1

A ROBUST SUBSET SELECTION PROCEDURE FOR LOCATION PARAMETER CASE BASED ON HODGES-LEHMANN ESTIMATORS *

by

K. S. Lee

Purdue University

Technical Report #88-60C

PURDUE UNIVERSITY

CENTER FOR STATISTICAL DECISION SCIENCES AND DEPARTMENT OF STATISTICS

1

A ROBUST SUBSET SELECTION PROCEDURE FOR LOCATION PARAMETER CASE BASED ON HODGES-LEHMANN ESTIMATORS [*]

by

K. S. Lee
Purdue University

Technical Report #88-60C

Department of Statistics
Purdue University

November 1988

[*] Research supported in part by the Korea Science & Engineering Foundation Overseas Fellowship, 1987. This research was done while the author was a visiting scholar in the Department of Statistics at Purdue University. Research also partially supported under the Office of Naval Research Contract N00014-88-K-0170 and NSF Grant DMS-8717799 at Purdue University.

A Robust Subset Selection Procedure for Location Parameter case based on Hodges–Lehmann Estimators[*]

by

Kang Sup Lee

Dankook University and Purdue University

Abstract

This paper deals with a robust subset selection procedure based on Hodges–Lehmann estimators of location parameters. An improved formula for the estimated standard error of Hodges–Lehmann estimators is considered. Also, the degrees of freedom of the studentized Hodges–Lehmann estimators are investigated and it is suggested to use $0.8n$ instead of $n-1$. The proposed procedure is compared with the other subset selection procedures and it is shown to have good efficiency for heavy–tailed distributions.

1 Introduction

Many classical subset selection procedures based on sample means have been developed under the assumption of normality. But, it is well known that the sample mean is very

[*] Research supported in part by the Korea Science & Engineering Foundation Overseas Fellowship, 1987. This research was done while the author was visiting scholar the Department of Statistics at Purdue University. Research also partially supported under the Office of Naval Research Contract N00014-88-K-0170 and NSF Grant DMS-8717799 at Purdue University.

sensitive to the departures from normality. We thus want some robust procedures which perform reasonably well over a wide range of underlying distributions and are insensitive to gross errors.

Robust subset selection procedures have been developed by using either rank scores or robust estimators. Subset selection procedures based on rank are investigated by Bartlett and Govindarajulu (1968) and Gupta and McDonald (1970). But, a critical difficulty of the procedures based on ranks is, in general, to find the least favorable configuration (LFC). For example, in the procedure proposed by Bartlett and Govindarajulu (1968) the LFC is not given by the equi-parameter configuration, which was proved by the counterexample of Rizvi and Woodworth (1970). To tide over this difficulty, some procedures based on robust estimators, such as sample medians, trimmed means, Huber's M-estimators and Hodges-Lehmann estimators, are considered. Some important contributions in the subset selection procedures based on sample medians have been made by Gupta and Leong (1979), Gupta and Singh (1980), Lorenzen and McDonald (1981) and Gupta and Sohn (1987) for several distributions. Subset selection procedures based on trimmed means have been proposed and studied by Song, Chung and Bae (1982) and Song and Kim (1984). Lee (1985) has considered a procedure based on Huber's M-estimators.

It is well known that under some regularity conditions, the Hodges-Lehmann (H-L) estimator derived from the Wilcoxon signed-rank test is an unbiased estimator of the location parameter and is robust with respect to contaminations and heaviness of distribution tails. Hence some subset selection procedures based on H-L estimators have been considered. Gupta and Huang (1974) have proposed some procedures based on one-sample H-L estimators assuming that the populations have a common known variance. For a two-way layout problem, Gupta and Leu (1987) have proposed an asymptotic distribution-free subset selection procedure based on H-L estimators. For the case of unknown variance, Song, Chung and Bae (1982) have studied the subset selection procedure based on the H-L estimators derived from the Wilcoxon signed-rank test. They used the median absolute deviation (MAD) to estimate the standard error of the H-L estimators. But, as pointed out by them, their proposed rule significantly violates the P^*-condition in heavy-tailed distributions since the MAD usually underestimates the standard error of the H–L estimators in heavy–tailed distributions. To overcome this violation, Song and Kim (1987) have developed a subset selection procedure based on the H–L estimators with the A–estimator which is an estimator of the standard error of the H–L estimator.

The purpose of this paper is to propose a robust subset selection procedure for the location parameter based on the H–L estimators. To derive a selection procedure we use a modified Sievers and McKean's (1986) estimator of the standard error of the H–L estimator rather than the A–estimator. Section 2 deals with a studentization of the H–L estimators. In Section 3, a subset selection procedure is proposed and compared with the other subset selection procedures through a small–sample Monte Carlo study. The results of the Monte Carlo study show that the proposed procedure is successful in satisfying the P^*–condition and also robust with respect to the heaviness of distribution of tails.

2 Studentizing Hodges–Lehmann Estimators

2.1 Estimation of the asymptotic standard error of Hodges–Lehmann Estimator

Let $X_1, \ldots, X_n$ be a random sample from a continuous and symmetric distribution $F(x-\theta)$ with a location parameter θ and density function $f(x-\theta)$. Under the regularity conditions, see Randles and Wolfe (1979) for details, the Hodges–Lehmann (H–L) estimator of θ based on the Wilcoxon signed–rank test is

$$\hat{\theta} = med_{i \leq j}\{(X_i + X_j)/2\}$$

and the asymptotic standard error σ_H of $\hat{\theta}$ is

$$\sigma_H = 1/(\sqrt{12n} \int f^2(x)\,dx). \tag{2.1.1}$$

Using the fact that $\sigma_H^2 = \pi\sigma^2/3n$ in the case of mormal distribution, Song and Kim (1987) proposed an estimator $\hat{\sigma}_S$ of σ_H

$$\hat{\sigma}_S = \sqrt{\pi/3nS_b} \qquad (2.1.2.)$$

where S_b is a biweight A–estimator of scale σ introduced by Lax (1985).

