BAYESIAN SUBSET SELECTION FOR ADDITIVE AND LINEAR LOSS FUNCTIONS*

by

Klaus-J. Miescke

University of Mainz and Purdue University

Department of Statistics
Division of Mathematical Sciences
Mimeograph Series #78-23

October 1978

* Research supported partly by the Office of Naval Research contract N00014-75-C-0455 at Purdue University. Reproduction in whole or in part is permitted for any purpose of the United States Government.

BAYESIAN SUBSET SELECTION FOR ADDITIVE AND LINEAR LOSS FUNCTIONS*

by

Klaus-J. Miescke

University of Mainz and Purdue University

1. Introduction and Summary.

Let $\{P_\theta\}_{\theta\in\Omega}$, $\Omega \subseteq \mathbb{R}$ be a given family of probability distributions over $(\mathbb{R}, \mathcal{B})$, where $\mathcal{B}$ denotes the corresponding Borel-sets, and let $\underline{X}_i = (X_{i1},\ldots,X_{in})$, $i = 1,\ldots,k$, be k independent samples of common size n from k populations π_i with distributions P_{θ_i}, $i = 1,\ldots,k$, where $\theta_1,\ldots,\theta_k \in \Omega$ are unknown. If our goal is to select a random non-empty subset $S(\underline{X}_1,\ldots,\underline{X}_k)$ of populations which is "good" (i.e. is associated with large θ-values in a certain sense), then there are several possible choices for a suitable criterion to make the term "good" more precise. One of these - a Bayes criterion with additive and especially linear loss functions - will be the topic of this paper.

For simplicity we assume that the rule S take values in $\mathcal{G} = \{s | s \subseteq \{1,\ldots,k\}, s \neq \phi\}$, where $S = s$ now means that all π_i are selected for which $i \in s$, $i = 1,\ldots,k$. Moreover for every $s \in \mathcal{G}$ let $|s|$ denote the size of s.

For a given loss function $L: \Omega^k \times \mathcal{G} \to \mathbb{R}$ and a given prior distribution τ over $(\Omega^k, \mathcal{B}_k)$ (for the now random parameter vector $\underline{\Theta}$ with values $\underline{\theta} \in \Omega^k$) every Bayes solution S* minimizes $E_\tau\, E_{\underline{\theta}}\, L(\underline{\Theta}, S(\underline{X}_1,\ldots,\underline{X}_k))$ among all $S: \mathbb{R}^{nk} \to \mathcal{G}$ by definition. (Here and later on we tacitly assume that L is intergrable properly.) The well known standard method to find the Bayes rules S* is to minimize, after observing $\underline{X}_i = \underline{x}_i$, $i = 1,\ldots,k$, the posterior expected loss $E(L(\underline{\Theta},s) | \underline{X}_i = \underline{x}_i,\ i = 1,\ldots,k)$ among all $s \in \mathcal{G}$, which always is possible since $\mathcal{G}$ is finite.

* Research supported partly by the Office of Naval Research contract N00014-75-C-0455 at Purdue University. Reproduction in whole or in part is permitted for any purpose of the United States Government.

The first paper within this framework is due to Deely and Gupta (1968) and deals with the "linear" loss function $L(\underline{\theta},s) = \sum_{i\in s} \alpha_{s,i}(\theta_{[k]}-\theta_i)$, where the α's are non-negative and $\theta_{[1]} \leq \cdots \leq \theta_{[k]}$ denote the ordered values of $\theta_1,\ldots,\theta_k$. (This notation will be utilized analogously for all vectors used in the sequel.). The result is that, under an additional assumption on the α's, one can choose a Bayes rule which always selects only one population.

Because of this somewhat undesirable property Goel and Rubin (1977) choose the loss function $L(\underline{\theta},s) = c|s| + \theta_{[k]} - \max_{i\in s} \theta_i$ and study the behavior of the corresponding Bayes rule in great detail. (Because of the complexity of the problem it is necessary to derive approximate solutions.). Other papers dealing with other non-additive loss functions are due to Bickel and Yahav (1977), where the Bayesian aspect, however, is not of primary interest, and due to Chernoff and Yahav (1977), where two rules (one of these is Gupta's means procedure (c.f. (3.12))) are compared with the Bayes rule in a "normal model" on the basis of Monte Carlo results. As a result Gupta's maximum means procedure turns out to be "remarkably efficient" w.r.t. the Bayes rule.

Finally three papers dealing with additive (non-linear) loss functions of the type $L(\underline{\theta},s) = \sum_{i\in s} (c_2-c_1 I_{\{\theta_{[k]}\}}(\theta_i))$ (where I denotes the indicator function) are due to Bratcher and Bhalla (1974) and Gupta and Hsu (1977,78). In the first the Bayes rule is derived and a binomial example is given. In the second similar Monte Carlo-studies are performed as in Chernoff and Yahav (1977), and again Gupta's maximum means procedures "do almost as well as the Bayes procedure". And in the last monotonicity of Bayes rules is the topic.

This paper serves two purposes: First (in Section 2) we show that the result of Deely and Gupta (1968) is not due to the linearity of the loss function but is due to the combined effect of the additivity of the loss function and the non-negativity of its terms. Then we study the case of additive and especially linear loss functions, thereby filling a gap lying in between the paper by Deely and Gupta (1968) and the others mentioned above. Then (in Section 3) we show that in the normal case with symmetric normal priors Gupta's maximum means procedure turns out to be asymptotically Bayes w.r.t. a class of additive loss functions, whereas on the other hand Seal's procedure turns out to be Bayes w.r.t. an unrealistic additive loss function. Finally (in the appendix) we derive some bounds for $E(\max_{j=1,\ldots,k}(\mu_j + \rho V_j))$ (where $\underline{\mu} \in \mathbb{R}^k$, $\rho \in \mathbb{R}$ are fixed known and $\underline{V} \sim N(\underline{0},I)$) to approximate the Bayes rules w.r.t. linear loss functions in cases where n is finite.

2. Additive and Linear Loss Functions.

As indicated above we first discuss the result of Deely and Gupta (1968). Let us consider the loss function $L(\underline{\theta},s) = \sum_{i\in s} \alpha_{s,i}\ell_i(\underline{\theta})$ and assume for simplicity that $\alpha_{s,i} = \alpha(|s|)$, $i = 1,\ldots,k$, $s \in G$, holds. (In fact our following result remains valid if we replace this assumption by the corresponding one of Deely and Gupta (1968).). Then we can state the following slight generalization of the theorem in Deely and Gupta (1968):

Theorem 1. Let $m\,\alpha(m) \geq \alpha(1)$, $m = 1,\ldots,k$. If the ℓ_i's are non-negative then there exists a Bayes rule which always selects exactly one population.

Proof: Given $\underline{X} = \underline{x}$ the aposteriori risk of any procedure S is given by $R(S|\underline{x}) = \alpha(|S(\underline{x})|) \sum_{i\in S(\underline{x})} A_i(\underline{x})$, where $A_i(\underline{x}) = E(\ell_i(\underline{\Theta}), \underline{x}) \geq 0$, $i = 1,\ldots,k$.

