〈Research Paper〉

Journal of the Korean
Statistical Society
Vol. 14, No. 1, 1985

A Study on Nonparametric Selection Procedures for Scale Parameters+

Moon Sup Song*, Han Young Chung* and Dong Jae Kim*

ABSTRACT

In this paper, we propose some nonparametric subset selection procedures for scale parameters based on rank-likes. The proposed procedures are compared to the Gupta-Sobel's parametric procedure through a small-sample Monte Carlo study. The results show that the nonparametric procedures are quite robust for heavy-tailed distributions, but they have somewhat low efficiencies.

1. Introduction

In many practical situations an experimenter is faced with the problem of selecting one or more out of several populations. The populations are usually characterized by a certain parameter and the experimenter is interested in choosing the best population associated with the largest or smallest parameter. For example, in a production line of measuring instrument the experimenter may want to choose the method associated with the least variability. In this paper, we are interested in selecting a subset which contains the best population in scale problem. The best population is the one having the smallest parameter.

Under the assumption of normality the subset selection procedures in terms of sample variances have been developed by Gupta and Sobel (1962) and Gupta and Huang (1976). McDonald (1977) investigated a class of selection procedures based on sample ranges.

* Department of Computer Science and Statistics, Seoul National University, Seoul 151, Korea
+ This research was partially supported by the Ministry of Education, through the Research Institute for Basic Sciences, SNU 1984—1985.

Kim and Song (1984) studied on the robustness of the Gupta-Sobel's procedure. The results show that the parametric procedure significantly violates the P^*-condition.

Gupta and McDonald (1970) studied on a class of subset selection procedures based on ranks for location parameters. In general, the least favorable configuration (LFC) is not given by the equal parameters configuration in selection procedures based on ranks as can be seen from the counter-examples of Rizvi and Woodworth (1970). To overcome this difficulty a nonparametric procedure based on pairwise ranks was proposed by Hsu (1980). In the Hsu's procedure, the LFC is given by an equal parameters configuration and the P^*-condition is exactly satisfied by the Hsu's procedure.

In this paper, we propose some nonparametric subset selection procedures for scale problem and compare the procedures with the parametric procedure by a samll-sample Monte Carlo study. The results show that the parametric procedure is too sensitive to the assumption of normality and the nonparametric procedures have somewhat low efficiencies.

2. Subset Selection Procedures Based on Rank-Likes

2.1. Preliminaries and Notations

We consider a set of k independent populations π_1, π_2, $\cdots$, π_k with cdf's $F((x-\theta)/\sigma_1)$, $F((x-\theta)/\sigma_2)$, $\cdots$, $F((x-\theta)/\sigma_k)$, respectively, where F is symmetric and absolutely continuous, θ is a common (known or unknown) location parameter and σ_i's are unknown scale parameters. Let $\sigma_{[1]} \leq \sigma_{[2]} \leq \cdots \leq \sigma_{[k]}$ be the ordered values of σ_i's. Our concern is to select a nonempty subset of populations containing the "best" one which has the smallest scale parameter $\sigma_{[1]}$. Thus, the correct selection(CS) means the selection of a subset which contains the population associated with the smallest scale parameter $\sigma_{[1]}$.

Gupta and Sobel (1962) proposed a selection procedure based on sample variances, which can be described as follows. Let X_{i1}, X_{i2}, $\cdots$, X_{in} be an independent sample from a normal population with mean θ_i and variance σ_i^2, $i=1, 2, \cdots, k$. Let S_i^2 be the sample variance defined by

$$S_i^2=\begin{cases}\frac{1}{n}\sum_{j=1}^{n}(X_{ij}-\theta_i)^2, & \text{if } \theta_i \text{ is known}\\ \frac{1}{n-1}\sum_{j=1}^{n}(X_{ij}-\bar{X}_i)^2, & \text{if } \theta_i \text{ is unknown,}\end{cases}$$

where $\bar{X}_i=\frac{1}{n}\sum_{j=1}^{n}X_{ij}$, $i=1,2,\cdots,k$. Let the ordered values of the k observed sample variances be denoted by