In (2.1.1), let $\gamma = \int f^2(x)\, dx$. Then the asymptotic standard error of the H–L estimator is proportional to γ^{-1}. There are some ways to estimate γ^{-1}. Lehmann (1963) proposed a consistent estimator of γ^{-1} based on the length of a distribution–free confidence interval for θ. Sievers and McKean (1986) proposed an estimator of γ^{-1} based on the difference between two ordered pairwise differences and showed that their estimator is consistent for both asymmetric and symmetric distributions. Sievers and McKean's estimator is given by

$$\hat{\gamma}^{-1} = \frac{2\hat{t}_\alpha/\sqrt{n}}{\hat{G}_n(\hat{t}_\alpha/\sqrt{n})}$$

where $\hat{t}_\alpha$ is the αth quantile of $\hat{G}_n(t)$, the empirical distribution function of the positive pairwise differences, that is,

$$\hat{G}_n(t) = \frac{2}{n(n-1)} \sum_{i<j} I(\mid X_i - X_j \mid \le t).$$

Therefore the standard error of $\hat{\theta}$ can be estimated by

$$\hat{\sigma}_H = \frac{1}{\sqrt{12n}} \hat{\gamma}^{-1}. \tag{2.1.3}$$

In the choice of the quantile α, Sievers and McKean (1986) recommended $\alpha = 0.8$.

But, as pointed out by Sievers and McKean (1986), the estimate $\hat{\sigma}_H$ in (2.1.3) require small sample corrections. Hence, in order to check the bias of the estimated standard error $\hat{\sigma}_H$, a Monte Carlo study was performed. To find empirical values of $\hat{\sigma}_H$ in (2.1.3), 1000 pseudo–random samples of size 10, 20 and 30 were generated from the normal, double exponential, contaminated normal, Cauchy, exponential, lognormal and skewed contaminated normal distributions. The subroutines GGNML, GGCAY, GGEXN and GGUBS in IMSL and inverse integral transformation were used. The cdf of contaminated normal and skewed contaminated normal distributions are given by

$$F(x) = (1-\epsilon)\Phi(x) + \epsilon\Phi(x/\sigma) \quad \text{and} \quad F(x) = (1-\epsilon)\Phi(x) + \epsilon\Phi((x-a)/\sigma),$$

respectively. The computations in this Monte Carlo study were carried out in double precision arithmetic on VAX–11/780 at Department of Statistics, Purdue University.

For a generated sample of size n, the values of $\hat{\sigma}_H$ in (2.1.3) were computed for different values of the quantile α. This process was repeated 1000 times for each values of $n = 10$, 20 and 30. The averages of these 1000 values of $\hat{\sigma}_H$ are summarized in Table 2.1.

The results in Table 2.1 show that $\hat{\sigma}_H$ in (2.1.3) significantly overestimates the standard error of $\hat{\theta}$. Hence some corrections are required. In fact, Sievers and McKean (1986) considered the standard least squares corrections for small sample, namely,

$$\hat{\sigma}_L = \sqrt{(n-1)/n}\hat{\sigma}_H. \tag{2.1.4}$$

But, as shown in Table 2.1, $\hat{\sigma}_L$ also overestimates the standard error of $\hat{\theta}$. Thus, to improve the behavior of $\hat{\sigma}_H$ in (2.1.3), we considered the following estimated standard error of $\hat{\theta}$ which is a slight modification of $\hat{\sigma}_H$:

$$\hat{\sigma}_M = \sqrt{(n-2)/n}\hat{\sigma}_H. \tag{2.1.5}$$

The results in Table 2.1 show that the modified standard error $\hat{\sigma}_M$ performs better than $\hat{\sigma}_H$ and $\hat{\sigma}_L$. Also, unlike Sievers and McKean's suggestion, the value $\alpha = 0.5$ produced good result in our study.

2.2 Studentization of Hodges–Lehmann Estimators

After the works of Tukey and McLaughlin (1963) and the conjectures of Huber (1970), some contributions in the studentization of robust estimators, especially M–estimators, have been made by Leone, Jayachandran and Eisenstat (1967), Gross (1973), Shorack (1976) and Lee(1985) for various ψ–functions. Recently, Song and Kim (1987) have considered a studentization of H–L estimators with biweight A–estimator of scale. The above researches are successful although the formulas of the number of degrees of freedom are unsound. The general philosophy of the studentization of robust estimators has been discussed by Huber (1970, 1981).

We now want to approximate the distribution of the quotient

$$\frac{\hat{\theta} - \theta}{\hat{\sigma}_M} \tag{2.2.1}$$

Table 2.1

A Comparision of the Asymptotic Standard Error σ_H

and Estimated Standard Errors $\hat{\sigma}$ of $\hat{\theta}$

Based on 1000 Replications.

(a) Normal Distribution

n	σ_H	α	$\hat{\sigma}_H$	$\hat{\sigma}_L$	$\hat{\sigma}_M$
10	0.3236	0.5	0.3824(0.0049)	0.3627(0.0046)	0.3420(0.0044)
		0.6	0.3710(0.0043)	0.3520(0.0041)	0.3319(0.0038)
		0.7	0.3662(0.0039)	0.3474(0.0037)	0.3275(0.0035)
		0.8	0.3618(0.0035)	0.3433(0.0033)	0.3236(0.0031)
		0.9	0.3597(0.0033)	0.3413(0.0031)	0.3218(0.0030)
20	0.2288	0.5	0.2476(0.0020)	0.2413(0.0019)	0.2349(0.0019)
		0.6	0.2446(0.0018)	0.2384(0.0018)	0.2321(0.0017)
		0.7	0.2429(0.0017)	0.2367(0.0017)	0.2304(0.0016)
		0.8	0.2425(0.0017)	0.2364(0.0016)	0.2301(0.0016)
		0.9	0.2429(0.0015)	0.2367(0.0015)	0.2304(0.0015)
30	0.1868	0.5	0.1955(0.0012)	0.1922(0.0012)	0.1889(0.0011)
		0.6	0.1943(0.0011)	0.1911(0.0011)	0.1877(0.0011)
		0.7	0.1940(0.0011)	0.1908(0.0010)	0.1875(0.0010)
		0.8	0.1940(0.0010)	0.1908(0.0010)	0.1875(0.0010)
		0.9	0.1947(0.0010)	0.1914(0.0010)	0.1881(0.0009)

Note: The numbers in parentheses are the estimated standard errors of $\hat{\sigma}$

Table 2.1 (continued)

A Comparision of the Asymptotic Standard Error σ_H

ɛnd Estimated Standard Errors $\hat{\sigma}$ of $\hat{\theta}$

Based on 1000 Replications.