Thus
$$R(S|\underline{x}) \geq \alpha(|S(\underline{x})|)\,|S(\underline{x})| \min_{i=1,\ldots,k} A_i(\underline{x})$$
$$\geq \alpha(1) \min_{i=1,\ldots,k} A_i(\underline{x}) = \min_{i=1,\ldots,k} R(\{i\}|\underline{x}),$$

which completes the proof.

Example 1. That the converse statement does not hold true can be demonstrated by the following example: Let

$$L(\underline{\theta},s) = |s|^{-1} \sum_{i\in s} (\theta_{[k]} - \theta_i - \varepsilon)$$

$$= \theta_{[k]} - |s|^{-1} \sum_{i\in s} \theta_i - \varepsilon,\ \varepsilon > 0.$$

Clearly $|S(\underline{x})|^{-1} \sum_{i\in S(\underline{x})} E(\theta_i|\underline{x}) \leq \max_{i=1,\ldots,k} E(\theta_i|\underline{x})$ holds, and even strict inequality occurs in many cases.

Definition 1. We call a loss function L additive, if

(2.1) $L(\underline{\theta},s) = \sum_{i\in s} \ell_i(\underline{\theta}),\ \ell_i\colon \Omega^k \to \mathbb{R},\ i = 1,\ldots,k,$

and linear, if

(2.2) $L(\underline{\theta},s) = c \sum_{i\in s} (\theta_{[k]} - \theta_i - \varepsilon)$, where $\varepsilon,\ c > 0$, and c clearly can be put equal to one.

In many situations the following assumptions assuring invariance under permutations and monotonicity of the losses seem to be quite natural:

(2.3) (a) $\ell_i(\underline{\theta}) = \ell(\{\theta_1,\ldots,\theta_k\},\ \theta_i),\quad \underline{\theta} \in \Omega^k.$

(b) $\ell_i(\underline{\theta}) \leq \ell_j(\underline{\theta})$ if $\underline{\theta} \in \Omega^k$ with $\theta_j \leq \theta_i,\ i,j \in \{1,\ldots,k\}.$

(c) $\ell_i(\underline{\theta}) \geq \ell_i(\underline{\theta}')$ if $\underline{\theta},\ \underline{\theta}' \in \Omega^k$ with $\theta_i \leq \theta_i'$ and

$$\theta_j \geq \theta_j',\ j \neq i,\ i = 1,\ldots,k.$$

Obviously these conditions are met by linear loss functions.

The following theorem, which to some extent can be viewed as being a special case of Lehmann's (1957) result, can be stated now without proof:

Theorem 2. *For an additive loss function of type* (2.1) *the subsets* $s(\underline{x}) \in \mathcal{G}$ *which a Bayes rule would select after observing* $\underline{x} \in \mathbb{R}^{nk}$ *satisfy the relation* $\underset{\sim}{S}^*(\underline{x}) \subseteq s(\underline{x}) \subseteq \tilde{S}^*(\underline{x}) \cup \{i\}$ *for at least one* $i \in M(\underline{x})$, *where*

$$\begin{aligned}
\tilde{S}^*(\underline{x}) &= \{j \,|\, E(\ell_j(\underline{\Theta})\,|\,\underline{x}) \le 0\}, \\
\underset{\sim}{S}^*(\underline{x}) &= \{j \,|\, E(\ell_j(\underline{\Theta})\,|\,\underline{x}) < 0\} \quad \underline{\text{and}} \\
M(\underline{x}) &= \{i \,|\, E(\ell_i(\underline{\Theta})\,|\,\underline{x}) = \min_{j=1,\ldots,k} E(\ell_j(\underline{\Theta})\,|\,\underline{x})\}.
\end{aligned} \tag{2.4}$$

Especially, for a linear loss function (2.4) *reads as follows:*

$$\begin{aligned}
\tilde{S}^*(\underline{x}) &= \{j \,|\, E(\Theta_j\,|\,\underline{x}) \ge E(\Theta_{[k]}\,|\,\underline{x}) - \varepsilon\}, \\
\underset{\sim}{S}^*(\underline{x}) &= \{j \,|\, E(\Theta_j\,|\,\underline{x}) > E(\Theta_{[k]}\,|\,\underline{x}) - \varepsilon\} \quad \underline{\text{and}} \\
M(\underline{x}) &= \{i \,|\, E(\Theta_i\,|\,\underline{x}) = \max_{j=1,\ldots,k} E(\Theta_j\,|\,\underline{x})\}.
\end{aligned} \tag{2.5}$$

Remark 1. Another type of loss function which turns out to be equivalent to our type of additive loss function is the following:

$$\hat{L}(\underline{\theta},s) = \sum_{i\in s} \ell_i^+(\underline{\theta}) + \sum_{i\notin s} \ell_i^-(\underline{\theta}), \tag{2.6}$$

where $\ell_i^+, \ell_i^-: \Omega^k \to [0,\infty)$, $i = 1,\ldots,k$, can be viewed as being losses for errors of the first and second kind in analogy to testing theory. Since such a loss function can be rewritten as

$$\hat{L}(\underline{\theta},s) = \sum_{i=1}^{k} \ell_i^-(\underline{\theta}) + \sum_{i\in s} (\ell_i^+(\underline{\theta}) - \ell_i^-(\underline{\theta})), \tag{2.7}$$

we arrive at the same set of Bayes solutions if we drop the first sum on the r.h.s. of (2.7). (Hereby in fact the overall risk is changed only by an additive constant. Note that the change of the conditional risk, given $\underline{X} = \underline{x}$, has no influence on the determination of the Bayes rules.).

Conversely if we start with an additive loss function of type (2.1) we can switch over to a loss function of type (2.6) analogously by choosing

ℓ_i^+ (ℓ_i^-) to be the usual positive (negative) part of ℓ_i, $i = 1,\dots,k$.

Thus Theorem 1 - applied to (2.6) - can be interpreted as follows: if we always have to pay for every population selected, then we take as few as possible: namely always one.

Example 2. Bratcher and Bhalla took a loss function $\hat{L}$ of type (2.6) with $\ell_i^+(\underline{\theta}) = \hat{c}_1(1-I_{\{\theta_{[\kappa]}\}}(\ell_i))$, $\ell_i^-(\underline{\theta}) = \hat{c}_2 I_{\{\theta_{[\kappa]}\}}(\theta_i)$, $\hat{c}_1,\hat{c}_2 > 0$, $i = 1,\dots,k$. On the other hand Gupta and Hsu (1977,78) took the loss function $L(\underline{\theta},s) = c_1(1 - I_{\{\theta_i | i \in s\}}(\theta_{[k]})) + c_2|s|$, c_1, $c_2 > 0$.

If we assume that the posterior distribution of $\underline{\Theta}$ assures that for every $\underline{x} \in \mathbb{R}^{nk}$ all the Θ_i's are distinct with probability one, then $L(\underline{\theta},s)$ can be replaced by

$$L(\underline{\theta},s) = c_1 + \sum_{i\in s} (c_2 - c_1\, I_{\{\theta_{[k]}\}}(\theta_i)).$$

If we take into account that we can multiply a loss function with a positive constant and moreoever can add any further constant to it without changing a given Bayes problem, then it is easy to see that the loss function of Bratcher and Bhalla (1974) and that of Gupta and Hsu (1978) are equivalent in the sense of Remark 1 and, especially, both are additive.

