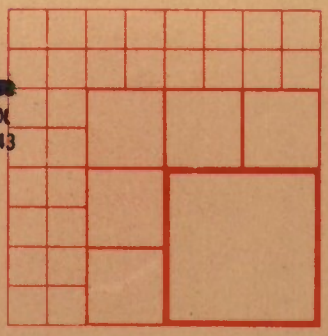


STATISTICS CENTER

Massachusetts Institute of Technology

77 Massachusetts Avenue, Rm. E40-111, Cambridge, Massachusetts 02139 (617) 253-8722

A ROBUST ESTIMATOR OF LOCATION USING AN ADAPTIVE SPLINE MODEL

BY

DONG YOON KIM
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

TECHNICAL REPORT NO. ONR-37

MARCH 1985

PREPARED UNDER CONTRACT

N00014-74-C-0555 (NR-609-001)

FOR THE OFFICE OF NAVAL RESEARCH

A ROBUST ESTIMATOR OF LOCATION USING AN ADAPTIVE SPLINE MODEL

by

Dong Yoon Kim

ABSTRACT

This paper contains a new approach toward the robust estimation of a location parameter. We propose NPS (Normal Pareto Spline) distribution which provides rough fit to density functions for arbitrary unimodal symmetric distributions. The bases of our NPS estimation are Pareto tails and spline constraints. Pareto tails can represent a diversity of tail behavior, and spline constraints ensure the smoothness of the density function.

We show that the NPS estimate of location has lower asymptotic variance than Huber's M-estimator in most cases, regardless of how Huber's trimmed constant k is chosen.

We also show that the NPS estimate of location can guarantee resistance for outliers.

For the generalized two sample location problem, where the scale parameters are unequal, we propose an iterative method to estimate the shift parameter and also have a proof that this iterative method converges to the desired M-estimate for an arbitrary scale location family of symmetric distributions.

Some Key Words: Pareto Tails, Spline, Outlier, Robust Adaptive Estimation

AMS 1980 subject classification. Primary 62F35

Table of Contents

Chapter 1

Introduction

Suppose we assume for our underlying statistical model a set of distributions $\{P_\theta\}$, $\theta \in \Xi$, that are fairly well specified, and then in terms of this model, find a good estimator for some characteristics of the true underlying distribution. If the true distribution of the population is not closely approximated by one of the set $\{P_\theta\}$, $\theta \in \Xi$, then the estimator can have a large error, no matter how large the sample size. To safeguard against this danger, we need a robust estimator.

Our approach for symmetric distributions is to use a sufficiently rich class of distributions to approximate the family of symmetric distributions. The distributions in our class will be called Normal Pareto Spline distributions. Given i.i.d. observations $x_1,\ldots,x_n$ from an arbitrary symmetric distribution g, we shall estimate g and its characteristics by using the maximum likelihood estimate (MLE) on the false assumption that $g \in$ NPS. The MLE gives rise to $\hat{\theta}$ and $P_{\hat{\theta}}$. This estimator will be called the NPS estimator. Insofar as the NPS family of distributions is very rich, we can expect that there will be a member f_0 close to g, and the NPS estimator will be close to f_0 and therefore to g. A problem with evaluating

this approach arises from the fact that it is not obvious what the choice for the closest f_0 to g should be.

In a sense we are moving a step toward nonparametric density estimation by estimating g through this rich three-parameter family of NPS distributions. In another sense, to be explained later, this approach may be regarded as an adaptive generalization of a version of Huber's M estimator.

In chapter 2, we summarize other approaches, both nonadaptive and adaptive, to robust estimation. In chapter 3, we define the NPS distribution, and explore the characteristics of the NPS family of distributions. In chapter 4, we derive asymptotic properites of NPS estimates and summarize simulation results. Also we show that the NPS estimate of location will usually perform better than Huber's M-estimator. In chapter 5, we discuss several variations of the two-sample location problem and also introduce an asymmetric NPS distribution. In chapter 6, we explain certain computational techniques used in this dissertation, including a simplex method, a Monte-Carlo swindle, and a variance reduction method for the logistic distribution. In chapter 7, we summarize all results. In the appendix, we have program lists for the NPS MLE.

Chapter 2

Background

While model building is certainly desirable, we know in practive that most models will not exactly fit the real situation. A realistic approach seeks statistical procedures good for a broad class of possible underlying models. Such statistical procedures are called robust.

2.1. Nonadaptive estimators

In 1964, Huber introduced M-estimates, which are flexible and can be generalized to multiparameter problems.

Any estimate T_n which minimizes $\Sigma\rho(x_i;T_n)$ where ρ is an arbitrary function is called an M-estimate.

Simplified versions, as location estimates, involve $\rho(x,T)$ of the form $\rho(x-T)$ for some function ρ, with $\rho(0) = 0$, $\rho(x) \geq 0$ for all x. Many nonadaptive robust estimates are M-estimates.

As examples of M-estimates;

(i) If $\rho(x) = x^2$, the corresponding estimator is the sample mean.

(ii) If $\rho(x) = |x|$ the corresponding estimator is the sample median.

(iii) In Huber's (1964) M-estimate

$$\rho(x) = \begin{cases} \frac{1}{2}x^2 & \text{if } |x| \leq k \\ k|x| - \frac{1}{2}k^2 & \text{if } |x| > k, \end{cases} \tag{2.1.1}$$

and the corresponding estimator is closely related to Winsorizing.

(iv) if

$$\rho(x) = \begin{cases} \frac{1}{2}x^2 & \text{if } |x| \leq k \\ \frac{1}{2}k^2 & \text{if } |x| > k, \end{cases}$$

the corresponding estimator is closely related to a trimmed mean.

As a multiparameter estimate, the choice

$$\rho(x;\theta) = - \log g(x;\theta)$$

gives the ordinary MLE where the underlying distribution is g.

Two other commonly used robust estimators are the L and R estimators. An estimate is an L-estimator

(Linear combination of order statistics) if it is of the form

$$T_n = \Sigma w_{in} X_{(i)}, \quad \Sigma w_{in} = 1$$

and the $X_{(i)}$ are the order statistics. The trimmed mean $\bar{X}_\alpha$ corresponds to

$$w_{in} = \lambda(i/(n+1)) \quad \text{where}$$

$$\lambda(t) = \begin{cases} (1-2\alpha)^{-1} & \text{if } \alpha < t < 1-\alpha \\ 0 & \text{if } t \leq \alpha \text{ or } t \geq 1-\alpha. \end{cases}$$

An estimator is an R-estimator if it is of the form

$$T_n = \text{median } \{w_{jk} \cdot X_{jk}\} \text{ where } (j=1,\ldots,n,\ k=j,\ldots,n)$$

$$w_{jk} = d_{n-(k-j)} / \sum_{i=1}^{n} i d_i, \quad d_i \geq 0 \text{ for all } i, \quad X_{jk} = (X_{(j)} + X_{(k)})$$

The Hodges-Lehmann estimator corresponds to $d_1 = \ldots = d_n = 1$.

These estimators can be modified a bit to be adaptive. For example, for the M-estimator, one may replace $\rho(x)$ by $\rho(x/s)$ where s is an estimate of scale

based on the data. The L and R estimators are scale invariant so no such modification is needed for these estimators. The usual meaning of "adaptive" as applied to an estimator is that the form of the estimator adopts according to the shape of the sample distribution, not merely the scale.

2.2. Adaptive estimation

In 1956, Stein published a paper which dealt with the problem of estimating and testing hypotheses about a parameter θ. The question he asked was "when can one estimate θ as well as asymptotically not knowing the true distribution of a population as knowing the true distribution." Stein gave a simple necessary condition for several important examples and he indicated a procedure for testing whether a center of symmetry has a specified value that should work.

Consider estimators $\hat{\theta}_n$ of the location parameter θ based on a sample $(X_1, X_2, \ldots, X_n)$ from an unknown distribution $G(x-\theta)$ which is symmetric about the origin, and has density $g(x-\theta)$. We can divide the previous literature on adaptive estimation methods into two main streams.

One stream of research on adaptive estimation finds an estimator $\hat{\theta}_n$ of θ such that $\hat{\theta}_n$ is asymptotically efficient under the model. I.e.,

$$\mathcal{L}(n^{\frac{1}{2}}(\hat{\theta}_n-\theta)) \to N(0, I^{-1}) \quad \text{as} \quad n \to \infty \tag{2.2.1}$$

where I denotes the Fisher information on θ from the distribution $G(x-\theta)$.