(b) Double Exponential Distribution

n	σ_H	α	$\hat{\sigma}_H$	$\hat{\sigma}_L$	$\hat{\sigma}_M$
10	0.3651	0.5	0.4588(0.0068)	0.4353(0.0065)	0.4104(0.0061)
		0.6	0.4502(0.0061)	0.4271(0.0058)	0.4027(0.0054)
		0.7	0.4488(0.0056)	0.4257(0.0053)	0.4014(0.0050)
		0.8	0.4503(0.0054)	0.4272(0.0051)	0.4027(0.0048)
		0.9	0.4580(0.0051)	0.4345(0.0049)	0.4096(0.0096)
20	0.2582	0.5	0.2883(0.0028)	0.2810(0.0027)	0.2735(0.0026)
		0.6	0.2869(0.0026)	0.2796(0.0026)	0.2722(0.0025)
		0.7	0.2862(0.0025)	0.2790(0.0024)	0.2716(0.0023)
		0.8	0.2879(0.0024)	0.2806(0.0023)	0.2731(0.0022)
		0.9	0.2929(0.0023)	0.2855(0.0022)	0.2778(0.0022)
30	0.2108	0.5	0.2243(0.0017)	0.2205(0.0016)	0.2167(0.0016)
		0.6	0.2238(0.0016)	0.2200(0.0016)	0.2162(0.0015)
		0.7	0.2241(0.0015)	0.2203(0.0015)	0.2165(0.0015)
		0.8	0.2256(0.0015)	0.2218(0.0015)	0.2179(0.0014)
		0.9	0.2285(0.0014)	0.2246(0.0014)	0.2207(0.0014)

Note: The numbers in parentheses are the estimated standard errors of $\hat{\sigma}$

Table 2.1 (continued)

A Comparision of the Asymptotic Standard Error σ_H

and Estimated Standard Errors $\hat{\sigma}$ of $\hat{\theta}$

Based on 1000 Replications.

(c) Contaminated Normal Distribution ($\epsilon = 0.1$, $\sigma = 5$)

n	σ_H	α	$\hat{\sigma}_H$	$\hat{\sigma}_L$	$\hat{\sigma}_M$
10	0.3754	0.5	0.4513(0.0063)	0.4281(0.0060)	0.4036(0.0057)
		0.6	0.4408(0.0056)	0.4182(0.0053)	0.3943(0.0050)
		0.7	0.4439(0.0053)	0.4211(0.0051)	0.3970(0.0048)
		0.8	0.4558(0.0055)	0.4324(0.0052)	0.4077(0.0049)
		0.9	0.4891(0.0066)	0.4640(0.0062)	0.4375(0.0059)
20	0.2655	0.5	0.2846(0.0025)	0.2774(0.0025)	0.2700(0.0024)
		0.6	0.2828(0.0023)	0.2756(0.0023)	0.2683(0.0022)
		0.7	0.2823(0.0022)	0.2752(0.0022)	0.2678(0.0021)
		0.8	0.2842(0.0022)	0.2770(0.0022)	0.2697(0.0021)
		0.9	0.3020(0.0023)	0.2944(0.0023)	0.2865(0.0022)
30	0.2168	0.5	0.2265(0.0016)	0.2227(0.0016)	0.2188(0.0015)
		0.6	0.2260(0.0015)	0.2222(0.0015)	0.2183(0.0015)
		0.7	0.2264(0.0015)	0.2226(0.0014)	0.2187(0.0014)
		0.8	0.2280(0.0015)	0.2242(0.0014)	0.2203(0.0014)
		0.9	0.2339(0.0015)	0.2300(0.0014)	0.2260(0.0014)

Note: The numbers in parentheses are the estimated standard errors of $\hat{\sigma}$

Table 2.1 (continued)

A Comparision of the Asymptotic Standard Error σ_H

and Estimated Standard Errors $\hat{\sigma}$ of $\hat{\theta}$

Based on 1000 Replications.

(d) Cauchy Distribution

n	σ_H	α	$\hat{\sigma}_H$	$\hat{\sigma}_L$	$\hat{\sigma}_M$
10	0.5736	0.5	0.7831(0.0162)	0.7429(0.0154)	0.7004(0.0145)
		0.6	0.7882(0.0163)	0.7478(0.0154)	0.7050(0.0146)
		0.7	0.9072(0.0255)	0.8607(0.0242)	0.8114(0.0228)
		0.8	2.0495(0.2120)	1.9443(0.2011)	1.8331(0.1896)
		0.9	3.3322(0.4223)	3.1612(0.4006)	2.9805(0.3777)
20	0.4056	0.5	0.4683(0.0055)	0.4564(0.0053)	0.4442(0.0052)
		0.6	0.4722(0.0054)	0.4602(0.0053)	0.4479(0.0052)
		0.7	0.4910(0.0061)	0.4786(0.0060)	0.4658(0.0058)
		0.8	0.5499(0.0091)	0.5360(0.0089)	0.5217(0.0086)
		0.9	2.3775(0.2276)	2.3173(0.2218)	2.2555(0.2159)
30	0.3312	0.5	0.3627(0.0034)	0.3566(0.0034)	0.3504(0.0033)
		0.6	0.3658(0.0034)	0.3596(0.0033)	0.3534(0.0033)
		0.7	0.3747(0.0036)	0.3684(0.0035)	0.3620(0.0035)
		0.8	0.4025(0.0043)	0.3958(0.0043)	0.3889(0.0042)
		0.9	0.6102(0.0184)	0.5999(0.0180)	0.5895(0.0177)

Note: The numbers in parentheses are the estimated standard errors of $\hat{\sigma}$

Table 2.1 (continued)

A Comparision of the Asymptotic Standard Error σ_H

and Estimated Standard Errors $\hat{\sigma}$ of $\hat{\theta}$

Based on 1000 Replications.