It seems to be worth mentioning that in the normal model (cf. (3.1)) the converse of Theorem 1 also holds true.

If the ℓ_i's mentioned there may also assume negative values (i.e. if $c_2 < c_1$ here), then there exists no Bayes rule which always selects exactly one population. Since aposteriori the Θ_i's now are jointly (non-degenerate) normally distributed, this follows from the fact that the Bayes rule S*, for all $\underline{x} \in \mathbb{R}^{nk}$ except possibly a null set, turns out to be

$$\text{(2.8)}\quad i \in S^*(\underline{x}) \text{ iff } P\{\Theta_i = \Theta_{[k]}|\underline{x}\} = \max_{j=1,\dots,k} P\{\Theta_j = \Theta_{[k]}|\underline{x}\}$$
$$\text{and/or } P\{\Theta_i = \Theta_{[k]}|\underline{x}\} > c_2/c_1.$$

Finally let us consider some other properties of the Bayes rules for additive loss functions. For this purpose let $Z_i = H(\underline{X}_i)$ be a one-dimensional sufficient statistic for θ_i with distribution Q_{θ_i}, say, $i = 1,\dots,k$. The rules S of course depend now on $Z_1,\dots,Z_k$ and are defined on $\mathbb{R}^k$.

Definition 2. We call a rule $S\colon \mathbb{R}^k \to \mathbb{G}$ *ordered* if for every

$\underline{z} \in \mathbb{R}^k$ $i \in S(\underline{z})$ and $z_i < z_j$ implies $j \in S(\underline{z})$.

We call it *monotone* if for every $i \in \{1,\dots,k\}$ and $\underline{z}, \underline{z}' \in \mathbb{R}^k$ with $z_i \le z_i'$ and $z_j \ge z_j'$, $j \ne i$, $i \in S(\underline{z})$ implies $i \in S(\underline{z}')$.

Since we only need to consider non-randomized rules in this Bayesian framework (cf. Goel and Rubin (1977)), Definition 2 for monotonicity and Definition 8 in Gupta and Hsu (1977) coincide. Now we can state:

Theorem 3. *Let* $\{Q_\theta\}_{\theta \in \Omega}$ *have densities* $\{f_\theta\}_{\theta \in \Omega}$ *w.r.t. the Lebesgue measure on* $(\mathbb{R},\mathbb{B})$ *with monotone non-decreasing likelihood ratios in* θ, *and let* L *be any additive loss function. Then the following statements hold true:*

(i) *If* L *satisfies* (2.3) (a) *and* (b) *and if the prior distribution of* $\underline{\theta}$ *is symmetric on* Ω^k, *then every Bayes rule can be assumed to be ordered.*

(ii) *If* L *satisfies* (2.3) (c) *and if*

$\Theta_1,\dots,\Theta_k$ *are apriori independently distributed, then every Bayes rule can be assumed to be monotone.*

Proof: The first part follows from Goel and Rubin (1977) (cf. Lemma 1 and Remark 1 cited there), since our loss function satisfies the conditions stated there.

The second part is a generalization of the Theorem 3 of Gupta and Hsu (1977). It is easy to see that their proof works with every loss function whose components $\ell_1,\dots,\ell_k$ satisfy the conditions stated in (II).

Note that additive loss functions of type (2.1) with $\ell_i(\underline{\theta}) = \ell(\theta_{[k]}-\theta_i)$, $i = 1,\dots,k$, for non-decreasing ℓ and thus in particular linear loss functions satisfy all the conditions stated in the theorem above. Thus we can state:

Corollary 1. *Let* $\{Q_\theta\}_{\theta \in \Omega}$ *be given as in Theorem 3 and let*

$$L(\underline{\theta},s) = \sum_{i \in s} \ell(\theta_{[k]} - \theta_i) \text{ with non-decreasing}$$

ℓ: $\mathbb{R} \to \mathbb{R}$. *Then, if* $\Theta_1,\dots,\Theta_k$ *are apriori independently identically distributed, every Bayes rule can be assumed to be ordered and monotone.*

Remark 2. In their Lemma 2 Goel and Rubin (1977) give an aid with which one can simplify the computation of the Bayes procedure if one adopts their loss function and assumes a symmetric prior distribution:

If $\underline{Z} = \underline{z} \in \mathbb{R}^k$ is observed and $\underline{z}$ is, without loss of generality, ordered in such a way that $z_1 \leq \dots \leq z_k$ holds, then the terms $r_i(\underline{z}) = E(L(\underline{\Theta},\{k,k-1,\dots,k-i+1\})|\underline{z})$, $i = 1,\dots,k$, have the property that $r_{j+1}(\underline{z}) - r_j(\underline{z})$ is non-decreasing in $j = 1,2,\dots,k-1$.

For additive loss functions this property reduces just to

$$(2.9) \qquad E(\ell_1(\underline{\theta})|\underline{z}) \geq \dots \geq E(\ell_k(\underline{\Theta})|\underline{z}),\ z_1 \leq \dots \leq z_k.$$

Though it is difficult to find general sufficient conditions for (2.9) to hold true, we can at least state the following:

Theorem 4. *Let* $\{Q_\theta\}_{\theta \in \Omega}$ *and* L *be given as in Corollary 1. Then each of the two conditions stated below is sufficient for (2.9) to hold true:*

(A.1) $\Theta_1,\ldots,\Theta_k$ <u>are apriori independently identically distributed.</u>

(A.2) L <u>is linear and the prior distribution</u> τ <u>is symmetric on</u> Ω^k.

<u>Proof</u>: Under (A.1) note that, given $\underline{Z} = \underline{z}$ with $z_1 \leq \ldots \leq z_k$, the Θ_i's are independent and stochastically ordered in the same order as the z_i's. With standard analysis one can show that this implies that even the dependent variables $\Theta_1 - \Theta_{[k]},\ldots,\Theta_k - \Theta_{[k]}$ are stochastically ordered in the same direction. Thus by the monotonicity of ℓ the assertion follows.

Under (A.2) for every $\underline{z} \in \mathrm{IR}^k$ with $z_1 \leq \ldots \leq z_k$ and $i \in \{1,\ldots,k-1\}$ we have

$$E(\ell_i(\underline{\Theta})|\underline{z}) - E(\ell_{i+1}(\underline{\Theta})|\underline{z}) = E(\Theta_{i+1} - \Theta_i|\underline{z}) = b \int_{\Omega^k} (\theta_{i+1} - \theta_i) \prod_{r=1}^{k} f_{\theta_r}(z_r) d\tau(\underline{\theta})$$

$$= b \int_{\{\theta_i<\theta_{i+1}\}} (\theta_{i+1}-\theta_i)[f_{\theta_i}(z_i)f_{\theta_{i+1}}(z_{i+1})-f_{\theta_i}(z_{i+1})f_{\theta_{i+1}}(z_i)] \prod_{r\neq i,i+1} f_{\theta_r}(z_r)d\tau(\underline{\theta}),$$

where b is a normalizing factor and the last identity follows by the method which Goel and Rubin used to prove their Lemma 2. Thus by the M.L.R. property of $\{f_\theta\}_{\theta \in \Omega}$ the proof is completed.

3. <u>The Normal Model</u>.