Takeuchi (1971) considers a fictitious random subsample of size k drawn from the original sample and constructs the best linear estimator based on the subsample. Since he estimates the variance-covariance matrix of the order statistics of the subsample, this method can be classed as an adaptive estimate.

Stone (1975) takes any estimator $\bar{\theta}_n$ of θ which satisfies $n^{\frac{1}{2}}(\bar{\theta}_n-\theta) = O_p(1)$ as $n \to \infty$. By using a nonparametric estimate of $L(x) = g'(x)/g(x)$, he imitates a single step of a Newton-Raphson iteration solving $\sum_{i=1}^{n} L(X-\hat{\theta}_n) = 0$ with $\bar{\theta}_n$ as the initial approximation. Since $L(x)$ is estimated from the data, $\hat{\theta}_n$ is an adaptive estimate resembling the MLE.

Van Eeden (1970) and Beran (1974) estimate $\phi(u,g) = - g'[G^{-1}(u)]/g[G^{-1}(u)]$ which provides the

approximate scores for the best linear rank statistic, using a window scheme and a Fourier transform respectively. We call these methods adaptive estimate because the $\phi(u,g)$'s are calculated from the data.

Beran (1978) estimates the density from the data in the sense of nonparametric minimum Hellinger distance. Using the estimated density, he estimates the location parameter θ.

Though all of these foregoing methods have very desirable mathematical properties, they are very hard to implement and require many calculations. Also, attainment of their asymptotic behavior seems to require very large sample sizes.

The adaptive estimation literature contains a second stream of papers describing methods which are not fully asymptotically efficient, but which are easy to implement and which require relatively small amounts of calculations.

Hogg (1974) (modified by De Wet and Van Qyk (1979), Harter et al. (1979)) uses the trimmed mean in the special symmetric case and shows how to select the amount of trimming. Since these methods use the trimmed mean, they do not satisfy the condition(2.2.1) of asymptotic efficiency except in the rare case in which a trimmed mean is asymptotically efficient.

2.3. Why NPS estimation?

Beran's, Stein's and similar methods have excellent asymptotic properties, but are not practical for everyday use. Hogg's method is easy to implement but it has less desirable theoretical properties. NPS estimation lies between these approaches. We might say that Hogg's method is discrete and restricted. (Since it is developed by considering only a few possible underlying distributions) The NPS estimate selects from a continuous range of distribuitonal shapes measured by a shape parameter γ and can adapt to a wide range of sample tail behaviors.

Except for Beran's (1978) method, only the NPS estimate suggests the rough shape of the underlying distribution.

Chapter 3

NPS (Normal Pareto Spline) Distributions

3.1. Definition of NPS distributions

A random variable X has <u>standard NPS</u> (Normal Pareto Spline) <u>distribution</u> with tail parameter γ, if it has a density of the form

$$f_0(x,\gamma) = \begin{cases} e^{ax^2+b} & \text{if } |x| \leq 1 \\ \frac{1}{10c}\{1+\frac{\gamma}{c}(|x|-1)\}^{-\frac{1}{\gamma}-1} & \text{if } 1 < |x| \leq A, \quad \neq 0 \end{cases} \tag{3.1.1}$$

If $\gamma > 0$ then A is ∞, and if $\gamma < 0$ then $A = 1 - \frac{c}{\gamma}$.
If $\gamma = 0$, then $A = \infty$ and

$$f_0(x,0) = \frac{1}{10c}\, e^{-\frac{1}{c}(|x|-1)} \quad \text{if } 1 < |x| \tag{3.1.1'}$$

The parameters a,b and c depend on γ, and are determined by the requirement that $\Pr\{|X| > 1\} = 0.2$ and the spline constraints that the density and the first derivative of the density are continuous everywhere. Thus the parameters a,b and c satisfy the following spline equations:

$$f_0(1^+) = f_0(1^-) \tag{3.1.2}$$

or

$$a + b = - \log (10c)$$

and

$$f_0'(1^+) = f_0'(1^-) \tag{3.1.3}$$

or

$$2ac = - (1+\gamma)$$

and

$$\int_{-1}^{1} e^{ax^2+b} dx = 0.8 \tag{3.1.4}$$

In addition, we often consider the family of variables $Y = \mu + \tau X$, where X has the standard NPS distribution. Then Y has the NPS distribution $NPS(\mu,\tau,\gamma)$, with center at μ and interdecile range equal to 2τ. An illustration appears in Fig. 3.1.5.

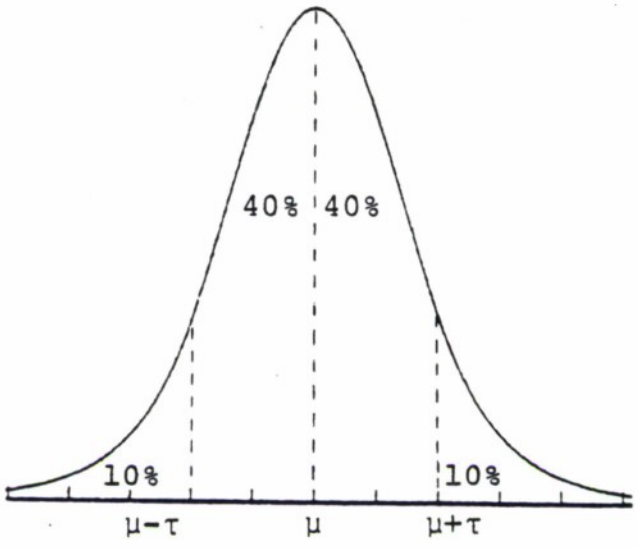


Fig. 3.1.5 Schematic illustration of the NPS distribution (density shown has $\gamma = 0$)

$\mu-\tau$: 10^{th} percentile

μ : 50^{th} percentile (median)

$\mu+\tau$: 90^{th} percentile

The three primary parameters are μ, the center of symmetry; τ the scale parameter; and γ, the tail shape parameter. The other parameters a,b,c are determined implicitly by γ, and are tabulated in Table 3.1.6.

Table 3.1.6 a,b,c as functions of γ

γ	a	b	c
-0.5	-0.6410	-0.7210	0.3900
-0.4	-0.7242	-0.6971	0.4142
-0.3	-0.7997	-0.6766	0.4377
-0.2	-0.8687	-0.6583	0.4604
-0.1	-0.9324	-0.6417	0.4826
0	-0.9915	-0.6265	0.5043
0.1	-1.0466	-0.6126	0.5255
0.2	-1.0983	-0.5998	0.5463
0.3	-1.1469	-0.5878	0.5667
0.4	-1.1927	-0.5767	0.5868
0.5	-1.2364	-0.5663	0.6066
0.6	-1.2778	-0.5565	0.6261
0.7	-1.3172	-0.5473	0.6453
0.8	-1.3549	-0.5386	0.6643
0.9	-1.3910	-0.5303	0.6830
1.0	-1.4256	-0.5225	0.7015
1.1	-1.4588	-0.5150	0.7198
1.2	-1.4907	-0.5079	0.7379
1.3	-1.5215	-0.5011	0.7558
1.4	-1.5513	-0.4946	0.7736
1.5	-1.5800	-0.4883	0.7912

The central portion of the density is Gaussian and the tails represent a reparametrization of Pareto densities. (See DuMouchel (1983)).

We will usually restrict ourselves to $\gamma > -0.5$ because for $\gamma \leq -0.5$, $f(x)$ does not have continuous derivatives at the end points $\mu \pm \tau \cdot A$.

A family of densities which are Gaussian in the middle and have a variety of tail behaviors are useful, realistic models for many kinds of data. Having heavy tails (γ larger than 0) allows us to model outlier-prone data, since, if X has an NPS distribution with $\gamma > 0$, then $E\{|X|^p\} = \infty$ for $p \geq \frac{1}{\gamma}$.