(e) Exponential Distribution

n	σ_H	α	$\hat{\sigma}_H$	$\hat{\sigma}_L$	$\hat{\sigma}_M$
10	0.1826	0.5	0.2483(0.0034)	0.2356(0.0033)	0.2221(0.0031)
		0.6	0.2519(0.0033)	0.2389(0.0032)	0.2253(0.0030)
		0.7	0.2591(0.0031)	0.2458(0.0029)	0.2318(0.0028)
		0.8	0.2692(0.0031)	0.2554(0.0029)	0.2408(0.0028)
		0.9	0.2956(0.0035)	0.2804(0.0033)	0.2644(0.0031)
20	0.1291	0.5	0.1526(0.0014)	0.1488(0.0013)	0.1448(0.0013)
		0.6	0.1551(0.0013)	0.1511(0.0013)	0.1471(0.0013)
		0.7	0.1593(0.0013)	0.1552(0.0013)	0.1511(0.0012)
		0.8	0.1649(0.0013)	0.1607(0.0013)	0.1564(0.0012)
		0.9	0.1762(0.0013)	0.1717(0.0013)	0.1672(0.0013)
30	0.1054	0.5	0.1186(0.0009)	0.1166(0.0008)	0.1146(0.0008)
		0.6	0.1202(0.0008)	0.1182(0.0008)	0.1162(0.0008)
		0.7	0.1227(0.0008)	0.1206(0.0008)	0.1185(0.0008)
		0.8	0.1266(0.0008)	0.1245(0.0008)	0.1223(0.0008)
		0.9	0.1338(0.0009)	0.1316(0.0008)	0.1293(0.0008)

Note: The numbers in parentheses are the estimated standard errors of $\hat{\sigma}$

Table 2.1 (continued)

A Comparision of the Asymptotic Standard Error σ_H

and Estimated Standard Errors $\hat{\sigma}$ of $\hat{\theta}$

Based on 1000 Replications.

(f) Lognormal Distribution

n	σ_H	α	$\hat{\sigma}_H$	$\hat{\sigma}_L$	$\hat{\sigma}_M$
10	0.2520	0.5	0.3379(0.0056)	0.3206(0.0053)	0.3022(0.0050)
		0.6	0.3467(0.0056)	0.3289(0.0053)	0.3101(0.0050)
		0.7	0.3734(0.0062)	0.3543(0.0059)	0.3340(0.0055)
		0.8	0.4108(0.0071)	0.3897(0.0067)	0.3675(0.0063)
		0.9	0.4920(0.0103)	0.4668(0.0098)	0.4401(0.0092)
20	0.1782	0.5	0.2043(0.0020)	0.1992(0.0020)	0.1939(0.0019)
		0.6	0.2073(0.0020)	0.2020(0.0020)	0.1967(0.0019)
		0.7	0.2151(0.0020)	0.2096(0.0020)	0.2040(0.0019)
		0.8	0.2274(0.0022)	0.2216(0.0021)	0.2157(0.0021)
		0.9	0.2791(0.0036)	0.2720(0.0035)	0.2647(0.0034)
30	0.1455	0.5	0.1587(0.0013)	0.1560(0.0013)	0.1533(0.0012)
		0.6	0.1611(0.0013)	0.1584(0.0012)	0.1556(0.0012)
		0.7	0.1650(0.0013)	0.1623(0.0012)	0.1594(0.0012)
		0.8	0.1735(0.0013)	0.1706(0.0013)	0.1676(0.0013)
		0.9	0.1954(0.0017)	0.1921(0.0017)	0.1887(0.0016)

Note: The numbers in parentheses are the estimated standard errors of $\hat{\sigma}$

Table 2.1 (continued)

A Comparision of the Asymptotic Standard Error σ_H

and Estimated Standard Errors $\hat{\sigma}$ of $\hat{\theta}$

Based on 1000 Replications.

(g) Skewed Contaminated Normal Distribution ($\epsilon = 0.18$, $\sigma = 13.5$, $a = 1.9$)

n	σ_H	α	$\hat{\sigma}_H$	$\hat{\sigma}_L$	$\hat{\sigma}_M$
10	0.4588	0.5	0.8004(0.0228)	0.7594(0.0217)	0.7159(0.0204)
		0.6	0.9525(0.0316)	0.9036(0.0300)	0.8520(0.0282)
		0.7	1.4783(0.0412)	1.4025(0.0391)	1.3223(0.0369)
		0.8	1.8297(0.0418)	1.7358(0.0396)	1.6365(0.0374)
		0.9	2.5759(0.0514)	2.4437(0.0487)	2.3040(0.0460)
20	0.3244	0.5	0.3872(0.0055)	0.3774(0.0054)	0.3674(0.0052)
		0.6	0.4603(0.0092)	0.4487(0.0090)	0.4367(0.0087)
		0.7	0.6297(0.0143)	0.6138(0.0140)	0.5974(0.0136)
		0.8	0.8624(0.0179)	0.8405(0.0174)	0.8181(0.0169)
		0.9	1.2930(0.0171)	1.2602(0.0167)	1.2266(0.0163)
30	0.2649	0.5	0.2961(0.0028)	0.2911(0.0028)	0.2860(0.0027)
		0.6	0.3224(0.0044)	0.3170(0.0043)	0.3114(0.0042)
		0.7	0.4107(0.0070)	0.4038(0.0069)	0.3967(0.0068)
		0.8	0.5962(0.0096)	0.5862(0.0095)	0.5760(0.0093)
		0.9	0.8894(0.0105)	0.8744(0.0103)	0.8592(0.0102)

Note: The numbers in parentheses are the estimated standard errors of $\hat{\sigma}$

by a t-distribution with appropriate degrees of freedom where $\hat{\theta}$ is the H-L estimator of θ and $\hat{\sigma}_M$,defined in (2.1.5), is an estimated standard error of $\hat{\theta}$. Huber (1970) suggested a method to determine an equivalent number of degrees of freedom by the asymptotic distribution of a consistent estimator of the asymptotic variance σ_H^2. He conjectured that the degrees of freedom are $(2/C)n$ with

$$C = 16\left(\frac{\int f^3(x)dx}{\left(\int f^2(x)dx\right)^2} - 1\right).$$

For the normal distribution, $2/C = 0.808$ which motivate us to consider the degrees of freedom in the subset selection procedures based on the H-L estimators with the estimated standard error $\hat{\sigma}_M$ defined in (2.1.5).