In this section we assume that apriori $\underline{X}_1,\ldots,\underline{X}_k$ are independent samples of common size n from k normal populations with unknown means $\theta_1,\ldots,\theta_k$ and a common known variance $\sigma^2 > 0$. By sufficiency we can reduce our set of data to $\underline{X} = (\bar{X}_1,\ldots,\bar{X}_k)$, where $\bar{X}_1,\ldots,\bar{X}_k$ are the corresponding sample means. As to the Θ's we assume an exchangeable normal prior. More precisely our "<u>normal model</u>" is as follows:

(3.1) (a) $\underline{X}|\underline{\Theta} = \underline{\theta} \sim N(\underline{\theta}, q\ I)$ and

(b) $\underline{\Theta} \sim N(m\ \underline{1},\ r\ I + t\ U)$, where

$q = \sigma^2/n$, $m \in \mathrm{IR}$, $r > 0$, $t > -r/k$, $\underline{1} = (1,\ldots,1)'$,

$U = \underline{1}\ \underline{1}'$ and 1 denotes the k×k identity matrix.

Note that $r > 0$ together with $t > -r/k$ is necessary and sufficient for

$r\,I + t\,U$ to be positive definite. This model was chosen by Chernoff and Yahav (1977) (with $t > 0$) and by Gupta and Hsu (1978).

By (3.1) it is easy to see that we have

(3.2) (a) $\underline{\theta}|\underline{X} = \underline{x} \sim N(\underline{\mu},\ a\,I + b\,U)$, where

$\underline{\mu} = r(q+r)^{-1}\,\underline{x} + q\ t((q+r)(q+r+kt))^{-1}\,U\underline{x} + q(q+r+kt)^{-1}\,m\,\underline{1}$,

$a = r\,q\,(q+r)^{-1}$ and $b = q^2 t((q+r)(q+r+kt))^{-1}$, and

(b) $\underline{X} \sim N(\underline{\tilde{\mu}},\ \tilde{a}\,I + \tilde{b}\,U)$, where

$\underline{\tilde{\mu}} = m\,\underline{1}$, $\tilde{a} = q+r$ and $\tilde{b} = t$.

Let us include non-additive loss functions $L(\underline{\theta},s)$ into our next considerations. Then one intuitively feels that one can put m and t equal to zero without changing the problem, if L is translation-invariant, i.e. if $L(\underline{\theta},s) = L(\underline{\theta} + \eta\,\underline{1},\ s)$ holds for all $\underline{\theta} \in \Omega^k$, $s \in \mathcal{A}$ and $\eta \in \mathbb{R}$. Moreover one should expect that the Bayes rules then are translation-invariant, too: $S^*(\underline{x}) = S^*(\underline{x} + \eta\,\underline{1})$ for all $\eta \in \mathbb{R}$ and all $\underline{x} \in \mathbb{R}^k$ except possibly a null set.

Ideas of this kind primarily are due to Chernoff and Yahav (1977) and also to Gupta and Hsu (1977). Nevertheless since the formulation and proof of a general theorem in this direction is missing up to now we feel that it is justified to do this in our present paper.

Now if any random vector $\underline{Y}$ is distributed according to some $N(\underline{\mu}, a\,I + b\,U)$ with $a > 0$ and $b > -a/k$, then at once one has in mind the following ("conditional i.i.d.") representation:

(3.3) $\underline{Y} = a^{1/2}\,\underline{V} + b^{1/2}\,W\,\underline{1} + \underline{\mu}$ where

$\underline{V} \sim N(\underline{0},I)$ and $W \sim N(0,1)$ are independent.

But this holds only in cases where $b \geq 0$ (i.e. where the correlations of the Y's are non-negative), and therefore does not help us in our more general setup. But, fortunately, there exists another representation, which always can be used for our purposes:

Lemma 1. *Let* $\underline{\mu} \in \mathbb{R}^k$, $a > 0$ *and* $b > -a/k$. *If*

(3.4) $\quad \underline{Y} = a^{1/2}\underline{V} + k^{-1}((a+k\ b)^{1/2} - a^{1/2})U\ \underline{V} + \underline{\mu},$

with $\underline{V} \sim N(\underline{0}, I)$,

then $\underline{Y} \sim N(\underline{\mu}, a\ I + b\ U)$.

The proof is standard and therefore omitted. Besides we remark that in (3.4) $(a + k\ b)^{1/2}$ can be replaced by $-(a + k\ b)^{1/2}$.

Theorem 5. *Under the normal model (3.1) for every loss function* $L(\underline{\theta},s)$ *which is translation-invariant the following three assertions hold true:*

(i) *For every rule* $S: \mathbb{R}^k \to \mathcal{A}$ *and every* $\underline{x} \in \mathbb{R}^k$ *the posterior risk* $E(L(\underline{\theta},S(\underline{x}))|\underline{X} = \underline{x})$ *does not depend on* m *and* t.

(ii) *Every Bayes rule can be assumed to be translation-invariant.*

(iii) *For every translation-invariant rule* S *the overall risk* $E(L(\underline{\theta},S(\underline{X})))$ *does not depend on* m *and* t.

Proof: Under the normal model (3.1) let $L(\underline{\theta},s)$ be translation-invariant.

(i) Given $\underline{X} = \underline{x}$, $\underline{\theta}$ is distributed according to $N(\underline{\mu}, a\ I + b\ U)$ where $\underline{\mu}$, a and b are given by (3.2) (a). In this situation we choose for $\underline{\theta}$ the representation given by Lemma 1.

Let S be any rule, $S(\underline{x}) = s$, say, and $\rho \in \mathbb{R}$. Then since $U\ \underline{y} = (y_1+\ldots+y_k)\ \underline{1}$ for every $\underline{y} \in \mathbb{R}^k$, we get by the translation-invariance of $L(\underline{\theta},s)$

(3.5) $\quad E(L(\underline{\Theta},s)\,|\,\underline{X} = \underline{x} + \rho\,\underline{1}) = E(L(r(q+r)^{-1}\underline{x} + (rq(q+r)^{-1})^{1/2}\,\underline{V},s)).$

Since the r.h.s. does not depend on m, t and ρ, part (I) is proved by putting $\rho = 0$.

(ii) If, given $\underline{X} = \underline{x}$, one $s^* \in \mathcal{A}$ minimizes the l.h.s. of (3.5) for one $\rho \in \mathbb{R}$, then it does it for all $\rho \in \mathbb{R}$ simultaneously. Thus every Bayes rule is translation-invariant, if one neglects possible pathological choices in cases where several solutions appear.

(iii) Let S be a translation-invariant rule and let for $\underline{x} \in \mathbb{R}^k$

$$R(S|\underline{x}) = E(L(\underline{\Theta},S(\underline{x}))\,|\,\underline{X} = \underline{x})$$

denote the conditional risk of S - given $\underline{X} = \underline{x}$. By (3.5) we see that $R(S|\underline{x})$ does not depend on m and t and moreover is translation-invariant in $\underline{x}$.