3.2. Tail behavior of the NPS distribution.

The family of NPS distributions can represent a diversity of tail behavior. (See Fig. 3.2.2). At $\gamma = 0$ we get exponential tails. Anscombe (1961) mentions that Generalized Pareto tails with $\gamma > 0$ can be generated by gamma mixtures of exponential distributions. When $\gamma < 0$, the distribution is truncated; when $\gamma = -1$, the uniform distribution on $(\mu-1.25\tau, \mu+1.25\tau)$ results, while $\gamma = 0.5$ leads to a triangular tail behavior. Since

$$f_0(x,\gamma) = \frac{1}{10c}\{1 + \frac{\gamma}{c}(x-1)\}^{-\frac{1}{\gamma}-1} \qquad \text{if} \quad 1 < x < A$$

$$f_0'(x,\gamma) = \frac{-(1+\gamma)}{10c^2}\{1 + \frac{\gamma}{c}(x-1)\}^{-\frac{1}{\gamma}-2} \qquad \text{if} \quad 1 < x < A$$

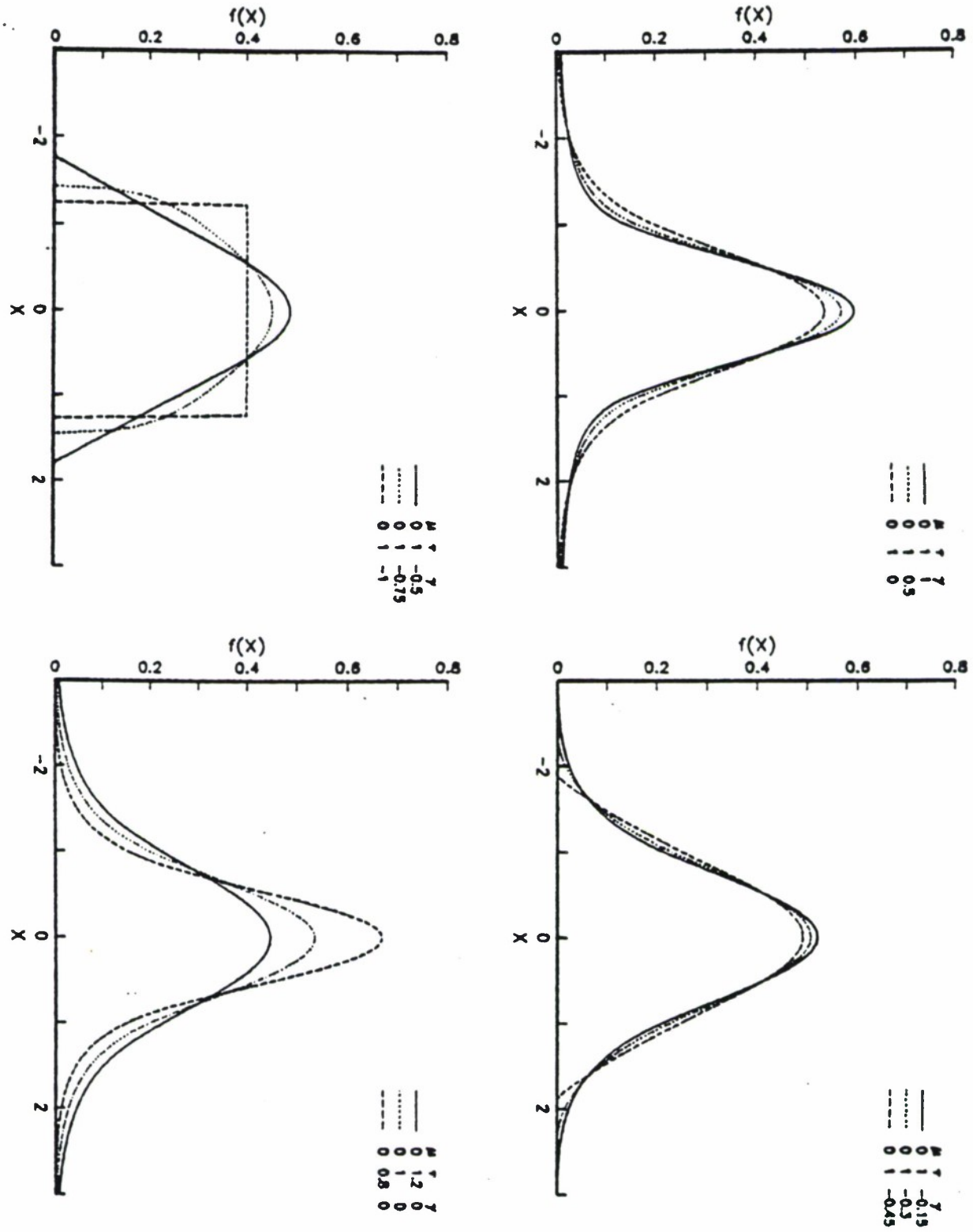
and

$$f_0''(x,\gamma) = \frac{(1+\gamma)(1+2\gamma)}{10c^3}\{1 + \frac{\gamma}{c}(x-1)\}^{-\frac{1}{\gamma}-3} \qquad \text{if} \quad 1 < x < A$$

tail behavior is as described in Table 3.2.1. Some graphs of NPS distributions are presented in Fig. 3.2.2.

Table 3.2.1 The tail behavior of NPS(0,1,γ) as a function of γ where $1 < x < A$.

Range of γ	$f_0'(x\gamma)$	$f_0''(x,\gamma)$	Tail behavior	Support
$0 \leq \gamma$	-	+	↘	Infinite
$-0.5 < \gamma < 0$	-	+	↘	Finite
$\gamma = -0.5$	-	0	↘	Finite
$-1 < \gamma < -0.5$	-	-	↘	Finite
$\gamma = -1$	0	0	→	Finite



DIFFERENT NPS DISTRIBUTIONS

Fig. 3.2.2

3.3. Moments of the NPS(0,1,γ) distribution

Let the random variable X have a standard NPS distribution. If $\gamma < \frac{1}{k}$, $E(X^k)$ exists. Since f is symmetric around 0,

$$E\{X^k\} = 0 \quad \text{if} \quad k \quad \text{is odd, and} \quad \gamma < \frac{1}{k}$$

So the mean is 0 (if $\gamma < 1$), and the skewness is 0 (if $\gamma < \frac{1}{3}$).

$$\text{Var}(X) = EX^2 = \int_{-1}^{1} x^2 \cdot e^{ax^2+b} dx + 2 \cdot \int_{1}^{A} x^2 \cdot \frac{1}{10c}\{1 + \frac{\gamma}{c}(x-1)\}^{-\frac{1}{\gamma}-1} dx$$

Integrating by parts, we have

$$\int_{-1}^{1} x^2 e^{ax^2+b} dx = \frac{1-4c}{10a \cdot c}$$

$$\int_{-1}^{1} x^2 \cdot \frac{1}{10c}\{1 + \frac{\gamma}{c}(x-1)\}^{-\frac{1}{\gamma}-1} dx = \begin{cases} \frac{1}{10} - \frac{c}{5(\gamma-1)} + \frac{c^2}{5(\gamma-1)(2\gamma-1)} & \text{if } \gamma < \frac{1}{2} \\ \infty & \text{if } \gamma \geq \frac{1}{2} \end{cases}$$

and finally

$$\text{Var}(X) = \begin{cases} \frac{1-4c}{10ac} + \frac{1}{5} - \frac{2c}{5(\gamma-1)} + \frac{2c^2}{5(\gamma-1)(2\gamma-1)} & \text{if } \gamma < \frac{1}{2} \\ \infty & \text{if } \gamma \geq \frac{1}{2} \end{cases} \qquad (3.3.1)$$

$$EX^4 = \int_{-1}^{1} x^4 \cdot e^{ax^2+b} dx + 2 \cdot \int_{1}^{A} x^4 \cdot \frac{1}{10c}\{1 + \frac{\gamma}{c}(x-1)\}^{-\frac{1}{\gamma}-1} dx;$$

$$\int_{-1}^{1} x^4 e^{ax^2+b} dx = \frac{2a-3(1-4c)}{20a^2c};$$

$$\int_{1}^{A} x^4 \frac{1}{10c}\{1 + \frac{\gamma}{c}(x-1)\}^{-\frac{1}{\gamma}-1} dx =$$

$$\frac{1}{10} - \frac{4c}{10(\gamma-1)} + \frac{12c^2}{10(\gamma-1)(2\gamma-1)} - \frac{24c^3}{10(\gamma-1)(2\gamma-1)(3\gamma-1)}$$

$$+ \frac{24c^4}{10(\gamma-1)(2\gamma-1)(3\gamma-1)(4\gamma-1)} \qquad \text{if } \gamma < \frac{1}{4}$$

$$\infty \qquad \text{if } \gamma \geq \frac{1}{4};$$

Finally

$$EX^4 = \frac{2a-3(1-4c)}{20a^2c} + \frac{1}{5} - \frac{4c}{5(\gamma-1)} + \frac{12c^2}{5(\gamma-1)(2\gamma-1)}$$

$$- \frac{24c^3}{5(\gamma-1)(2\gamma-1)(3\gamma-1)} + \frac{24c^4}{5(\gamma-1)(2\gamma-1)(3\gamma-1)(4\gamma-1)} \qquad \text{if } \gamma < \frac{1}{4} \qquad (3.3.2)$$

$$= \infty \qquad \gamma \geq \frac{1}{4},$$

and the kurtosis of X is given by

$$k = \frac{EX^4}{\{\text{Var } X\}^2}$$

The variance and kurtosis are tabulated in Table 3.3.3.