To check the goodness-of-fit of the studentized H-L estimator (2.2.1), we performed a small-sample simulation study. For each sample of size $n = 10$ and 20, three cases of the degrees of freedom, that is, $n-1$, $n-2$ and $0.8n$, are considered. To drive comparative studentization, we included the studentization of the sample means with usual sample standard deviation, H-L estimator with $\hat{\sigma}_H$ defined in (2.1.3) and H-L estimator with $\hat{\sigma}_S$ defined in (2.1.2). That is, in our simulation we included the following six studentizations:

$$T_1 = \frac{\overline{X} - \theta}{S/\sqrt{n}} \quad \text{with } df = n-1; \qquad T_2 = \frac{\hat{\theta} - \theta}{\hat{\sigma}_S} \quad \text{with } df = n-1;$$

$$T_3 = \frac{\hat{\theta} - \theta}{\hat{\sigma}_H} \quad \text{with } df = n-1; \qquad T_4 = \frac{\hat{\theta} - \theta}{\hat{\sigma}_M} \quad \text{with } df = n-1;$$

$$T_5 = \frac{\hat{\theta} - \theta}{\hat{\sigma}_M} \quad \text{with } df = n-2; \qquad T_6 = \frac{\hat{\theta} - \theta}{\hat{\sigma}_M} \quad \text{with } df = 0.8n;$$

where $\overline{X}$ is the sample mean and S is the usual sample standard deviation. And the other notations are as defined in Section 2.1. Note that $T_5 = T_6$ for sample size $n = 10$.

For each distribution of the normal, double exponential, contaminated normal and Cauchy the simulation was repeated 1,000 times with sample of size $n = 10$ and 20. The probabilty $P(T \geq t(\nu, p))$ is estimated by the number of values exceeding $t(\nu, p)$ divided by 1,000 , where $t(\nu, p)$ is the $100(1-p)$ percentile of the t-distribuion with ν degrees of freedom and T is one of the quotients mentioned above. These estimated probabilities are summarized in Table 2.2.

The results in Table 2.2 show that the t-distribution approximation of the quotient T_6, H-L estimators with $\hat{\sigma}_M$ and the degrees of freedom $\nu = 0.8n$, is good. If the underlying distribution is normal, T_1 and T_2 gave good results. However, in the heavy-tailed distributions, T_6 are better than T_1 or T_2. T_5 and T_6 gave almost the same results, however, the usage of T_6 looks slightly better than T_5.

3 A Robust Procedure Based on Hodges-Lehmann Estimator for Selecting the Best Location Parameter

3.1 Subset Selection Procedures

Let $\pi_1, \ldots, \pi_k$ be k independent populations with cdf's $F(\frac{x-\theta_1}{\sigma}), \ldots, F(\frac{x-\theta_k}{\sigma})$, respectively, unknown location parameters θ_i and a common unknown variance σ^2. Let $X_{i1}, \ldots, X_{in}$ be a random sample of size n from the population π_i, $i = 1, \ldots, k$. We assume that the experimenter has no prior knowledge concerning the pairing of the π_i with the jth ranked value $\theta_{[j]}$ of the θ_i's , $i = 1, \ldots, k, j = 1, \ldots, k$. The goal of the experimenter is to select the 'best' population associated with the largest location parameter $\theta_{[k]}$. If more than one population are best , we tag one of them and consider it as the 'best'.

Gupta (1956, 1965) has suggested the following subset selection procedure R_G based on the sample means.

Gupta's procedure (R_G): Select π_i if and only if

$$\overline{X}_i \geq \max_{1 \leq j \leq k} \overline{X}_j - \frac{dS}{\sqrt{n}}$$

where $\overline{X}_i$ is the sample mean of the ith population, $d = d(k, n, P^*)$ is chosen so as to satisfy the P^*-condition, and S^2 is the usual pooled sample variance with $\nu = k(n-1)$ degrees of freedom.

Table 2.2

Estimated Probability of $P(T \geq t(\nu, p))$

Based on $1,000$ Replication

(a) Sample size $n = 10$

Distribution	T	p:	0.400	0.250	0.100	0.050	0.025	0.010	0.005
Normal	T_1		0.377	0.246	0.118	0.057	0.033	0.014	0.007
	T_2		0.392	0.240	0.115	0.058	0.035	0.016	0.007
	T_3		0.369	0.208	0.090	0.053	0.030	0.013	0.008
	T_4		0.384	0.228	0.111	0.064	0.041	0.021	0.013
	T_6		0.384	0.227	0.109	0.064	0.036	0.018	0.011
Double Exponential	T_1		0.407	0.243	0.088	0.034	0.020	0.004	0.003
	T_2		0.380	0.221	0.065	0.027	0.013	0.005	0.003
	T_3		0.369	0.196	0.061	0.025	0.010	0.005	0.003
	T_4		0.383	0.226	0.080	0.039	0.016	0.009	0.005
	T_6		0.383	0.226	0.080	0.037	0.014	0.009	0.003
Contaminated Normal ($\epsilon = 0.1, \sigma = 5$)	T_1		0.428	0.261	0.091	0.041	0.020	0.007	0.004
	T_2		0.380	0.227	0.089	0.047	0.021	0.009	0.006
	T_3		0.375	0.205	0.081	0.046	0.017	0.008	0.004
	T_4		0.379	0.236	0.104	0.060	0.028	0.014	0.007
	T_6		0.379	0.234	0.102	0.056	0.025	0.011	0.005
Cauchy	T_1		0.418	0.304	0.095	0.028	0.012	0.002	0.001
	T_2		0.389	0.236	0.079	0.033	0.011	0.002	0.000
	T_3		0.387	0.216	0.070	0.037	0.014	0.004	0.002
	T_4		0.402	0.241	0.091	0.049	0.025	0.009	0.004
	T_6		0.402	0.240	0.090	0.048	0.024	0.009	0.004

Note: For sample size $n = 10$, $T_5 = T_6$

Table 2.2 (continued)