Now, marginally, $\underline{X}$ is distributed according to $N(\underline{\tilde{\mu}},\tilde{a}\ I + \tilde{b}\ U)$ where $\underline{\tilde{\mu}}$, $\tilde{a}$ and $\tilde{b}$ are given by (3.2) (b). If we choose now for $\underline{X}$ the representation given by Lemma 1 then we see that the overall risk $E(L(\underline{\Theta}, S(\underline{X}))) = E(R(S|\underline{X}))$ does not depend on m and t. Thus the proof of the theorem is completed.

Remark 3. In the present (Bayesian) framework Theorem 5 clearly fits all our needs. But it should be pointed out that the following simple fact can be viewed as being the basis of this theorem:

"Let $\underline{Y} \sim N(\underline{\mu} + \rho\ \underline{1},\ a\ I + b\ U)$ with $\underline{\mu} \in \mathbb{R}^k$, $\rho \in \mathbb{R}$, $a > 0$ and $b > -a/k$. Then there exists a random vector $\underline{Z}$ with $\underline{Z} \sim N(\underline{\mu},\ a\ I\ )$ such that $h(\underline{Y}) = h(\underline{Z})$ everywhere for every translation-invariant $h\colon \mathbb{R}^k \to \mathbb{R}^k$".

For the remainder of this paper we restrict our considerations to additive loss functions with translation-invariant $\ell_i(\underline{\theta})$, $i = 1,\dots,k$. Since by Theorem 5 we can put $m = t = 0$ without loss of generality, our model can considerably be simplified to

$$(3.6) \qquad \underline{X}|\underline{\theta} = \underline{\theta} \sim N(\underline{\theta}, q\ I), \quad \underline{\theta} \sim N(\underline{0}, r\ I),$$

which in turn implies

$$(3.7) \qquad \underline{\theta}|\underline{X} = \underline{x} \sim N(r(q+r)^{-1}\ \underline{x},\ r\ q(q+r)^{-1}\ I),\ \underline{X} \sim N(\underline{0},(q+r)I).$$

Thus, given $\underline{X} = \underline{x}$, $\underline{\theta}$ has the representation:

$$(3.8) \qquad \underline{\theta} = r(q+r)^{-1}\underline{x} + (r\ q(q+r)^{-1})^{1/2}\ \underline{V} \text{ with } \underline{V} \sim N(\underline{0},I).$$

The next result is to some extent sharper than Theorems 3 and 4:

Theorem 6. _If under the normal model (3.1)_

$L(\underline{\theta},s) = \sum_{i \in s} \ell\ (\theta_{[k]} - \theta_i)$ _with nondecreasing_ ℓ, _then (2.9) holds and every Bayes rule can be assumed to be ordered and monotone_.

Proof: By Theorem 2 every Bayes rule selects according to small values of $E(\ell(\theta_{[k]} - \theta_i)|\underline{x})$. Since ℓ is non-decreasing, we have for every $\underline{\theta} \in \Omega^k$ $\ell(\theta_{[k]} - \theta_i) = \max_{j=1,\dots,k} \ell(\theta_j - \theta_i)$, $i = 1,\dots,k$. But by (3.8) we get for every $\underline{x} \in \mathbb{R}^k$

$$(3.9) \qquad E(\ell(\theta_{[k]} - \theta_i)|\underline{x}) = E(\max_{j=1,\dots,k} \ell[r(q+r)^{-1}(\bar{x}_j - \bar{x}_i) + (r\ q(q+r)^{-1})^{1/2}(V_j - V_i)]),$$

which clearly implies the desired result.

Example 3. Under the normal model (3.1) consider the additive loss function

$$(3.10) \qquad L(\underline{\theta},s) = \sum_{i \in s} [k^{-1} \sum_{j=1}^{k} \theta_j - \theta_i - \varepsilon],\ \varepsilon > 0.$$

Since it is translation-invariant, by Theorem 4 we can assume that (3.8) holds, and the unique Bayes rule S* turns out to be

$$(3.11) \qquad i \in S^*(\underline{X}) \text{ iff } \bar{X}_i \geq k^{-1} \sum_{j=1}^{k} \bar{X}_j - r^{-1}(q+r)\varepsilon,$$

which is the well known procedure of Seal (1957).

There is a long story about the question how good different well established procedures like this perform in certain circumstances. One result is that Seal's procedure is not safe to use and especially inferior to Gupta's maximum means procedures (cf. Definition 3) with respect to many aspects. This is shown for example in Seal (1957), Deely and Gupta (1968) and Gupta and Miescke (1978). One new argument in this direction now seems to be to us that Seal's procedure under the normal model turns out to be a Bayes solution w.r.t. a very unrealistic loss function.

Perhaps the best known subset selection procedure is due to Gupta (1956, 65):

Definition 3. Gupta's maximum means procedure is given by

(3.12) $i \in S_d(\underline{X})$ iff $\bar{X}_i \geq \bar{X}_{[k]} - q^{1/2} d$, $d > 0$.

In the classical (non-Bayesian) approach due to Gupta one has to choose an $P^* > k^{-1}$ and then $d(k,P^*)$ is given by the requirement that S_d should contain the best population with probability at least P^* for every fixed $\underline{\theta} \in \Omega^k$. (Conversely if d is predetermined then $P^*(k,d)$ is fixed.)

It is conjectured and partially proved by many authors that S_d performs well or even is optimal in many situations. But up to now it was not possible to find S_d to be close to a Bayes rule in any given model, except perhaps in the Monte Carlo-studies of Chernoff and Yahav (1977) and Gupta and Hsu (1977). In this spirit our following results seem to be interesting.

Theorem 7. If under the normal model (3.1)

$L(\underline{\theta},s) = \sum_{i \in s} \ell(\theta_{[k]} - \theta_i - \varepsilon)$, where $\varepsilon > 0$ is fixed, ℓ is non-decreasing, continuous, bounded and satisfies $\ell(\rho) = 0$ if and only if $\rho = 0$, then the following procedure S is the limit of Bayes rules for large n:

(3.13) $\quad i \in S(\underline{X})$ *iff* $\bar{X}_i \geq \bar{X}_{[k]} - \varepsilon$.

Proof: Given $\underline{X} = \underline{x}$, by Theorem 2 and 5 and especially (3.8) the following rules are Bayes rules for $n = 1,2,\ldots$:

$$i \in S_n^*(\underline{x}) \text{ iff } E(\ell[\max_{j=1,\ldots,k} \{r(q+r)^{-1}(\bar{x}_j - \bar{x}_i) + (r\;q\;(q+r)^{-1})^{1/2}\,(V_j - V_i)\} - \varepsilon]) \leq 0.$$

Since ℓ is bounded and continuous, for large n (by Lebesgue's dominated convergence theorem) the expectation converges to $\ell[\max_{j=1,\ldots,k}\{\bar{x}_j - \bar{x}_i\} - \varepsilon]$.

Thus in view of the additional assumptions which we imposed upon ℓ the theorem is proved.

Though we have seen that the limit of Bayes rules is very similar to rules of type (3.12), this does not completely satisfy our requests, since procedures of type (3.13) have a $P^* = P^*(k, q^{-1/2}\,\varepsilon)$ which tends to one for large n. Thus for a moment we alternatively take another loss function as given in the theorem below:

Theorem 8. *If under the normal model* (3.1)

$L(\underline{\theta}, s) = \sum_{i \in s} (\theta_{[k]} - \theta_i - q^{1/2}\,d)$ *with a fixed* $d > 0$, *then Gupta's maximum means procedure* S_d *is the limit of the (unique) Bayes rules as* n *tends to infinity*.