3.4 Information matrix for the NPS(μ,σ,γ) distribution

The random variable $Y = \mu + \tau X$ has density $f(y,\theta) = \frac{1}{\tau} f(\frac{y-\mu}{\tau},\gamma)$. By definition, the information matrix I is given by

$$I(\mu,\tau,\gamma) = \begin{bmatrix} \int \left(\frac{\partial \log f}{\partial \mu}\right)^2 f\,dx & \int \left(\frac{\partial \log f}{\partial \tau}\right)\left(\frac{\partial \log f}{\partial \mu}\right) f\,dx & \int \left(\frac{\partial \log f}{\partial \gamma}\right)\left(\frac{\partial \log f}{\partial \mu}\right) f\,dx \\ \int \left(\frac{\partial \log f}{\partial \mu}\right)\left(\frac{\partial \log f}{\partial \tau}\right) f\,dx & \int \left(\frac{\partial \log f}{\partial \tau}\right)^2 f\,dx & \int \left(\frac{\partial \log f}{\partial \gamma}\right)\left(\frac{\partial \log f}{\partial \tau}\right) f\,dx \\ \int \left(\frac{\partial \log f}{\partial \mu}\right)\left(\frac{\partial \log f}{\partial \gamma}\right) f\,dx & \int \left(\frac{\partial \log f}{\partial \tau}\right)\left(\frac{\partial \log f}{\partial \gamma}\right) f\,dx & \int \left(\frac{\partial \log f}{\partial \gamma}\right)^2 f\,dx \end{bmatrix}$$

We restrict ourselves to $\gamma > -0.5$, because $I_{\mu\mu} = \infty$ for $\gamma \leq -0.5$. Since $I(\mu,\tau,\gamma) = \tau^{-2} I(0,1,\gamma)$, we evaluate I for $\mu = 0$ and $\tau = 1$.

$$\frac{\partial \log f}{\partial \mu} = \begin{cases} -2ay & \text{if } |y| \leq 1 \\ \dfrac{1+\gamma}{c\{1 + \frac{\gamma}{c}(|y|-1)\}} \cdot \operatorname{sgn}(y) & \text{if } 1 < |y| < A \end{cases} \qquad (3.4.1)$$

$$\frac{\partial \log f}{\partial \tau} = -1 + y \cdot \frac{\partial \log f}{\partial \mu} \qquad (3.4.2)$$

and

$$\frac{\partial \log f}{\partial \gamma} = \begin{cases} \frac{\partial a}{\partial \gamma} y^2 + \frac{\partial b}{\partial \gamma} & \text{if } |y| \leq 1 \\ -\frac{1}{c}\frac{\partial c}{\partial \gamma} + \gamma^{-2}\log\{1+\frac{\gamma}{c}(|y|-1)\} & \\ \quad -\frac{(1+\gamma^{-1})(|y|-1)}{\{1+\frac{\gamma}{c}(|y|-1)\}}(c^{-1}-\gamma\cdot c^{-2}\frac{\partial c}{\partial \gamma}) & \text{if } 1 < |y| < A \end{cases} \tag{3.4.3}$$

Since $\frac{\partial \log f}{\partial \mu}$ is an odd function of y and $\frac{\partial \log f}{\partial \tau}$ and $\frac{\partial \log f}{\partial \gamma}$ are even, $I_{\mu\tau} = I_{\mu\gamma} = 0$ and

$$I(\mu,\tau,\gamma) = \begin{vmatrix} I_{\mu,\mu} & 0 & 0 \\ 0 & I_{\tau,\tau} & I_{\tau,\gamma} \\ 0 & I_{\gamma,\tau} & I_{\gamma,\gamma} \end{vmatrix}$$

In particular,

$$I_{\mu,\mu} = -\frac{2}{\tau^2}\{0.8a + \frac{\gamma(1+\gamma)}{10c^2(1+2\gamma)}\} \tag{3.4.4}$$

$$= \frac{1+\gamma}{\tau^2}\{\frac{4}{5c} - \frac{\gamma}{5c^2(1+2\gamma)}\}$$

where c is a function of γ tabulated in Table 3.1.6,

Table 3.4.1 Information matrix

γ	$I_{\mu,\mu}$	$I_{\tau,\tau}$	$I_{\tau,\gamma}$	$I_{\gamma,\gamma}$
-0.4	2.557	4.581	3.015	3.6
-0.3	1.828	2.876	1.581	2.359
-0.2	1.642	2.04	0.764	1.344
-0.1	1.588	1.636	0.4072	0.8807
0	1.586	1.403	0.2215	0.6224
0.1	1.608	1.254	0.1067	0.4273
0.2	1.642	1.151	0.02767	0.2766
0.3	1.683	1.078	-0.02421	0.1785
0.4	1.728	1.024	-0.05766	0.1182
0.5	1.774	0.985	-0.07814	0.08141
0.6	1.822	0.9562	-0.09034	0.0585
0.7	1.869	0.9351	-0.09738	0.04387
0.8	1.917	0.9197	-0.1008	0.03425
0.9	1.964	0.9086	-0.102	0.02772
1	2.01	0.9008	-0.1015	0.02314
1.1	2.055	0.8955	-0.1002	0.01982
1.2	2.1	0.8923	-0.09823	0.01733
1.3	2.144	0.8906	-0.09583	0.0154
1.4	2.186	0.8902	-0.09359	0.01388
1.5	2.228	0.8908	-0.0909	0.01263

Table 3.4.2 Asymptotic standard deviations and correlation coefficient of MLE based on inverse of Information matrix

γ	σ_μ	σ_τ	$\rho_{\tau,\gamma}$	σ_γ
-0.4	0.6253	0.6974	-0.7424	0.7867
-0.3	0.7397	0.7418	-0.6068	0.819
-0.2	0.7805	0.7891	-0.4613	0.972
-0.1	0.7934	0.8312	-0.3392	1.133
0	0.794	0.8689	-0.237	1.305
0.1	0.7886	0.9025	-0.1457	1.546
0.2	0.7803	0.9331	-0.04904	1.904
0.3	0.7708	0.9648	0.0552	2.37
0.4	0.7608	1.002	0.1657	2.949
0.5	0.7507	1.048	0.2759	3.646
0.6	0.7409	1.107	0.382	4.474
0.7	0.7314	1.179	0.4808	5.445
0.8	0.7223	1.267	0.568	6.565
0.9	0.7136	1.37	0.6428	7.841
1	0.7054	1.481	0.7029	9.242
1.1	0.6975	1.604	0.7523	10.78
1.2	0.6901	1.727	0.79	12.39
1.3	0.683	1.843	0.8183	14.02
1.4	0.6763	1.965	0.842	15.73
1.5	0.6699	2.056	0.857	17.27

Chapter 4

The One Sample Location Problem: Estimation

4.1. Problem description

Let $y_1, y_2, \ldots, y_n$ denote a random sample from a continuous population with symmetric unimodal distribution function $G(y-\mu)$. In this chapter we shall deal with the problem of estimating the location parameter μ. Our estimator will be that derived from computing the maximum likelihood estimate of $\theta = (\mu, \tau, \gamma)'$ based on the assumption that the distribution belongs to the NPS family. We shall call this the NPS estimate. By estimating γ we describe the tail behavior of the distribution g. Inasmuch as this estimate of γ affects our procedure for estimating the location parameter μ, we may regard the NPS estimate of μ as adaptive.

For large samples one should expect the NPS estimate to be close to that value of θ that corresponds to the NPS distribution that is closest to G. However there are several notions of a closest distribution which may be considered and we shall describe three in the next section.

4.2. Three concepts of closest NPS distribution

The first concept we introduce is that of the NPS distribution with the same median, variance and

90-th percentile as $G(x-\mu)$. Since these parameters of the NPS distributions are μ, τ^2 multiplied by Var (X) derived in section 3.3, and $\mu+\tau$ respectively, these matching conditions can be used to determine $\theta_M = (\mu_M, \tau_M, \gamma_M)'$. This concept seems to be rather naive. It is based on some arbitrary choices such as the 90-th percentile and is unlikely to have sound theoretical justification. Moreover θ_M is not defined if the variance of Y is infinite.