$$S_{[1]}^2\leq S_{[2]}^2\leq\cdots\leq S_{[k]}^2.$$

Then the Gupta-Sobel's parametric selection procedure R_1 is defined by

$$R_1: \text{Select } \pi_i \quad \text{iff} \quad S_i^2\leq S_{[1]}^2/c,$$

where $c=c(k,\ n,\ P^*)$ is determined to satisfy the P^*-condition

$$\inf_{\Omega} P(CS|R)\geq P^* \tag{2.1}$$

and $\Omega=\{\sigma^2=(\sigma_1^2,\sigma_2^2,\cdots,\sigma_k^2)\ :\ \sigma_i>0,\ i=1,2,\cdots,k\}$

2.2. A Procedure Based on Combined Rank-Likes

In this section we consider a nonparametric procedure based on rank-likes. We let

$$V_{i\beta}=|X_{i\beta}-\theta|,\ i=1,2,\ \cdots,\ k,\ \beta=1,2,\ \cdots,\ n,$$

and let $R_{i\beta}$ denote the rank of $V_{i\beta}$ in $C=\{V_{11},\ \cdots,\ V_{kn}\}$. Then, a class of statistics based on $R_{i\beta}$ is distribution-free under equal parameters configuration. We also let

$$R_i=\sum_{\beta=1}^{n}R_{i\beta},\ i=1,2,\cdots,k,$$

which is the rank sum of $V_{i1},\ \cdots,\ V_{in}$. We now propose a subset selection procedure R_2 based on the combined rank-likes, which is a scale problem analogue of the Gupta-McDonald procedure in location problem, as follows.

$$R_2: \text{Select } \pi_i \quad \text{iff} \quad R_i\leq D \text{ and/or } R_i=\min_{1\leq j\leq k}R_j,$$

where the constant D is to be determined to satisfy the P^*-condition (2.1). In the procedure R_2, the role of "select π_i if $R_i=\min_{1\leq j\leq k}R_j$" is to ensure that a nonempty subset is selected. Note that

$$P(R_i\leq D \text{ and/or } R_i=\min_{1\leq j\leq k}R_j)\geq P(R_i\leq D).$$

Therefore according to Theorem 3.2 of Gupta and McDonald (1970) we have

$$\inf P(CS|R_2)\geq P(W\leq D),$$

where W is the Wilcoxon rank sum statistic associated with samples of size n and $(k-1)n$ taken from identically distributed populations. We thus determine the value of D by solving the equation

$$P(W \leq D) = P^*.$$

The values of D can be obtained from the table of Wilcoxon rank sum statistic. For large values of k and n, we may use the normal approximation of the Wilcoxon statistic.

In constructing the selection rule R_2 we assumed that the common median θ is known. But, in most practical situations, the common median θ is unknown. We thus want to modify the procedures by using a robust estimator of θ. Let $M=\text{med}\ (X_{11}, \cdots, X_{1n}, \cdots, X_{k1} \cdots, X_{kn})$ be the combined sample median, and let

$$V_{i\beta}^* = |X_{i\beta} - M|, \quad i=1, 2, \cdots, k, \quad \beta=1, 2, \cdots, n.$$

Then the modified (approximate) procedure of R_2 using the ranks of $V_{i\beta}^*$'s is given by the following:

$$R_2^*: \text{ Select } \pi_i \text{ iff } R_i^* \leq D \text{ and/or } R_i^* = \min_{1 \leq j \leq k} R_j^*$$

where D is the value defined in R_2.

It is not proved that R_2^* satisfies the basic probability requirement. But, we can expect that the procedure at least approximately satisfies the P^*-condition. Note that, since M is symmetric in its arguments, according to Fligner, Hogg and Killeen (1976), any statistic based on the ranks of $V_{i\beta}^*$ is distribution-free under the equal parameters configuration. We investigate the small properties of R_2^* in Section 3.