Estimated Probability of $P(T \geq t(\nu, p))$ Based on 1,000 Replication

(b) Sample size $n = 20$

Distribution	T	p: 0.400	0.250	0.100	0.050	0.025	0.010	0.005
Normal	T_1	0.402	0.239	0.097	0.053	0.031	0.008	0.003
	T_2	0.383	0.242	0.101	0.055	0.029	0.008	0.003
	T_3	0.380	0.228	0.094	0.047	0.024	0.009	0.007
	T_4	0.387	0.241	0.103	0.057	0.027	0.013	0.008
	T_5	0.387	0.241	0.102	0.057	0.027	0.012	0.008
	T_6	0.387	0.241	0.102	0.056	0.026	0.011	0.008
Double Exponential	T_1	0.428	0.256	0.089	0.043	0.021	0.009	0.002
	T_2	0.407	0.212	0.064	0.028	0.015	0.005	0.002
	T_3	0.408	0.226	0.072	0.035	0.015	0.009	0.001
	T_4	0.416	0.236	0.080	0.040	0.017	0.010	0.003
	T_5	0.416	0.236	0.080	0.040	0.017	0.010	0.003
	T_6	0.416	0.235	0.080	0.040	0.017	0.010	0.003
Contaminated Normal ($\epsilon = 0.1, \sigma = 5$)	T_1	0.412	0.266	0.093	0.042	0.015	0.007	0.002
	T_2	0.386	0.222	0.082	0.040	0.018	0.005	0.001
	T_3	0.386	0.229	0.094	0.041	0.018	0.007	0.004
	T_4	0.395	0.238	0.104	0.048	0.023	0.008	0.005
	T_5	0.395	0.238	0.104	0.047	0.022	0.008	0.005
	T_6	0.395	0.238	0.103	0.046	0.022	0.008	0.005
Cauchy	T_1	0.413	0.318	0.104	0.037	0.012	0.004	0.002
	T_2	0.396	0.228	0.073	0.029	0.009	0.003	0.001
	T_3	0.404	0.244	0.096	0.045	0.023	0.008	0.004
	T_4	0.412	0.251	0.113	0.052	0.025	0.011	0.005
	T_5	0.412	0.251	0.113	0.052	0.024	0.011	0.004
	T_6	0.412	0.251	0.111	0.051	0.023	0.011	0.004

If we assume that π_i is a normal population, then the constant d is a solution of

$$\int_0^\infty \int_{-\infty}^\infty \Phi^{k-1}(u+dw)\phi(u)q_\nu(w)dudw = P^* \tag{3.1.1}$$

where Φ and ϕ are cdf and density function of standard normal distribution, respectively, and $q_\nu(w)$ is density function of $\chi_\nu/\sqrt{\nu}$. The values of d have been tabulated by Gupta and Sobel (1957) and also by Gupta, Panchapakesan and Sohn (1985) (see $\rho = 0.5$ in this paper) for various combinations of k, ν and P^*.

Since Gupta's procedure R_G is based on the sample means and variances, it is sensitive to extreme observations. We thus want some robust selection procedures which are insensitive to outliers. As a robust procedure, Song and Kim (1987) have proposed the following subset selection rule R_S based on the H–L estimators with the biweight A-estimators of scale.

Song and Kim's procedure (R_S): Select π_i if and only if

$$\hat{\theta}_i \geq \max_{1\leq j\leq k} \hat{\theta}_j - d_b S_b \tag{3.1.2}$$

where $\hat{\theta}_i$ is the H–L estimator of θ_i and S_b is the pooled sample estimated standard error of the H–L estimator, that is, $S_b^2 = \sum_{i=1}^k \hat{\sigma}_{iS}^2/k$ with $\hat{\sigma}_{iS}$ defined in (2.1.2) for the ith population. In (3.1.2), Song and Kim (1987) used d values of Gupta's procedure as given by (3.1.1); they provide approximate values of d_b.

However, as shown in the above section, the modified standard error $\hat{\sigma}_M$ in (2.1.5) of the H–L estimator $\hat{\theta}$ with the degrees of freedom $\nu = 0.8n$ has a good behavior in the heavy-tailed distributions. We thus want to propose an improved selection procedure based on H–L estimators. The proposed selection procedure is as follows.

Proposed procedure (R_M): Select π_i if and only if

$$\hat{\theta}_i \geq \max_{1\leq j\leq k} \hat{\theta}_j - d_m S_m \tag{3.1.3}$$

where $\hat{\theta}_i$ is the H–L estimator of θ_i and S_m is the pooled sample estimated standard error of the H–L estimator, that is, $S_m^2 = \sum_{i=1}^k \hat{\sigma}_{iM}^2/k$ with $\hat{\sigma}_{iM}$ defined in (2.1.5) for the ith population.

The constant d_m is also to be determined to satisfy the P^*–condition. But, since the distribution of $\hat{\theta}_M$ and S_m are too complicated to determine d_m, the exact values of d_m to satisfy the P^*–condition are not available. However, the results of the above section imply that we may use the constants d in (3.1.1) for the constants d_m in (3.1.3) after changing the degrees of freedom from $k(n-1)$ to $k(0.8n)$ as the studies of Lee (1985) and Song and Kim (1987).

3.2 An Empirical Study on the Procedures

This section treats the results of a Monte Carlo study to compare the three subset selection procedures, Gupta's procedure R_G based on the sample means, Song and Kim's procedure R_S based on the H–L estimators with A–estimator for scale and the proposed procedure R_M based on the H–L estimators with modified estimated standard error and degrees of freedom. The purpose of this Monte Carlo study is to compare the three procedures for various underlying distributions including the normal, double exponential, contaminated normal and Cauchy distributions.

To investicate the performance of the three procedures, equally–spaced –parameter case is considered, that is,

$$\theta_i = \theta_0 + (i-1)\delta\sigma, i = 1, \dots, k$$

where $\delta > 0$ is a given constant and σ is the standard deviation of each population. When the distribution does not possess the second moment, the value of $F^{-1}(0.84) - F^{-1}(0.5)$ is used instead of the value of standard deviation. The constants used in our simulation study are $k = 5$, $n = 10$. For the contaminated normal distributions, $\epsilon = 0.1$ and $\sigma = 5$ are considered.