Proof: Given $\underline{X} = \underline{x}$, by Theorems 2 and 5 and especially (3.8) the unique Bayes rules for $n = 1,2,\ldots$ turn out to be

(3.14) $\quad i \in S_n^*(\underline{x}) \quad$ iff

$$\bar{x}_i \geq E(\max_{j=1,\ldots,k} \{\bar{x}_j + (r^{-1}q(q+r))^{1/2} V_j\}) - r^{-1}(q+r)q^{1/2}\,d,$$

if the $\bar{x}_j$'s are distinct. By (A.9) we have

$$(3.15)\quad \bar{x}_{[k]} \leq E(\max_{j=1,..,k}\{\bar{x}_j + (r^{-1}q(q+r))^{1/2}\, V_j\}) \leq \bar{x}_{[k]} + \tilde{A}_2(\underline{x}),\quad \text{where}$$

$$\tilde{A}_2(\underline{x}) = (2r^{-1}q(q+r))^{1/2} \sum_{j=1}^{k-1} T((2r^{-1}q(q+r))^{-1/2}(\bar{x}_{[j]}-\bar{x}_{[k]}))$$

and the function T is given by (A.1).

Since by (3.7) $\underline{X}$, unconditionally, has non-degenerate normal distribution, we can assume that all the $\bar{x}_j$'s are distinct. But then by (A.2) we have $\tilde{A}_2(\underline{x}) = o(q^{1/2})$ and the proof is completed by noting that r remains fixed and $q = \delta^2/n$ tends to zero if n tends to infinity.

At least we shall study the case of linear loss functions in more detail. Here we have to distinguish between two possibilities:

$$(3.16)\quad L_1(\underline{\theta},s) = \sum_{i\in s} (\theta_{[k]} - \theta_i - \theta),\ \text{with fixed } \varepsilon > 0 \text{ and}$$

$$(3.17)\quad L_2(\underline{\theta},s) = \sum_{i\in s} (\theta_{[k]} - \theta_i - q^{1/2} d),\ \text{with fixed } d > 0.$$

By Theorems 7 and 8 we know that asymptotically L_2 leads to exactly one population (the best in every situation where $\underline{\Theta} = \underline{\theta}$ is fixed) whereas L_1 leads to a screening procedure even in the limit. On the other hand for every finite n both L_1 and L_2 give us screening procedures.

For the case of finite n we give now some approximations to the Bayes rule which apply to L_1 and L_2 simultaneously. These approximations are very easy to handle since they do not involve any integral and can in fact be evaluated with the help of the function T alone. For convenience we formulate our results in terms of L_1.

The results for L_2 follow easily since one has to replace ε by $q^{1/2}d$ only.

<u>Corollary 2.</u> <u>If under the normal model (3.1)</u> $L_1(\underline{\theta},s)$ <u>is given by (3.16) and</u> $\varepsilon > 0$ <u>and</u> n <u>are fixed, then with probability one we have for</u> $\alpha = 1,2$ <u>and</u> $\beta = 1,\ldots,4$

(3.18) $\underset{\sim}{S}_{\alpha}(\underline{X}) \subseteq S^*(\underline{X}) \subseteq \tilde{S}_{\beta}(\underline{X})$,

where for $\underline{x} \in \mathrm{IR}^k$ $i \in \underset{\sim}{S}_{\alpha}(\underline{x})$ *(or* $\tilde{S}_{\beta}(\underline{x})$*) iff*

$\bar{x}_i \geq \bar{x}_{[k]} - r^{-1}(q+r)\varepsilon + \tilde{A}_{\alpha}(\underline{x})$ *(or* $\underset{\sim}{A}_{\beta}(\underline{x})$*),*

and $\tilde{A}_{\alpha}$, $\underset{\sim}{A}_{\beta}$ *are given by* (A.8-11) *and* (A.17-18).

Proof: The assertion follows from (3.14), where $q^{1/2}$ d is to be replaced by ε, and by the fact that for all α,β and $\underline{x} \in \mathrm{IR}^k$ we have

$$\bar{x}_{[k]} + \underset{\sim}{A}_{\beta}(\underline{x}) \leq E(\max_{j=1,\dots,k} (\bar{x}_j + (r^{-1}q(q+r))^{1/2}\, V_j)) \leq \bar{x}_{[k]} + \tilde{A}_{\alpha}(\underline{x}),$$

which is proved in the appendix.

Let us explicitly point out the following special case of (3.18) if we take $\tilde{A}_1$ and $\underset{\sim}{A}_1$: Then with probability one we have

(3.19) $S_{d_1}(\underline{X}) \subseteq S^*(\underline{X}) \subseteq S_{d_2}(\underline{X})$,

where $d_2 = q^{-1/2}\, r^{-1}(q+r)\varepsilon$, $d_1 = q^{-1/2}\, r^{-1}(q+r)\varepsilon - (r^{-1}(q+r))^{1/2}\, a_k$,

and a_k is given by (A.2).

Here we can expect that for moderate n in many cases of $\underline{X} = \underline{x}$ S_{d_1} and S_{d_2} coincide, so that the experimenter finds the Bayes rule with the help of two means procedures, which are very easy to compute. This idea of course analogously applies to the other approximations, which, however, are no longer means procedures of type (3.12). Finally it should be pointed out that it is always possible to use the exact Bayes procedure (3.14) (with $q^{1/2}$ d possibly replaced by ε) if one is willing to evaluate either (A.4) or (A.5) with the help of a computer program.

Appendix: Expectation of the Largest of k Independently Distributed Normal Random Variables.

We derive now lower and upper bounds for

$$E(\max_{j=1,\ldots,k} (\mu_j + \rho V_j)) = E_k(\underline{\mu},\rho), \text{ say, where}$$

$\mu_j = \bar{x}_j$, $j = 1,\ldots,k$ and $\rho^2 = r^{-1} q(q+r)$ are fixed known and

$\underline{V} = (V_1,\ldots,V_k)' \sim N(\underline{0}, I)$.

Remark 4. Let us mention briefly that

$$r(q+r)^{-1} E(\max_{j=1,\ldots,k} (\mu_j + \rho\, V_j))$$

turns out to be Bayes estimate for the largest mean (i.e. $\theta_{[k]}$) under model (3.6) w.r.t. squared error loss. Thus our bounds (to be derived in the sequel) multiplied with $r(q+r)^{-1}$ can be utilized also as bounds for these Bayes estimates.

Let φ and Φ denote the density and distribution function of the one-dimensional standard normal distribution, and let T be the following auxiliary function which previously was used also by Goel and Rubin (1977):

$$\text{(A.1)} \quad T(\xi) = \int_{-\infty}^{\xi} \Phi(\eta)d\eta = \varphi(\xi) + \xi\Phi(\xi), \ \xi \in \mathbb{R}.$$

T is strictly increasing, strictly convex and satisfies

$$\text{(A.2)} \quad \lim_{\xi \to -\infty} T(\xi) = 0, \ T(0) = (2\pi)^{-1/2}, \ \lim_{\xi\to\infty} (T(\xi)-\xi) = 0.$$

Moreover let

$$\text{(A.3)} \quad a_k = E(\max_{j=1,\ldots,k} V_j), \quad k = 1,2,\ldots .$$

For further study and especially tables of the a_k's see David (1970).