A second concept derives from the fact that under suitable regularity conditions, the NPS estimator will converge to the NPS distribution which is closest in the Kullback-Leibler sense. That is, we select θ_{KL} to minimize

$$I(g_\mu, f_\theta) = \int g(y-\mu) \log [\frac{g(y-\mu)}{f(y|\theta)}] dy \qquad (4.2.1)$$

where $f_\theta(y) = f(y|\theta)$ represents the density of the NPS(μ,τ,γ) distribution and g_μ the density $g(y-\mu)$ of $G(y-\mu)$.

This concept introduces some difficulties. For $\gamma < 0$, f is a distribution of bounded range and $I(g_\mu, f_\theta) = \infty$ if g_μ has infinite support. Thus, even though for some $\gamma < 0$, f_θ may resemble g_μ very closely, f_θ will not be

a candidate for the closest NPS distribution to g_μ. In particular, numerical calculations in Table 4.2.2 demonstrate that the closest NPS distribution in the Kulback-Leibler sense to the standard normal $N(0,1)$ is of the form $NPS(0,\tau_{KL},0)$ with $\tau_{KL} = 1.2508$. Also the closest for the standard logistic is $NPS(0,\tau_{KL},0)$ with $\tau_{KL} = 2.171$. (See Table 4.2.3)

As we shall see, simulations of NPS estimates from $N(0,1)$ data yield values of $\hat{\gamma}$ around -0.2. In contrast to θ_{KL}, our first relatively naive concept gives $(\mu_M,\tau_M,\gamma_M) = (0,1.282,-.218)$ as the parameter of the closest NPS estimator. Thus θ_M seems to be more relevant than θ_{KL}.

Finally we introduce a third concept. Let $y_{i,n}$ be expected i^{th} order statistics among n samples of $G(y-\mu)$ for $1 \leq i \leq n$. The closest distribution will be $NPS(\tilde{\mu}_n,\tilde{\tau}_n,\tilde{\gamma}_n)$ where $\tilde{\theta}_n = (\tilde{\mu}_n,\tilde{\tau}_n,\tilde{\gamma}_n)$ is the NPS estimate based on the synthetic (nonrandom) sample $y_{1,n},y_{2,n},\ldots,y_{n,n}$. The vector $\tilde{\theta}_n$ will be called the synthetic parameter. It is clear that $\tilde{\theta}_n$ depends on n and one would expect $\tilde{\gamma}_n$ to converge to 0 as $n \to \infty$ if g_μ corresponds to $N(0,1)$ or $L(0,1)$. However we shall see that for moderately large n, $\tilde{\gamma}_n$ will be about

Table 4.2.2 $\max_{\tau} \int (\log f) \cdot g \, dx$ for various γ's

when $g \sim N(0,1)$

γ	τ	$\int (\log f) \cdot g dx$
0	1.25078	-1.43038
0.001	1.25078	-1.43046
0.002	1.25097	-1.43055
0.003	1.25097	-1.43063
0.005	1.25097	-1.43079
0.01	1.25117	-1.43121
0.02	1.25175	-1.43205
0.03	1.25234	-1.43289
0.04	1.25312	-1.43375
0.05	1.25390	-1.43461
0.1	1.25898	-1.43899
0.2	1.27304	-1.44789
0.3	1.29082	-1.45664
0.4	1.31074	-1.46504
0.5	1.33222	-1.47298

Table 4.2.3 $\max_{\tau} \int (\log f) \cdot g \, dx$ for various γ's

when $g \sim$ logistic $(0,1)$

γ	τ	$\int (\log f) \cdot g dx$
0	2.17101	-2.00040
0.001	2.17082	-2.00041
0.002	2.17042	-2.00043
0.003	2.17023	-2.00044
0.005	2.16964	-2.00047
0.01	2.16828	-2.00056
0.02	2.16593	-2.00077
0.03	2.16398	-2.00102
0.04	2.16222	-2.00131
0.05	2.16066	-2.00163
0.1	2.15675	-2.00366
0.2	2.16164	-2.00911
0.3	2.17765	-2.01553
0.4	2.20050	-2.02230
0.5	2.22765	-2.02913

-0.2, -0.05 for the N(0,1) and L(0,1) respectively. In Table 4.2.4 and 4.2.5, we list the synthetic parameter $\tilde{\theta}_n$ based on the synthetic sample from the N(0,1) and L(0,1) respectively.

Table 4.2.4 Synthetic parameter $\tilde{\theta}_n$ based on N(0,1)

n	$\tilde{\mu}_n$	$\tilde{\tau}_n$	$\tilde{\gamma}_n$
20	0	1.364	-.761
50	0	1.297	-.391
100	0	1.289	-.305
500	0	1.281	-.227
1000	0	1.280	-.214

Table 4.2.5 Synthetic parameter $\tilde{\theta}_n$ based on L(0,1)

n	$\tilde{\mu}_n$	$\tilde{\tau}_n$	$\tilde{\gamma}_n$
20	0	2.292	-.481
50	0	2.208	-.187
100	0	2.195	-.113
500	0	2.186	-.054
1000	0	2.184	-.046

We shall refer to the parameters of the NPS distributions corresponding to these concepts as

(1) Variance and 90-th percentile matching or just plain matching, θ_M,

(2) closest Kulback-Leibler, θ_{KL}, and

(3) synthetic NPS parameter, $\tilde{\theta}_n$

In Table 4.2.6 we list three special distributions. These are the standard Gaussian, Logistic and Slash, distributions.

Table 4.2.6 Standard distributions

Name	Notation	Density	Support
Gaussian	N(0,1)	$\frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$	$-\infty < x < \infty$
Logistic	L(0,1)	$\frac{e^{-x}}{\{1 + e^{-x}\}^2}$	$-\infty < x < \infty$
Slash	S(0,1)	$\frac{1}{\sqrt{2\pi}} \frac{1 - e^{-\frac{x^2}{2}}}{}$	$-\infty < x < \infty$

In Figures 4.2.8-4.2.10 we present the densities of the closest NPS distributions to these standard distributions for each concept, and in Table 4.2.7, we tabulate the corresponding values of $\theta = (\mu,\tau,\gamma)'$.

Table 4.2.6 $\theta_M, \theta_{KL}, \tilde{\theta}_n$ for N(0,1), L(0,1), S(0,1)

For N(0,1)		μ	τ	γ
	θ_M	0	1.282	-0.218
	θ_{KL}	0	1.251	0
	$\tilde{\theta}_{1000}$	0	1.280	-0.214
For L(0,1		μ	τ	γ
	θ_M	0	2.197	-0.043
	θ_{KL}	0	2.171	0
	$\tilde{\theta}_{1000}$	0	2.184	-0.046
For S(0,1)		μ	τ	γ
	θ_M*	-	-	-
	θ_{KL}	0	3.335	1.246
	$\tilde{\theta}_{1000}$	0	3.335	1.246