2.3. Procedures Based on Pairwise Rank-Likes

In most of the nonparametric selection procedures using combined sample rank statistics, the LFC is not known. But, according to Theorem 3.1 of Gupta and McDonald (1970), for $k=2$ the infimum of $P(CS)$ for the procedure based on the combined sample ranks occurs at an equal parameters configuration. Using this property, Hsu (1980) proposed a nonparametric procedure based on pairwise ranks in location problem. Kim and Song (1984) also considered a nonparametric procedure in scale problem, which is an analogue of the Hsu's procedure, as follows. Let

$$V_{i\beta} = |X_{i\beta} - \theta|, \quad i=1, 2, \cdots, k, \quad \beta=1, 2, \cdots, n,$$

and $C_{ij}=\{V_{i1},\cdots, V_{in}, V_{j1},\cdots, V_{jn}\}$, $i\neq j$, $i,j=1,2,\cdots,k$. Let $R_{j\beta}^{(i)}$ denote the rank of $V_{j\beta}$ in C_{ij} and let $R_j^{(i)}$ be the rank sum of $V_{j1},\cdots, V_{jn}$ in C_{ij}, i.e. $R_j^{(i)}=\sum_{\beta=1}^{n}R_{j\beta}^{(i)}$. We also let $D_{(1)}^{(ji)}<D_{(2)}^{(ji)}<\ldots<D_{(n^2)}^{(ji)}$ denote the n^2 ordered differences $V_{i\alpha}-V_{j\beta}$, $\alpha,\beta=1,2,\cdots,n$. Note that these are distinct with probability one. Let

$$D_{\text{med}}^{(ji)}=\begin{cases}D_{(p+1)}^{(ji)} & \text{if } n^2=2p+1\\ [D_{(p)}^{(ji)}+D_{(p+1)}^{(ji)}]/2 & \text{if } n^2=2p\end{cases}$$

and let

$$T_i=\sum_{j=1}^{k}D_{\text{med}}^{(ji)}/k, \text{ for } i=1,2,\cdots,k,$$

with $D_{\text{med}}^{(ii)}=0$. The proposed procedure R_3 is as follows:

$$R_3: \text{Select } \pi_i \text{ iff } \min_{j\neq i} R_j^{(i)}>r_n(P^*) \text{ and/or } T_i=\min_{1\leq j\leq k} T_j,$$

where $r_n(P^*)$ is the smallest integer such that $P_0[\min_{j\neq i} R_j^{(i)}\leq r_n(P^*)]\leq 1-P^*$ and P_0 indicates that the probability is to be computed at equal parameters configuration. Assuming that θ is known, it can be shown that $\inf P[\min_{j\neq i} R_j^{(i)}>r_n(P^*)]$ occurs at the equal parameters configuration. (The proof is analogous to that of Hsu (1980).) The values of $r_n(P^*)$ can be obtained from Miller (1966).

We now consider another nonparametric procedure based on pairwise rank-likes with normal scores. Let

$$H_j^{(i)}=\sum_{\beta=1}^{n}E[Z_{(R_{j\beta}^{(i)})}], \quad i\neq j=1,2,\cdots,k,$$

where $Z_{(i)}$ is the i-th order statistic of a random sample of size 2n from a standard normal distribution. The procedure based on normal scores is defined by

$$R_4: \text{Select } \pi_i \text{ iff } \min_{j\neq i} H_j^{(i)}>h_n(P^*) \text{ and/or } T_i=\min_{1\leq j\leq k} T_j,$$

where $h_n(P^*)$ is the smallest integer such that $P_0[H_j^{(i)}\leq h_n(P^*)]\leq 1-(P^*)^{1/(k-1)}$.

To show that the procedure R_4 satisfies the P^*-condition we use the terminologies and techniques in Hsu (1980). Assuming, without loss of generality, that π_1 is the best, we can see that $\min_{j\geq 2} H_j^{(1)}$ is nondecreasing in $B=\{1,2,\cdots,n\}^c$. Let $G_{\sigma i}(x)$ be the cdf of $V_{i\beta}=|X_{i\beta}-\theta|$. Then it can be easily shown that $G_{\sigma_1}^n\cdots G_{\sigma_k}^n$ is stocastically larger than $G_{\sigma^*}^{kn}$ in B. Hence, the following inequalities can be verified without much difficulties:

$$P_\sigma(CS|R_4)\geq P_0[\min_{j\geq 2} H_j^{(1)}>h_n(P^*)]$$

$$\geq \prod_{j=2}^{k} P_0[H_i^{(1)} > h_n(P^*)]$$

$$\geq P^*$$

Thus, the procedure R_4 satisfies the P^*-condition.