1,000 replications were performed for each value of $\delta\sqrt{n} = 0, 2$ and 4. When $\delta\sqrt{n} = 0$, the average number of selected populations divided by 1,000 can be interpreted as the empirical P^*. These values are given in Table 3.1. The empirical results show that the proposed procedure R_M successfully satisfies the P^*–condition for various distributions.

Table 3.1

Empirical P^* Based on 1,000 Replications

Distribution	Rule	P^*:	0.750	0.900	0.950	0.975	0.990
Normal	R_G		0.7424	0.8982	0.9498	0.9756	0.9902
	R_S		0.7448	0.9018	0.9536	0.9764	0.9894
	R_M		0.7914	0.9236	0.9658	0.9830	0.9918
Double	R_G		0.7552	0.8990	0.9510	0.9756	0.9882
Exponential	R_S		0.7982	0.9308	0.9674	0.9830	0.9942
	R_M		0.8020	0.9240	0.9628	0.9836	0.9932
Contaminated	R_G		0.7484	0.9082	0.9552	0.9806	0.9940
Normal	R_S		0.8050	0.9370	0.9730	0.9872	0.9952
$(\epsilon = 0.1, \sigma = 5)$	R_M		0.7948	0.9286	0.9658	0.9828	0.9912
Cauchy	R_G		0.6820	0.9066	0.9636	0.9832	0.9942
	R_S		0.8112	0.9234	0.9574	0.9756	0.9906
	R_M		0.7870	0.9074	0.9474	0.9684	0.9824

To compare the efficiencies of selection procedures, we use the following definition of the relative efficiency of the procedure R_1 to the procedure R_2 suggested by Song and Oh (1981):

$$e(R_1, R_2) = \frac{E(S|R_2)}{E(S|R_1)} \times \frac{P(CS|R_1)}{P(CS|R_2)}$$

where $E(S|R)$ is the expected number of populations to be retained in the selected suset for a given procedure R. To estimate the relative efficiency, empirical relative efficiencies of R_M relative to R_G are computed from the number of times that each population is selected in 1,000 replications. The results are summarized in Table 3.2.

The results in Table 3.2 show that the performances of the robust selection procedures R_S and R_M are satisfactory. For the normal distribution, Gupta's rule R_G is better than R_S and R_M. However, the rules R_S and R_M are quite robust with respect to contaminations and heaviness of distribution tails. Also, we find that the rule R_M is slightly better than the rule R_S for heavy–tailed distributions.

Table 3.2

Empirical Relative Efficiencies Based on 1,000 Replications

Distribution	Efficiency	$\delta\sqrt{n}$	P^*: 0.750	0.900	0.950	0.975	0.990
Normal	$e(R_S, R_G)$	2	0.985	0.973	0.967	0.971	0.966
		4	0.984	0.980	0.981	0.969	0.970
	$e(R_M, R_G)$	2	0.936	0.905	0.892	0.891	0.884
		4	0.955	0.931	0.918	0.886	0.900
Double Exponential	$e(R_S, R_G)$	2	1.043	1.020	1.023	1.019	1.016
		4	1.018	1.025	1.001	0.999	1.020
	$e(R_M, R_G)$	2	1.032	1.011	1.014	1.002	1.004
		4	1.012	1.013	0.984	0.991	1.001
Contaminated Normal	$e(R_S, R_G)$	2	1.137	1.143	1.150	1.136	1.136
		4	1.153	1.183	1.190	1.211	1.212
	$e(R_M, R_G)$	2	1.142	1.158	1.169	1.148	1.145
		4	1.166	1.183	1.190	1.211	1.212
Cauchy	$e(R_S, R_G)$	2	1.213	1.240	1.184	1.142	1.098
		4	1.623	1.695	1.644	1.576	1.470
	$e(R_M, R_G)$	2	1.265	1.295	1.241	1.182	1.127
		4	1.670	1.785	1.727	1.659	1.546

4 Acknowledgements

The author wishes to thank to Professor Shanti S. Gupta and Professor Moon Sup Song for their extensive comments for this work. Also thanks are to the Department of Statistics Purdue University for giving me an opportunity to visit during 1988–89 and for providing facilities for my research.

References

[1] Bartlett, N. S. and Govindarajulu, Z. (1968). Some distribution-free statistics and their application to the selection problem. *Ann. Inst. Statist. Math.*, 20, 79-97.

[2] Gross, A. M. (1973). A robust confidence interval for location for long-tailed distributions. *Proceedings of the National Academy of Science*, USA, 70, 1995-1997.

[3] Gupta, S. S. (1956). On a decision rule for a problem in ranking means. Ph. D. Thesis *(Mimeo. Ser. No. 150).* Inst. of Statist., Univ. of North Carolina, Chapel Hill.

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

[5] Gupta, S. S. and Huang, D. Y. (1974). Nonparametric subset selection procedures for the t best populations. *Bull. Inst. Math. Acad. Sincia*, 2, 377-386.

[6] Gupta, S. S. and Leong, Y. -K. (1979). Some results on subset selection procedures for double exponential populations. *Decision Information* (Tsokos, C. P. and Thrall, R. M. Ed.), Academic Press, New York, 277-305.

[7] Gupta, S. S. and Leu, L. -Y. (1987). An asymptotic distribution-free selection procedure for a two-way layout problem. *Commun. Statist.-Theory Meth.*, 16(8), 2313-2325.

[8] Gupta, S. S. and McDonald, D. C. (1970). On some classes of selection procedures based on ranks. *Nonparametric Techniques in Statistical Inference* (Puri, M. L., Ed.), Cambridge University Press, Cambridge, England, 491-514.

[9] Gupta, S. S., Panchapakesan, S. and Sohn, J. K. (1985). On the distribution of the studentized maximum of equally correlared normal random variables. *Commun. Statist.-Simula. Computa.*, 14(1), 103-135.

[10] Gupta, S. S. and Singh, A. K. (1980). On rules based on sample medians for selection of the largest location parameter. *Commun. Statist.-Theor. Meth.*, A9(14), 1277-1298.

[11] Gupta, S. S. and Sobel, M. (1957). On a statistic which arises in selection and ranking problems. *Ann. Math. Statist.*, 28, 957–967.