*not available because Var[S(0,1)] = ∞

NPS APPROXIMATIONS FOR NORMAL DISTRIBUTION

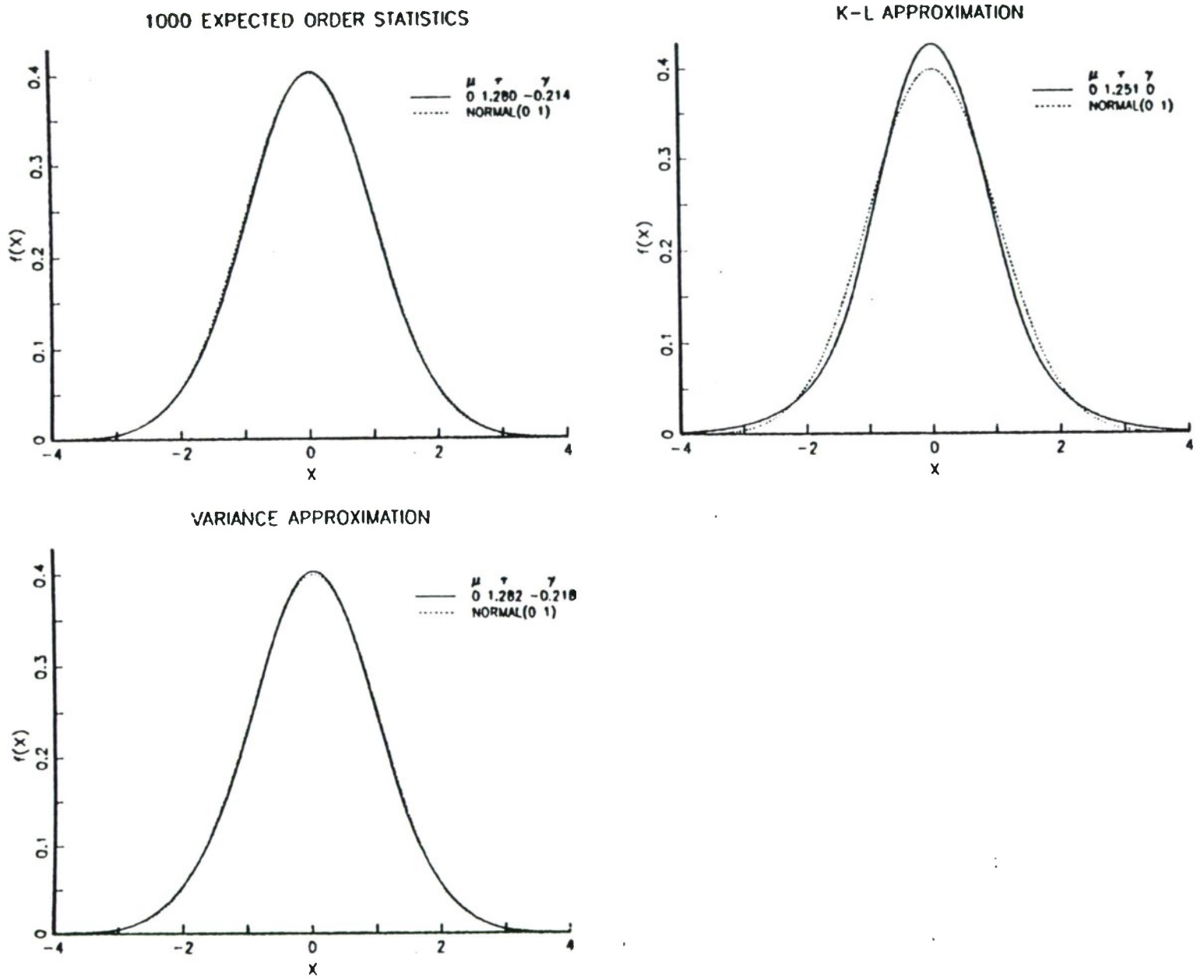
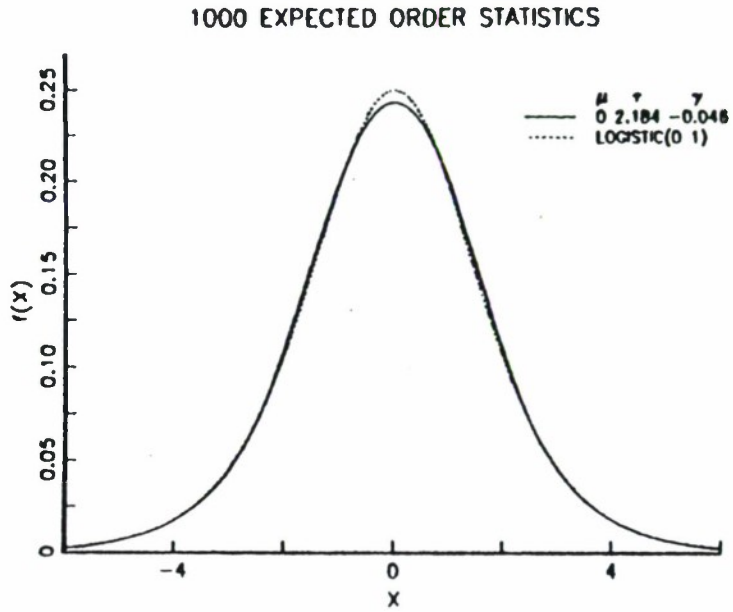
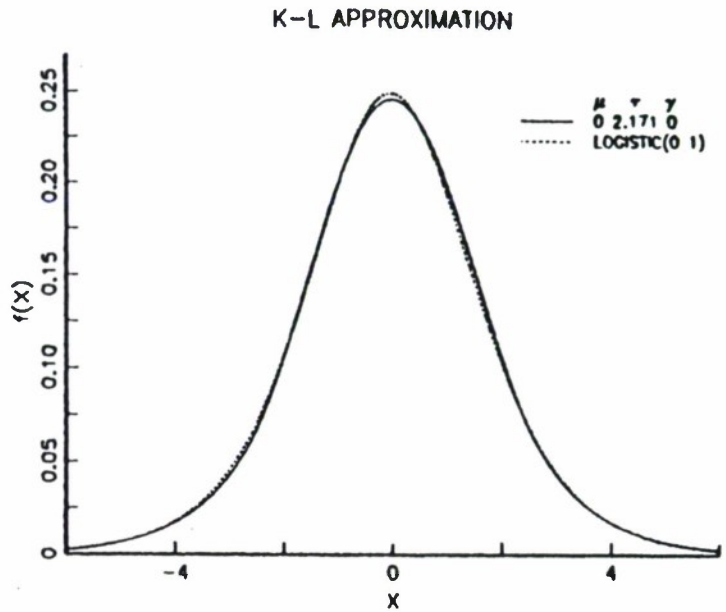


Fig. 4.2.8

NPS APPROXIMATIONS FOR LOGISTIC DISTRIBUTION





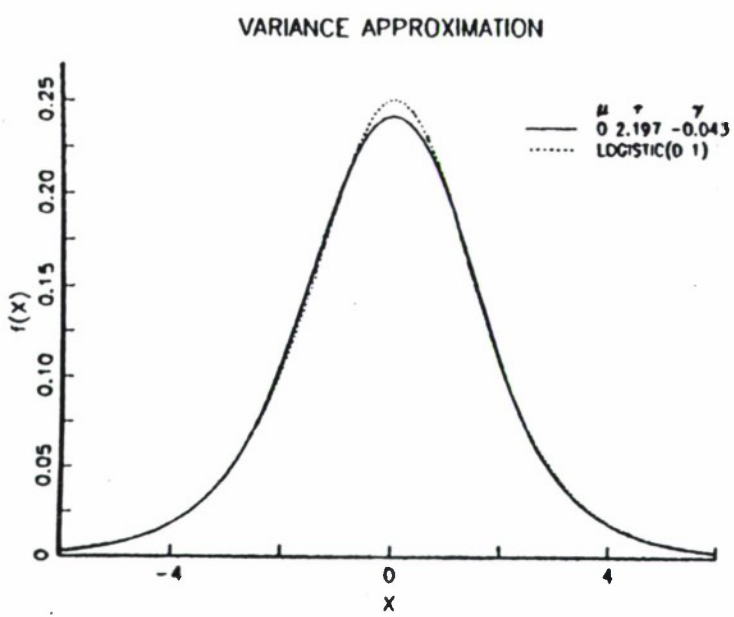


Fig. 4.2.9

NPS APPROXIMATIONS FOR SLASH DISTRIBUTION

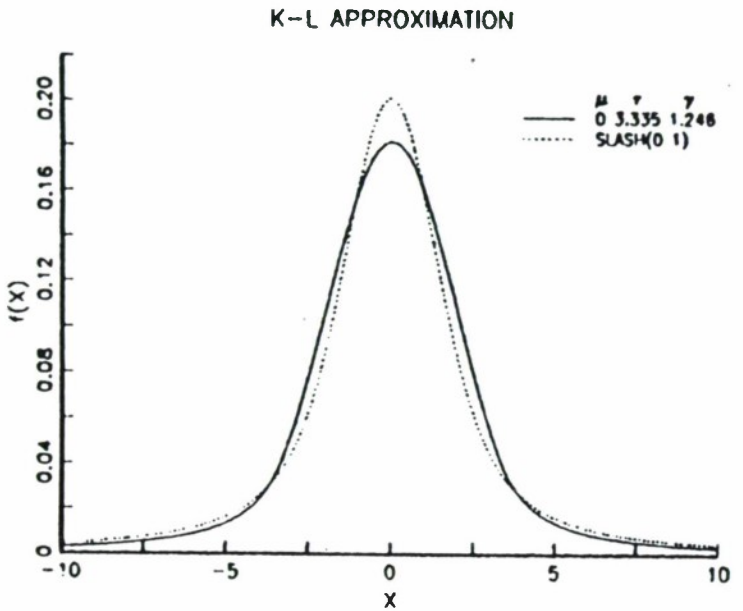


Fig. 4.2.10

When the underlying distribution is g, let $f(y;\theta_{KL})$ be the result from concept 2 (closest Kullback-Leibler) and let $f(y;\tilde{\theta}_{1000})$ be the result from concept 3 (synthetic NPS parameter, where $n = 1000$) of defining the "closest" NPS distribution to g. We are especially interested in the behaviors of γ_{KL}, $\tilde{\gamma}_{1000}$ for various commonly used distributions g which have infinite support. Table 4.2.11 lists some possible g's for various combinations of γ_{KL}'s and $\tilde{\gamma}_{1000}$'s.

