n^{1/2}[(12)^{1/2}\int f^2(x)dx]\,(D^{(ji)}_{(a_n)} - (\theta_i-\theta_j),\ j\neq i)$ and $n^{1/2}\sigma^{-1}(\overline{X}_i-\overline{X}_j + d_n S/n^{1/2},\ j\neq i)$ have the same limiting distribution. Hence if $n'(n)$ is such that $\lim_{n\to\infty} n'(n)/n = 1/12\sigma^2[\int f^2(x)dx]^2$, then $\lim_{n\to\infty} P_{F,\theta}\{CS|R_R(n')\}/P_{F,\theta}\{CS|R_N(n)\} = 1$ and $\lim_{n\to\infty} E_{F,\theta}[S|R_R(n')]/E_{F,\theta}[S|R_N(n)] = 1$ for any F,θ. Therefore if we define asymptotic relative efficiency $e_{R,N}$ of R_R to R_N as the limit of the reciprocal of the sample sizes required such that the two procedures have the same asymptotic performance, where performance is measured in terms of $E(S|R)$ for controlled $P\{CS|R\}$, $P\{CS|R\}$ for controlled $E(S|R)$, or a linear combination of $P\{CS|R\}$ and $E(S|R)$, then $e_{R,N} = 12\sigma^2[\int f^2(x)dx]^2$. It is well known that $12\sigma^2[\int f^2(x)dx]^2 \geq 0.864$ for all f, and $12\sigma^2[\int f^2(x)dx]^2 = 3/\pi \approx 0.955$ when f is the normal density.

BIBLIOGRAPHY

BIBLIOGRAPHY

Bickel, P.J. (1967). Some contributions to the theory of order statistics. Proc. Fifth Berkeley Symp. Math. Stat. Prob. 1:575-591.

Bickel, P.J. and Yahav, J.A. (1977). On selecting a set of good populations. Statistical Decision Theory and Related Topics II. (ed. Shanti S. Gupta and David S. Moore), 37-55. Academic Press, N.Y.

Brown, L.D. (1974). Notes on statistical decision theory, revised.

Brown, L.D., Cohen, A. and Strawderman, W.E.A. (1976). A complete class theorem for strict monotone likelihood ratio with applications. Ann. Statist. 4:712-722.

Chernoff, H. and Yahav, J.A. (1977). A subset selection problem employing a new criterion. Statistical Theory and Related Topics II. (ed. Shanti S. Gupta and David S. Moore), 93-119. Academic Press, N.Y.

Eaton, M.L. (1967). Some optimum properties of ranking procedures. Ann. Math. Statist. 38:124-137.

Goel, P.K. and Rubin, H. (1975). On selecting a subset containing the best population--A Bayesian Approach. Mimeograph Series No. 432, Dept. of Statistics, Purdue University, W. Laf., IN.

Gupta, S.S. (1956). On a decision rule for a problem in ranking means. Mimeograph Series No. 150, Institute of Statistics, University of North Carolina, Chapel Hill, N.C.

Gupta, S.S. (1965). On some multiple decision (selection and ranking) rules. Technometrics 7:225-245.

Gupta, S.S. and Huang, D.Y. (1976). On some methods for constructing optimal subset selection procedures. Mimeograph Series #470, Dept. of Statistics, Purdue University, W. Laf., IN.

Gupta, S.S. and Nagel, K. (1971). On some contributions to multiple decision theory. In Statistical Decision Theory and Related Topics (ed. Shanti S. Gupta and James Yackel), 79-102, Academic Press, N.Y.

Hall, I.J. and Kudo, A. (1968). On slippage tests-(I) A generalization of Neyman-Pearson's lemma. Ann. Math. Statist. 39:1693-1699.

Hodges, J.L. and Lehmann, E.L. (1956). The efficiency of some nonparametric competitors of the t-test. Ann. Math. Statist. 27:324-335.

Huber, P.J. (1964). Robust estimation of a location parameter. Ann. Math. Statist. 35:73-101.

Jackel, L.A. (1971). Robust estimates of location: Symmetric and asymmetric contamination. Ann Math. Statist. 42:1020-1034.

Karlin, S. and Rubin, H. (1956). The theory of decision procedures for distributions with monotone likelihood ratio. Ann. Math. Statist. 27:272-299.

Karlin, S. and Truax, D. (1960). Slippage problems. Ann. Math. Statist. 31:296-394.

Lehmann, E.L. (1955). Ordered families of distributions. Ann. Math. Statist. 26:399-419.

Lehmann, E.L. (1963a). Robust estimation in analysis of variance. Ann. Math. Statist. 34:957-966.

Lehmann, E.L. (1963b). Nonparametric confidence intervals for a shift parameter. Ann. Math. Statist. 34:1507-1512.

Lehmann, E.L. (1963c). A class of selection procedures based on ranks. Math. Annalen. 150:268-275.

Lehmann, E.L. (1966). On a theorem of Bahadur and Goodman. Ann. Math. Statist. 37:1-6.

Lehmann, E.L. (1975). Nonparametrics: Statistical Methods Based on Ranks. Holden-Day, San Francisco.

McDonald, G.C. (1969). On some distribution-free ranking and selection procedures. Mimeograph Series No. 174, Dept. of Statistics, Purdue University, W. Laf., IN.