The values of $h_n(P^*)$ can be obtained from the table of two-sample expected normal scores statistic. For example, the tables of $h_n(P^*)$ for $n=3$ (1) 10 and $(P^*)^{1/(k-1)}=.90$, .925, .95, .975, .99, .995, .999 were given by Bradley (1968).

In most practical situations, the common median θ is unknown. As mentioned in Section 2.2, we may use the combined sample median $M=\text{med}\ (X_{11}, \cdots, X_{1n}, \cdots, X_{k1}, \cdots, X_{kn})$, instead of θ. Let $V_{i\beta}^* = |X_{i\beta} - M|$, $i=1, 2, \cdots, k$, $\beta=1, 2, \cdots, n$, and we use an analogous notations. Also, according to Fligner, Hogg and Killeen (1976), $R_j^{*(i)}$ and $H_j^{*(i)}$ are distribution-free under equal parameters configuration. The modified (approximate) procedures of R_3 and R_4 are given by

$$R_3^* : \text{Select } \pi_i \quad \text{iff} \quad \min_{j \neq i} R_j^{*(i)} > r_n(P^*) \quad \text{and/or} \quad T_i^* = \min_{1 \leq j \leq k} T_j^*$$

and

$$R_4^* : \text{Select } \pi_i \quad \text{iff} \quad \min_{j \neq i} H_j^{*(i)} > h_n(P^*) \quad \text{and/or} \quad T_i^* = \min_{1 \leq j \leq k} T_j^*,$$

respectively, where $r_n(P^*)$ and $h_n(P^*)$ are the values defined in R_3 and R_4.

As mentioned for the procedure R_2^* in Section 2.2, R_3^* and R_4^* cannot satisfy the basic probability requirement, but we expect that the procedures at least approximately satisfy the P^*-condition. We investigate the small sample properties of R_3^* and R_4^* in Section 3.

3. Small-Sample Monte Carlo Study

In this section we compare the selection procedures R_1, R_2^*, R_3^* and R_4^* for various underlying distributions through a small-sample simulation study. The underlying distributions considered are normal, double exponential, contaminated normal and Cauchy distributions. Here, the ε-contaminated normal distribution has a pdf such that

$$f(x) = (1-\varepsilon)\phi(x) + \frac{\varepsilon}{a}\phi\left(\frac{x}{a}\right),$$

where ϕ is the pdf of the standard normal distribution.

We use the subroutine GGNML in IMSL (VAX 780) in generating normal samples with and without contamination. The other samples are also generated by using the

subroutine GGUBT in IMSL and the inverse integral transformations.

In the simulation study, we consider the equal-ratio configuration

$$\sigma_i=\sigma_{i-1}\delta,\ i=2,3,\cdots,k.$$

500 replications were performed for each value of $\delta=1.0,\ 1.5,\ 2.0,\ 3.0$. The constants used in our simulation study are $k=5$, $n=9$, $\sigma_1=1$. For the contaminated normal samples, $\varepsilon=0.1$ and $a=3,\ 5$ are taken. The exact values for c, $r_n(P^*)$ and $h_n(P^*)$ are used and a normal approximate value for D is used.

When $\delta=1$, the average number of selected populations divided by 500 can be interpreted as the empirical P^*. These values are given in Table 1. The procedures R_2^*, R_3^* and R_4^* almost satisfy the P^*-condition for all underlying distributions. But the Gupta-Sobel's procedure does not satisfy the P^*-condition except for the normal distribution. For example, the empirical P^* in the contaminated normal case with $a=5$ is 0.657, which is supposed to be 0.90.

To compare the efficiencies of the four procedures, we use the relative efficiency of the rule R_i^* to R_1, $i=2,3,4$, defined by

$$e(R_i^*,\ R_1)=\frac{E(S|R_1)}{E(S|R_i^*)}\cdot\frac{P(CS|R_i^*)}{P(CS|R_1)},$$

where S is the number of populations to be selected. To estimate this relative efficiency, the empirical relative efficiencies of R_i^* relative to R_1 are computed from the number of time that each population is selected in our simulation study. These results are summarized in Table 2.