[12] Gupta, S. S. and Sohn, J. K. (1987). Selection and ranking procedures for Tukey's generalized lambda distributions. Technical Report No. 85–29, Dept. of Statistics, Purdue University, West Lafayette, Indiana.

[13] Huber, P. J. (1970). Studentizing robust estimates. *Nonparametric Techniques in Statistical Inference* (Puri, M. L., Ed.), Cambridge University Press, Cambridge, England, 453–463.

[14] Huber, P. J. (1981). *Robust Statistics.* John Wiley and Sons, Inc., New York.

[15] Lax, D. A. (1985). Robust estimators of scale: finite–sample performance in long-tailed symmetric distributions. *J. Amer. Statist. Assoc.*, 80, 736–741.

[16] Lee, K. S. (1985). A study on selection procedures based on Huber's M–estimators. Ph. D. Thesis, Seoul National University.

[17] Lehmann, E. L. (1963). Nonparametric confidence intervals for a shift parameter. *Ann. Math. Statist.*, 34, 1507–1512.

[18] Leone, F. C., Jayachandran, T. and Eisenstat, S. (1967). A study of robust estimators. *Technometrics*, 9, 652–660.

[19] Lorenzen, T. J. and McDonald, G. C. (1981). Selecting logistic populations using the sample medians. *Commun. Statist.-Theor. Meth.*, A10(15), 101–124.

[20] Randles, R. H. and Wolfe, D. A. (1979). *Introduction to the Theory of Nonparametric Statistics.* John Wiley and Sons, New York.

[21] Rizvi, M. H. and Woodworth, G. G. (1970). On selection procedures based on ranks: Counter–examples concerning least favorable configurations. *Ann. Math. Statist.*, 41, 1942–1951.

[22] Shorack, G. R. (1976). Robust studentization of location estimates. *Statistica Neerlandica*, 30, 119–141.

[23] Sievers, G. L. and McKean, J. W. (1986). On the robust rank analysis of linear models with nonsymmetric error distributions. *J. Statist. Plann. Inference*, 13, 215–230.

[24] Song, M. S., Chung, H. Y. and Bae, W. S. (1982). Subset selection procedures based on some robust estimators. *J. Korean Statist. Soc.*, 11, 109–117.

[25] Song, M. S. and Kim, S. -K. (1987). On a subset selection procedure based on Hodges–Lehmann estimators. *J. Korean Statist. Soc.*, 17, (to appear).

[26] Song, M. S. and Kim, Y. W. (1984). On the subset selection procedure based on trimmed means. *Proc. Coll. Natur. Sci.*, SNU, 9, 17–25.

[27] Song, M. S. and Oh, C. H. (1981). On a robust subset selection procedure for the slope of regression equations. *J. Korean Statist. Soc.*, 10, 105–121.

[28] Tukey, J. W. and McLaughlin, D. H. (1963). Less vulnerable confidence and significance procedures for location based on a single sample. *Sankhyā*, Ser. A25, 331-352.

REPORT DOCUMENTATION PAGE

1a. REPORT SECURITY CLASSIFICATION

1b. RESTRICTIVE MARKINGS

2a. SECURITY CLASSIFICATION AUTHORITY
Unclassified

2b. DECLASSIFICATION / DOWNGRADING SCHEDULE

3. DISTRIBUTION / AVAILABILITY OF REPORT

4. PERFORMING ORGANIZATION REPORT NUMBER(S)
Technical Report #88-60C

5. MONITORING ORGANIZATION REPORT NUMBER(S)

6a. NAME OF PERFORMING ORGANIZATION
Purdue University

6b. OFFICE SYMBOL *(If applicable)*

7a. NAME OF MONITORING ORGANIZATION

6c. ADDRESS *(City, State, and ZIP Code)*
Department of Statistics
West Lafayette, IN 47907

7b. ADDRESS *(City, State, and ZIP Code)*

8a. NAME OF FUNDING / SPONSORING ORGANIZATION
Office of Naval Research

8b. OFFICE SYMBOL *(If applicable)*

9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER
KOSEF Overseas Fellowship
N00014-88-K-0170 and DMS-8717799

8c. ADDRESS *(City, State, and ZIP Code)*
Daejeon, Chungnam 200, Korea
Arlington, VA 22217-5000

10. SOURCE OF FUNDING NUMBERS

PROGRAM ELEMENT NO.	PROJECT NO.	TASK NO.	WORK UNIT ACCESSION NO.

11. TITLE *(Include Security Classification)*
A robust subset selection procedure for location parameter case based on Hodges-Lehmann estimators

12. PERSONAL AUTHOR(S)
Kang Sup Lee

13a. TYPE OF REPORT
Technical

13b. TIME COVERED
FROM ________ TO ________

14. DATE OF REPORT *(Year, Month, Day)*
November 29, 1988

15. PAGE COUNT
23

16. SUPPLEMENTARY NOTATION

17. COSATI CODES

FIELD	GROUP	SUB-GROUP

18. SUBJECT TERMS *(Continue on reverse if necessary and identify by block number)*
Robust Subset Selection Procedure, location parameter, Hodges-Lehmann Estimator, Estimated Standard Error of H-L Estimator

19. ABSTRACT *(Continue on reverse if necessary and identify by block number)*

This paper deals with a robust subset selection procedure based on Hodges-Lehmann estimators of location parameters. An improved formula for the estimated standard error of Hodges-Lehmann estimators is considered. Also, the degrees of freedom of the studentized Hodges-Lehmann estimators are investigated and it is suggested to use 0.8n instead of n-1. The proposed procedure is compared with the other subset selection procedures and it is shown to have good efficiency for heavy-tailed distributions.

20. DISTRIBUTION / AVAILABILITY OF ABSTRACT
☒ UNCLASSIFIED/UNLIMITED □ SAME AS RPT. □ DTIC USERS

21. ABSTRACT SECURITY CLASSIFICATION

22a. NAME OF RESPONSIBLE INDIVIDUAL
Professor Shanti S. Gupta

22b. TELEPHONE *(Include Area Code)*
(317) 494-6031

22c. OFFICE SYMBOL

83 APR edition may be used until exhausted.
All other editions are obsolete.

SECURITY CLASSIFICATION OF THIS PAGE
UNCLASSIFIED