Marshall, A.W. and Olkin, I. (1974). Majorization in multivariate distributions. Ann. Statist. 2:1189-1200.

Mehra, K.L. and Sarangi, J. (1967). Asymptptic efficiencies of certain rank tests for comparative experiments. Ann. Math. Statist. 38:90-107.

Nagel, K. (1970). On subset selection rules with certain optimality properties, Mimeograph Series No. 222, Dept. of Statistics, Purdue University, West Lafayette, IN.

Olshen, R.A. (1967). Sign and Wilcoxon test for linearity. Ann. Math. Statist. 38:1759-1769.

Oosterhoff, J. (1969). Combination of One-Sided Statistical Tests, Mathematical Centre Tracts. Amsterdam.

Paulson, E. (1949). A multiple decision procedure for certain problems in the analysis of variance. Ann. Math. Statist. 20:95-98.

Pratt, J. (1956). Some results in the decision theory of one-parameter multivariate Polya type distributions. Ph.D. Thesis, Dept. of Statistics, Stanford University, Stanford, CA.

Puri, M.L. and Puri, P.S. (1969). Multiple decision procedures based on ranks for certain problems in analysis of variance. Ann. Math. Statist. 40:619-632.

Puri, P.S. and Puri, M.L. (1968). Selection procedures based on ranks: Scale parameter case. Sankhya Ser. A 30:291-302.

Rizvi, M.H. and Woodworth, G.G. (1970). On selection procedures based on ranks: countersamples concerning the least favorable configurations. Ann. Math. Statist. 41:1942-1951.

Seal, K.C. (1955). On a class of decision procedures for ranking means of normal populations. Ann. Math. Statist. 26:387-398.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: Mimeograph Series #491	2. GOVT ACCESSION NO.
	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle): ON SOME DECISION-THEORETIC CONTRIBUTIONS TO THE PROBLEM OF SUBSET SELECTION	5. TYPE OF REPORT & PERIOD COVERED: Technical
	6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s): Jason C. Hsu	8. CONTRACT OR GRANT NUMBER(s): N0014-75-C-0455
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Department of Statistics, Purdue University, West Lafayette, IN 47907	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: NR 042-243
11. CONTROLLING OFFICE NAME AND ADDRESS: Office of Naval Research, Washington, DC	12. REPORT DATE: June 1977
	13. NUMBER OF PAGES: 73
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, decision-theoretic, Bayes procedure, essentially complete class, multivariate monotone likelihood ratio, monotone procedures, Monte Carlo, robust subset selection procedures, nonparametric subset selection procedures.

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Ranking and selection procedures, subset selection procedures in particular, are procedures that provide in a realistic manner attractive ways of handling problems that are commonly treated by the 2-action procedure of a global F-test, and the many-action procedure of a typical multiple range test. Consider the usual one-way layout situation in analysis of variance. Usually the experimenter wants to know more than just whether all the treatment effects are equal, but he may not want to make inferences concerning all

DD FORM 1473, 1 JAN 73 EDITION OF 1 NOV 65 IS OBSOLETE

S/N 0102-014-6601

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered) linear contrasts of means. One of the more frequently occuring situations for which this is so is where the experimenter simply wishes to know which of the treatments gives the best product. In this situation, formulating the problem as a selection problem is appropriate. Subset selection procedures are often thought of as screening procedures. If the data indicates several treatments are better than the remaining treatments but no treatment is clearly the best, then perhaps the experimenter ought to retain all of the better treatments for future considerations.

It is generally recognized that for multivariate problems uniformly best procedures do not exist. Hence it is reasonable to look for procedures that do well on the average, averaged over the parameter space by some prior. This approach has been taken in the first two chapters. The essentially complete class of Bayes procedures and their limits is investigated. The concept of Total Monotone Likelihood Ratio is introduced as the multivariate analog of univariate monotone likelihood ratio. Then a multivariate analog of the classical result of Karlin and Rubin (1956), that monotone procedures form an essentially complete class, is proved for a loss function which seems natural to the subset selection problem by proving that Bayes procedures are monotone.

Bayes procedures typically require numerical integrations to implement and this makes them sometimes unsuitable for practical use. Besides, the use of Bayes procedures is by no means universally accepted. So if there is available an easy to implement procedure whose performance is close to that of the Bayes procedure, then this procedure ought to be used. This possibility is explored in Chapter 3 for the case of normal populations problem and normal exchangeable priors. As it turns out, for each prior and loss function pair there is always a Gupta type procedure that performs almost as well as the Bayes procedure, while this being true for Seal type procedures only when the normal prior is very informative. As of yet we do not know how these procedures perform when the prior is not normal. Nevertheless we recommend the use of Gupta type procedures when the observations arise from normal distributions as procedures that have at least some near optimality properties.

In the case where the parameter of interest is a location parameter and the underlying distribution is not entirely known there are good robust estimators of the parameter. Under mild regularity conditions they are asymptotically normal. From the results of Chapter 3, we would expect that Gupta type procedures based on these estimators to have good asymptotic performance. In Chapter 4 robust and nonparametric Gupta type procedures are proposed. One procedure in particular, the procedure based on simultaneous confidence bounds derived from rank tests, is nonparmatric. It controls the probability of a correct selection for any sample size. Since it is based essentially on the Hodges-Lehmann estimator, it inherits the high asymptotic relative efficiency of that estimator.