Lemma 2.

$$\text{(A.4)} \quad E_k(\underline{\mu},\rho) = \mu_{[k]} + \sum_{j=1}^{k-1} \int_R \prod_{i=1}^{j} \Phi(\rho^{-1}(\xi-\mu_{[k-i+1]}))[1-\Phi(\rho^{-1}(\xi-\mu_{[k-j]}))]d\xi$$

$$\text{(A.5)} \quad E_k(\underline{\mu},\rho) = \mu_{[k]} + \int_R \Phi(\rho^{-1}(\xi-\mu_{[k]}))[1-\prod_{i=1}^{k-1} \Phi(\rho^{-1}(\xi-\mu_{[i]}))]d\xi.$$

Proof: By Chernoff and Yahav (1977) we have the following recursive relations for $j = 1,\dots,k-1$:

$$E(\max_{i\geq k-j} (\mu_i + \rho V_i)) - E(\max_{i\geq k-j+1} (\mu_i + \rho V_i))$$

$$= \int_{\mathbb{R}} \prod_{i=1}^{j} \Phi(\rho^{-1}(\xi-\mu_{[k-i+1]}))[1-\Phi(\rho^{-1}(\xi-\mu_{[k-j]}))]d\xi,$$

which clearly imply (A.4). And (A.5) follows by the telescopic property of the sum in (A.4).

For $k = 2$ this reduces to

Lemma 3.

$$\text{(A.6)} \quad E_2(\underline{\mu},\rho) = \mu_{[2]} + 2^{1/2}\rho\, T(2^{-1/2}\rho^{-1}(\mu_{[1]}-\mu_{[2]})).$$

Proof:

$$\begin{aligned} E_2(\underline{\mu},\rho) &= E(\mu_{[2]} + \rho V_2 + \max(0,\mu_{[1]} + \rho V_1 - \mu_{[2]} - \rho V_2)) \\ &= E(\mu_{[2]} + \rho V_2 + \max(0,\mu_{[1]} - \mu_{[2]} + 2^{1/2} \rho V_1)) \\ &= \mu_{[2]} + E((\mu_{[1]} - \mu_{[2]} + 2^{1/2} \rho V_1)^{+}). \end{aligned}$$

Applying Lemma 7 of Goel and Rubin (1977) the proof is completed. Besides we remark that (A.6) in another context (cf. Remark 4) was also derived by Blumenthal and Cohen (1968).

For later applications let us rewrite Lemma 3 in the following way:

(A.7) $$\int_{\mathbb{R}} \Phi(\rho^{-1}(\xi+\Delta))\Phi(-\xi^{-1}\xi)d\xi = 2^{1/2}\rho\, T(2^{-1/2}\rho^{-1}\Delta) \text{ for all } \Delta \in \mathbb{R} .$$

Then we state:

Lemma 4. $E_k(\underline{\mu},\rho) \leq \mu_{[k]} + \tilde{A}_\alpha$, $\alpha = 1,2$, _where_

(A.8) $\tilde{A}_1 = \rho\, a_k$, _and_

(A.9) $$\tilde{A}_2 = 2^{1/2}\rho \sum_{j=1}^{k-1} T(2^{-1/2}\rho^{-1}(\mu_{[j]} - \mu_{[k]})).$$

Proof: (A.8) is immediate. In view of (A.4) we have

$$E_k(\underline{\mu},\rho) \leq \mu_{[k]} + \sum_{j=1}^{k-1} \int_{\mathbb{R}} \Phi(\rho^{-1}(\xi-\mu_{[k]}))\Phi(-\rho^{-1}(\xi-\mu_{[k-j]}))d\xi.$$

Application of (A.7) and reordering of indices thus leads to (A.9).

Lemma 5. $E_k(\underline{\mu},\rho) \geq \mu_{[k]} + \underset{\sim}{A}_\beta$, $\beta = 1,2$, _where_

(A.10) $\underset{\sim}{A}_1 = 0$, _and_

(A.11) $$\underset{\sim}{A}_2 = 2^{1/2}\rho\, T(2^{-1/2}\rho^{-1}(\mu_{[k-1]} - \mu_{[k]})).$$

Proof: (A.10) is immediate. In view of (A.5) we have

$$E_k(\underline{\mu},\rho) \geq \mu_{[k]} + \int_{\mathbb{R}} \Phi(\rho^{-1}(\xi-\mu_{[k]}))\Phi(-\rho^{-1}(\xi-\mu_{[k-1]}))d\xi,$$

and thus application of (A.7) leads to (A.11).

By the convexity of T we get two further bounds with the help of Jensen's inequality. At first we state

Lemma 6.

(A.12) $$E_k(\underline{\mu},\rho) = \mu_{[k]} + \rho E(T(\rho^{-1}[\max_{j=1,\ldots,k-1}(\mu_j+\rho\, V_j)-\mu_{[k]}]))$$

(A.13) $$E_k(\underline{\mu},\delta) \geq \mu_{[k]} + \rho\, T(\rho^{-1}[E(\max_{j=1,\dots,k-1} (\mu_j + \rho\, V_j))-\mu_{[k]}]).$$

Proof: Integrating (A.5) by parts leads to

(A.14) $$E_k(\underline{\mu},\rho) = \mu_{[k]} + \rho \int_{\mathbb{R}} T(\rho^{-1}(\xi-\mu_{[k]}))d\{\prod_{i=1}^{k-1} \Phi(\rho^{-1}(\xi-\mu_{[i]}))\},$$

which equals to (A.12). And by Jensen's inequality (A.13) follows.

Remark 5. Though it is not necessary for our purposes, we point out that Lemma 6 can be generalized considerably as follows:

If $Y_1,\dots,Y_k$ are independent random variables with continuous distribution functions $F_1,\dots,F_k$, then under mild conditions on $F_1,\dots,F_k$ we have

(A.15) $$E(\max_{j=1,\dots,k} Y_j \mid \max_{j=1,\dots,k-1} Y_j = \xi) = \xi + \int_{\xi}^{\infty}[1-F_k(\eta)]d\eta,\ \xi \in \mathbb{R},$$

where the r.h.s. is convex in $\xi \in \mathbb{R}$. Thus Jensen's inequality implies

(A.16) $$E(\max_{j=1,\dots,k} Y_j) \geq E(\max_{j=1,\dots,k-1} Y_j) + \int_{E(\max_{j=1,\dots,k-1} Y_j)}^{\infty} [1-F_k(\eta)]d\eta.$$

On the other extreme, for standard normal variables, we have

$$E(\max_{j=1,\dots,k} V_j \mid \max_{j=1,\dots,k-1} V_j) = T(\max_{j=1,\dots,k-1} V_j) \quad \text{and}$$

$$E_k(\underline{0},\rho) = E(T(\max_{j=1,\dots,k-1} V_j)) \geq T(E_{k-1}(\underline{0},\rho))\dots \geq T\ \underbrace{o\dots o}_{k-1}\ T(0).$$

Lemma 7. $E_k(\underline{\mu},\rho) \geq \mu_{[k]} + \underset{\sim}{A}_{\beta}$, $\beta = 3,4$, *where*