Since the Gupta-Sobel's procedure significantly violates the P^*-condition, it may not be reasonable to compute the relative efficiency by the above definition. But, the values

Table 1. Empirical P^* based on 500 replications (with $k=5$, $n=9$, $\varepsilon=0.1$)

P^*	rule	normal	double exponential	contaminated normal $a=3.0$	contaminated normal $a=5.0$	Cauchy
.90	R_1	.910	.762	.785	.657	.471
	R_2^*	.892	.888	.900	.897	.888
	R_3^{*+}	.898	.921	.900	.903	.901
	R_4^*	.923	.915	.908	.923	.928
.95	R_1	.949	.838	.838	.734	.522
	R_2^*	.946	.946	.953	.947	.948
	R_3^*	.960	.970	.960	.951	.961
	R_4^*	.960	.959	.967	.952	.962

+ The value of $r_n(P^*)$ for $P^*=0.90$ was obtained by simulation.

Table 2. Empirical relative efficiencies based on 500 replications (with $k=5$, $n=9$, $\varepsilon=0.1$)

P^*	efficiency	δ	normal	double exponential	contaminated normal $a=3.0$	contaminated normal $a=5.0$	Cauchy
.90	$e(R_2^*, R_1)$	1.5	.636	.623	.631	.648	.684
		2.0	.479	.498	.508	.492	.590
		3.0	.369	.405	.404	.434	.511
	$e(R_3^*, R_1)$	1.5	.779	.708	.779	.801	.765
		2.0	.794	.721	.768	.792	.763
		3.0	.774	.694	.802	.813	.823
	$e(R_4^*, R_1)$	1.5	.766	.712	.745	.753	.718
		2.0	.742	.697	.747	.739	.751
		3.0	.741	.719	.749	.791	.815
.95	$e(R_2^*, R_1)$	1.5	.642	.619	.648	.667	.677
		2.0	.510	.496	.513	.510	.553
		3.0	.388	.389	.419	.433	.480
	$e(R_3^*, R_1)$	1.5	.760	.672	.739	.774	.707
		2.0	.743	.660	.742	.678	.685
		3.0	.736	.646	.749	.687	.750
	$e(R_4^*, R_1)$	1.5	.744	.662	.733	.739	.695
		2.0	.728	.664	.716	.661	.668
		3.0	.707	.661	.738	.728	.723

in Table 2 still indicate that the nonparametric procedures are quite robust with respect to the heaviness of distribution tails and the contamination. In general, the procedures based on pairwise rank-likes are better than the procedures based on combined rank-likes. Further studies on the asymptotic properties of the nonparametirc procedures may also be desirable.

REFERENCES

(1) Bradley, J.V. (1968), *Distribution-Free Statistical Tests*, Prentice-Hall, Inc., Englewood Cliffs, N.J..

(2) Fligner, N.A., Hogg, R.A. and Killeen, T.J. (1976), Some distribution-free rank-like statistics having the Mann-Whitney-Wilcoxon null distribution, *Commun. Statist.-Theor. Meth. A5*, 373—376.

(3) Gupta, S.S. and Huang, D.Y. (1976), On some methods for constructing optimal subset selection procedures, *Mimeo. Ser. No. 470*, Dept. of Statist., Purdue Univ., West Lafayette, Indiana.

(4) Gupta, S.S. and McDonald, G.C. (1970), On some classes of selection procedures based on ranks, *Nonparametric Techniques in Statistial Inference*, 491—514.

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

(6) Hsu, J.C. (1980), Robust and nonparametric subset selection procedures, *Commun. Statist.-Theor. Meth. A9*, 1439—1459.

(7) Kim, D.J. and Song, M.S. (1984), On the robustness of subset selection rules in scale problem, *Proc. Coll. Natur. Sci., SNU*, 9, No.2, 71—75.

(8) McDonald, G.C. (1977), Subset selection procedures based on sample ranges from a uniform population, *Technometrics, 18*, 343—349.

(9) Miller, R.G., Jr. (1966), *Simultaneous Statistical Inference*. McGraw Hill, New York.

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