Table 4.2.11 Classification of values of γ_{KL}'s and $\tilde{\gamma}_{1000}$'s for some distributions g which have infinite support

	sign of γ_{KL} +	0	-
+	NPS($\gamma > 0$) Slash Cauchy	Center : Normal** for tail : Cauchy	***
0	Center : NPS(γ=0) for tail : Cauchy	NPS(γ=0)	**
-	Center : Normal* for tail : Cauchy	Normal logistic	***

*NPS($\mu,\tau,0$) (Normal) center from minimum expected order statistics to maximum expected order statistics where n = 1000, with Cauchy for tails.

** c(<.999), c of Normal center and (1-c) of Cauchy far tails.

*** $\gamma_{KL} < 0$ is impossible if g has infinite support

4.3. The Huber M-estimator

We shall compare the NPS estimator with Huber's M-estimator. The M-estimator consists of selecting μ to minimizing $\Sigma\rho(y_i,\mu)$ or equivalently to set $\Sigma\psi(y_i,\mu) = 0$ where $\psi = \partial\rho/\partial\mu$. One typically is concerned with those examples where ρ and ψ may be written in the form $\rho(y-\mu)$ and $\psi(y-\mu)$ respectively. Then, it is known (Huber (1964)) that under regularity conditions on the symmetric distribution $G(y-\mu)$, the M-estimator T_n is consistent, and as $n \to \infty$, $L(\sqrt{n}(T_n-\mu)) \to N(0,\sigma_T^2)$ where

$$\sigma_T^2 = \frac{\int\psi^2(y)\cdot g(y)\,dy}{[\int\psi^1(y)\cdot g(y)\,dx]^2}$$

By Huber's M-estimator we mean the M-estimator where

$$\rho(y) = \begin{cases} y^2/2 & \text{if } |y| \leq k \\ k|y|-k^2/2 & \text{if } |y| > k \end{cases} \tag{4.3.1}$$

and

$$\psi(y) = \begin{cases} y & \text{if } |y| \leq k \\ k\cdot\text{sgn}\ (y) & \text{if } |y| > k \end{cases} \tag{4.3.2}$$

Then the asymptotic variance is given by

$$\sigma_H^2(g,k) = \frac{\int \min(y^2,k^2)\, g(y)\,dy}{\left[\int_{-k} g(y)\,dy\right]^2} \qquad (4.3.3)$$

This estimator may be regarded as a variation of the NPS estimator where the maximization with respect to $\theta = (\mu,\tau,\gamma)'$ is carried out subject to the restrictions $\tau = k$ and $\gamma = 0$. In this sense the NPS estimator is an adaptive generalization of Huber's M-estimator, where the data are used to estimate τ and γ. As we pointed out in section 2.1, it is not uncommon to use an adaptive version of the Huber estimator where the scale parameter is estimated.

For later comparisons, we tabulate the asymptotic variance of Huber's M-estimator for several values of k for normal, logistic and slash distributions in Table 4.3.4.

Table 4.3.4. Asymptotic variance of Huber's M-estimator

Distribution	Normal	logistic	slash
k			
.5	1.2625	3.4816	5.6867
1	1.1073	3.1947	5.4896
1.5	1.0371	3.0595	5.5961
2	1.0104	3.0178	5.9294
2.5	1.0023	3.0283	6.4283
3	1.0004	3.0637	7.0249
3.5	1.0001	3.1071	7.6922
4	1.0000	3.1490	8.3999

4.4. Asymptotic distribution of NPS estimates

If G has infinite support and the closest KL distribution to $G(y-\mu)$ has $\gamma > 0$, then from Huber (1967) we see that under mild regularity conditions,

$$L[\sqrt{n}(\hat{\theta}_n - \tilde{\theta}_{KL})] \to N(0, \Sigma(\theta_{KL}))$$

where $\hat{\theta}_n$ is the NPS estimator based on a sample of size n from G,

$$\Sigma(\theta) = B^{-1}AB^{-1}, \tag{4.4.1}$$

$$A = A(\theta) = E[(\frac{\partial \log f(Y,\theta)}{\partial\theta})(\frac{\partial \log f(Y,\theta)}{\partial\theta})'] \tag{4.4.2}$$

and

$$B = B(\theta) = -E(\frac{\partial^2 \log f(Y,\theta)}{\partial\theta^2}). \tag{4.4.3}$$

It should be noted that these expectations are with respect to the distribution G.

However the case where G is normal and $\gamma_{KL} = 0$ does not satisfy the regularity conditions. Indeed our calculations indicate that $\tilde{\theta}_n$ approaches 0 so slowly that it would be unreasonable to expect the limiting distribution of $\sqrt{n}(\hat{\theta}_n - \theta_{KL})$ to be normal. Instead we

shall present a heuristic derivation to the effect that $\sqrt{n}(\hat{\theta}_n-\tilde{\theta}_n)$ is asymptotically normal. This derivation, which follows, involves the expansion of the log of the likelihood about $\theta = \tilde{\theta}_n$ and $Y_{in} = y_{in}$ where Y_{in} are the order statistics and y_{in} are their expected values.

$$0 = \sum_{i=1}^{n} \frac{\partial \log f(Y_i,\hat{\theta}_n)}{\partial\theta}$$

$$= \sum_{i=1}^{n} \frac{\partial \log f(y_{in},\tilde{\theta}_n)}{\partial\theta} + A_{in} - n(\hat{\theta}_n-\tilde{\theta}_n)B_{in} + \text{higher order terms} \tag{4.4.4}$$

where

$$A_{in} = \sum_{i=1}^{n} \frac{\partial^2 \log f(y_{in},\tilde{\theta}_n)}{\partial\theta\partial y}(Y_{in}-y_{in}) = A_{in}(\tilde{\theta}_n) \tag{4.4.5}$$

and

$$B_{in} = -\frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2 \log f(Y_{in},\hat{\theta}_n)}{\partial\theta^2} = B(\tilde{\theta}_n)+O_p(1) = B(\theta_{KL})+O_p(1) \tag{4.4.6}$$

and the sum on the right hand side of (4.4.4) vanishes. Thus

$$\sqrt{n}(\hat{\theta}_n-\tilde{\theta}_n) = -B_{in}^{-1}\frac{1}{\sqrt{n}}A_{in} + O_p(1) \tag{4.4.7}$$

But A_{in} is a linear function of the order statistics,

and by the Chernoff, Gastwirth, Johns (1967) theorem, it is asymptotically normal. To be more specific, given a vector function $\underline{H}(y)$, the distribution of

$$T_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \underline{H}(y_{in})(Y_{in}-y_{in})$$

converges to $N(0,\Sigma^*)$ where

$$\Sigma^* = \text{cov } (\underline{C}(Y))$$

and

$$C(y) = \int H(y)\,dy$$

In our particular application

$$\underline{H}(y) = \frac{\partial^2 \log f(y,\tilde{\theta}_n)}{\partial y \partial \theta}$$

and

$$\underline{C}(Y) = \frac{\partial \log f(Y,\tilde{\theta}_n)}{\partial \theta}$$

Thus

$$L(\sqrt{n}(\hat{\theta}_n-\tilde{\theta}_n)) = N(0,\Sigma_{in}) \qquad (4.4.8)$$

where

$$\Sigma_{1n} = B_{1n}^{-1} A_{2n} B_{1n}^{-1} \tag{4.4.9}$$

and

$$A_{2n} = \frac{1}{n}\Sigma\left(\frac{\partial \log f(y_{in}, \tilde{\theta}_n)}{\partial \theta}\right)\left(\frac{\partial \log f(y_{in}, \tilde{\theta}_n)}{\partial \theta}\right)' \tag{4.4.10}$$

Since $\tilde{\theta}_n$ converges slowly to θ_{KL}, we tabulate $\tilde{\theta}_n$ and the elements of Σ_{1n} for various values of n. We include the limit $\Sigma(\theta_{KL})$. These tabulations appear in Table 4.4.11 for the normal and logistic where $\gamma_{KL} = 0$. For the slash distribution where $\gamma_{KL} > 0$, we simply present $\Sigma(\theta_{KL})$.