(A.17) $$\underset{\sim}{A}_3 = \rho\ T(\rho^{-1}[a_{k-1}-k(k-1)^{-1}(\mu_{[k]}-\bar{\mu})])$$

with $\bar{\mu} = k^{-1}(\mu_1+\dots+\mu_k)$, *and*

(A.18) $$\underset{\sim}{A}_4 = \rho\ T\{\rho^{-1}(\mu_{[k-1]} - \mu_{[k]}) + 2^{1/2}T(2^{-1/2}\rho^{-1}(\mu_{[k-2]}-\mu_{[k-1]}))\}.$$

Proof: By (A.5)

$$E_k(\underset{\sim}{\mu},\rho) = \mu_{[k]} + \int_{\mathbb{R}} \Phi(\rho^{-1}\xi)[1-\prod_{i=1}^{k-1} \Phi(\rho^{-1}(\xi+\mu_{[k]} - \mu_{[i]}))]d\xi,$$

which by the log-concavity of Φ is bounded from below by

$$\mu_{[k]} + \int_{\mathbb{R}} \Phi(\rho^{-1}\xi)[1-\Phi(\rho^{-1}(\xi+\mu_{[k]} - (k-1)^{-1}(\mu_{[1]}+\dots+\mu_{[k-1]})))^{k-1}]d\xi.$$

Integrating by parts this in turn equals to

$$\mu_{[k]} + \rho\int_{\mathbb{R}} T(\rho^{-1}\xi)d\{\Phi(\rho^{-1}(\xi+k(k-1)^{-1}(\mu_{[k]} - \bar{\mu})))^{k-1}\}$$

$$= \mu_{[k]} + \rho E(T(\rho^{-1}\{\max_{j=1,\dots,k-1} V_j-k(k-1)^{-1}(\mu_{[k]}-\bar{\mu})\})),$$

and by Jensen's inequality this is bounded from below by (A.17). (A.18) follows immediately by applying (A.11) (for k-1 instead of k) to (A.13).

Remark 6. One can get more lower bounds of type (A.18) by iterating (A.13) N times (N = 1,2,...,k-2) before applying (A.11) (for k-N instead of k).

Note that among $\underset{\sim}{A}_2$, $\underset{\sim}{A}_3$ and $\underset{\sim}{A}_4$ (or among $\tilde{A}_1$ and $\tilde{A}_2$) none of the bounds is uniformly better than the others. One reason is that for every fixed $\eta \in \mathbb{R}$, $u\, T(u^{-1}\eta)$ is strictly increasing in u.

Acknowledgment. I would like to thank Professor Shanti S. Gupta and Woo-Chul Kim of Purdue University for their valuable comments and suggestions which led to an improvement of this paper.

REFERENCES

[1] Bickel, P. J. and Yahav, J. A. (1977). On selecting a set of good populations. Statistical Decision Theory and Related Topics II. (Gupta, S. S. and Moore, D. S. eds.). Academic Press, 37-55.

[2] Blumenthal, S. and Cohen, A. (1968). Estimation of the larger of two normal means. J. Amer. Statist. Assoc. 63, 861-876.

[3] Bratcher, T. L. and Bhalla, P. (1974). On the properties of an optimal selection procedure. Comm. Statist. 3, 191-196.

[4] Chernoff, H. and Yahav, J. A. (1977). A subset selection problem employing a new criterion. Statistical Decision Theory and Related Topics II. (Gupta, S. S. and Moore, D. S. eds.). Academic Press, 93-119.

[5] David, H. A. (1970). Order Statistics. John Wiley, New York.

[6] Deely, J. J. and Gupta, S. S. (1968). On the properties of subset selection procedures. Sankhya Ser. A 30, 37-50.

[7] Goel, P. K. and Rubin, H. (1977). On selecting a subset containing the best population - A Bayesian approach. Ann. Statist. 5, 969-983.

[8] Gupta, S. S. (1956). On a decision rule for a problem in ranking means. Mimeo. Series No. 150, Inst. of Statist., Univ. of North Carolina, Chapel Hill, North Carolina.

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

[10] Gupta, S. S. and Hsu, J. C. (1978). On the performance of some subset selection procedures. Communications in Statistics, Vol. B7, No. 6.

[11] Gupta, S. S. and Hsu, J. C. (1977). On the monotonicity of Bayes subset selection procedures. Book 4, Proceedings of the 41st Session of the International Statistical Institute held in New Delhi, December 1977, 208-211.

[12] Gupta, S. S. and Miescke, K. J. (1978). On subset selection procedures for ranking means of three normal populations. Mimeo. Series No. 78-19, Dept. of Statist., Purdue University, W. Lafayette, Indiana.

[13] Lehmann, E. L. (1957). A theory of some multiple decision problems I. Ann. Math. Statist. 28, 1-25.

[14] Seal, K. C. (1957). An optimum decision rule for ranking means of normal populations. Calcutta Statist. Assoc. Bull. 7, 131-150.

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER Mimeograph Series #78-23	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle)
Bayesian Subset Selection for Additive and Linear Loss Functions

5. TYPE OF REPORT & PERIOD COVERED
Technical

6. PERFORMING ORG. REPORT NUMBER
Mimeo. Series #78-23

7. AUTHOR(s)
Klaus-J. Miescke

8. CONTRACT OR GRANT NUMBER(s)
ONR N00014-75-C-0455

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

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

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

12. REPORT DATE
October 1978

13. NUMBER OF PAGES
23

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, Bayes Rules, Additive Linear Loss Function, Gupta's Maximum Means Procedure, Approximate Bayes Solutions, Expected Values of the Maximum.

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

Given k independent samples $\underline{X}_1,\ldots,\underline{X}_k$ of common size n from k populations $\pi_1,\ldots,\pi_k$ with distributions P_{θ_i}, $\theta_i \in \Omega \subseteq \mathbb{R}$, $i = 1,\ldots,k$, the problem is to select a non-empty subset $S(\underline{X}_1,\ldots,\underline{X}_k)$ from $\{\pi_1,\ldots,\pi_k\}$, which is associated with "good" (large) θ-values. More precisely, we consider this problem from a Bayesian approach. (over)

By choosing additive ($L(\underline{\theta},s) = \sum_{i\in s} \ell_i(\underline{\theta})$) and especially linear ($L(\underline{\theta},s) = \sum_{i\in s} (\theta_{[k]} - \theta_i - \varepsilon)$) loss functions we try to fill a gap lying in between the results of Deely and Gupta (1968) and more recent papers due to Goel and Rubin (1977), Gupta and Hsu (1978) and other authors. It is shown that under a certain "normal model" Seal's procedure turns out to be Bayes w.r.t. an unrealistic loss function whereas Gupta's maximum means procedure turns out to be (for large n) asymptotically Bayes w.r.t. more realistic additive loss functions. Finally, in the appendix some bounds for $E(\max_{j=1,\ldots,k} (\mu_j + \rho V_j))$ are derived (where $\underline{\mu} \in \mathbb{R}^k$, $\rho \in \mathbb{R}$ are fixed known and $\underline{V} \sim N(\underline{0},I)$) to approximate the Bayes rules w.r.t. linear loss functions in cases where n is finite.