Table 4.4.1 Various asymptotic variances for Normal, Logistic and Slash distributions

Σ_{KL} for normal

1.0647	0	0
0	0.8668	-.1290
0	-.1290	0.2431

where $\theta_{KL} = 10, 1.251, 0)'$

Σ_{KL} for logistic

3.0116	0	0
0	3.3921	-.1997
0	-.1997	.4515

where $\theta_{KL} = (0,2.171,0)'$

Σ_{KL} for slash

5.375	0	0
0	20.4431	-.3318
0	-.3318	.0582

where $\theta_{KL} = (0,3.335,1.246)'$

$\Sigma_{1,1000}$ for normal

.9919	0	0
0	.8869	.2364
0	.2364	.4596

where $\tilde{\theta}_{1,1000} = (0, 1.280, -.214)'$

$\Sigma_{1,1000}$ for logistic

3.0110	0	0
0	3.3419	-.2763
0	-.2763	.3872

where $_{1,1000} = (0, 2.184, -.046)'$

4.5. The Normal-like distribution

We describe here another approach to analysing the asymptotic properties of the NPS estimator. This approach has some theoretical shortcomings, which are emphasized by the comparison of the theory with the simulations for moderately large sample size.

A major theoretical problem has been that the Kullback-Leibler information $I(g,f) = \infty$ for g with infinite support and f NPS with $\gamma < 0$. Our approach is to replace g when it has finite support by a distribution g_ε which "look like" g over most of its range but which differs in the far tails, in that it has finite support. Since g is close to g_ε one may hope that the NPS estimator applied to g would have similar properties to that when applied to g_ε. Since g_ε has finite support, the difficulty with the Kullback-Leibler information will be alleviated. For the theoretical comparison using the "look alike" distribution in simulations, we need to go through the following steps. Suppose t_ε is defined by

$$\int_{-\infty}^{t_\varepsilon} g(y)\,dy = 1 - \varepsilon/5 \quad \text{and} \quad A_\varepsilon = \tilde{\tau}_n\left(1 - \frac{\tilde{c}_n}{\tilde{\gamma}_n}\right) \quad \text{where} \quad n = \frac{1}{\varepsilon}.$$

STEP 1 Define g_ε as follows

$$g_\varepsilon(y) = \begin{cases} g(y) & |y| \leq t_\varepsilon \\ h_\varepsilon(y) & t_\varepsilon < |y| < A_\varepsilon \\ 0 & |y| \geq A_\varepsilon \end{cases}$$

As an example, we will take g to be the normal distribution and we call g_ε the normal-like distribution, indexed by the parameter ε, and schematically represented in Fig. 4.5.1.

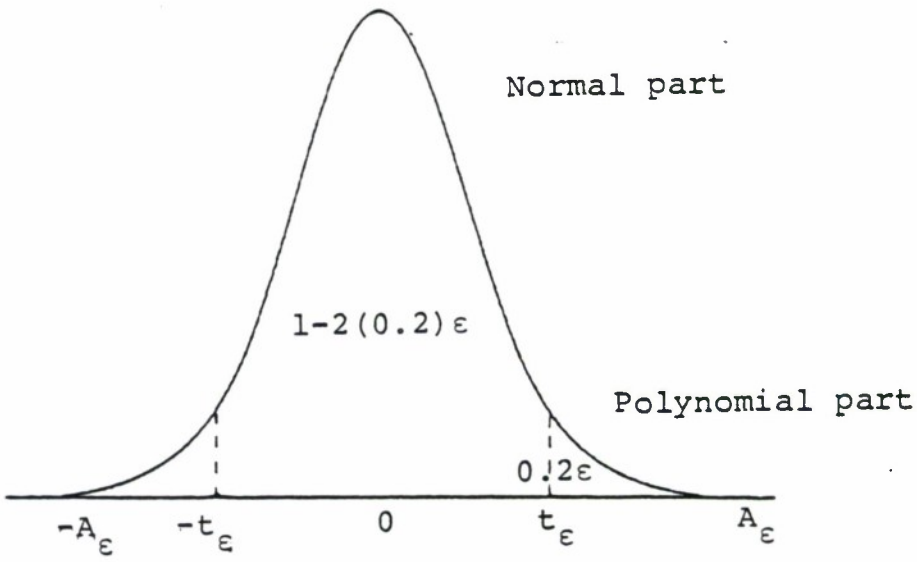


Fig. 4.5.1. Normal-like density

Here,

$$g(y) = \phi(y) \quad \text{where} \quad \phi(y) = \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}}$$

and

$$h_\varepsilon(y) = (A_\varepsilon - Y)^2\{a_\varepsilon + b_\varepsilon(A_\varepsilon - y) + c_\varepsilon(A_\varepsilon - y)^2\}$$

The conditions are that each "far tails" have probability 0.2ε and that the density and its first derivative be continuous at t_ε and A_ε. Here t_ε is $\Phi(1-0.2\varepsilon)$. The parameters t_ε, A_ε, a_ε, b_ε and c_ε are tabulated in Table 4.5.2.

Table 4.5.2 Parameters of the Normal-like distribution

ε	t_ε	A_ε	a_ε	b_ε	c_ε
0.002	3.353	3.841	0.0454	0.1566	0.1558
0.001	3.540	4.011	0.0251	0.0899	0.0930

STEP 2 Calculate $\theta_{KL\varepsilon} = (0, \tau_{KL\varepsilon}, \gamma_{KL\varepsilon})'$, the parameter of the NPS $g_{KL\varepsilon}$ which is closest to g_ε in the Kullback-Leibler sense.

To do so we maximize

$$\int (\log f)\cdot g_\varepsilon dy = \int_{-\tau}^{\tau} \{a(\tfrac{y}{\tau})^2 + b - \log \tau\}\cdot\phi(y)dy$$

$$+ 2\int_{\tau}^{t_\varepsilon} \phi(y)[\log \frac{1}{10\tau c} - (1 + \tfrac{1}{\gamma})\log \{1 + \tfrac{\gamma}{c}(\tfrac{y}{\tau} - 1)\}]dy$$

$$= (\frac{a}{\tau^2} + b - \log \tau)\{2\Phi(\tau) - 1\} - 2\cdot\frac{a}{\tau}\phi(t)$$

$$+ 2 \log \frac{1}{10\tau c}\{\Phi(t_\varepsilon) - \Phi(\tau) + .2\varepsilon\}$$

$$- 2(1 + \tfrac{1}{\gamma})\int_{\tau}^{t_\varepsilon} \phi(y)\log \{1 + \tfrac{\gamma}{c}(\tfrac{y}{\tau} - 1)\}]dy$$

$$- 2(1 + \tfrac{1}{\gamma})\int_{t_\varepsilon}^{A_\varepsilon} [(A_\varepsilon - y)^2\{a_\varepsilon + b_\varepsilon(A_\varepsilon - y) + c_\varepsilon(A_\varepsilon - y)^2\}]\cdot\log \{1 + \tfrac{\gamma}{c}(\tfrac{y}{\tau} - 1)\}dy$$

Table 4.5.3 lists $\tau_{KL\varepsilon}$, $\gamma_{KL\varepsilon}$ which were computed for various values of ε.

Table 4.5.3 $\tau_{KL\varepsilon}$, $\gamma_{KL\varepsilon}$ which minimizes $I(g_\varepsilon, f)$ for various ε's.

ε	$\tau_{KL\varepsilon}$	$\gamma_{KL\varepsilon}$
0.002	1.309	-.106
0.001	1.291	-.127

STEP 3 Calculate asymptotic variance of $f_{KL\varepsilon}$ when underlying distribution is g_ε.

From Huber (1967) we see that for very large samples from g_ε, the asymptotic distribution of the NPS estimator $\hat{\theta}_{n\varepsilon}$ satisfies

$$\mathcal{L}(\sqrt{n}(\hat{\theta}_{n\varepsilon} - \theta_{KL\varepsilon})) \to N(0, \Sigma_\varepsilon) \tag{4.5.4}$$

where

$$\Sigma_\varepsilon = B_\varepsilon^{-1} A_\varepsilon B_\varepsilon \tag{4.5.5}$$

$$A_\varepsilon = E\left[\left(\frac{\partial \log f(Y_\varepsilon, \theta_{KL\varepsilon})}{\partial \theta}\right)\left(\frac{\partial \log f(Y_\varepsilon, \theta_{KL\varepsilon})}{\partial \theta}\right)'\right] \tag{4.5.6}$$

and

$$B_\varepsilon = -E\left[\frac{\partial^2 \log f(Y_\varepsilon, \theta_{KL\varepsilon})}{\partial \theta^2}\right] \tag{4.5.7}$$

and these expectations are with respect to the distribution g_ε of Y_ε. In Table 4.5.8 the asymptotic variance of $\hat{\mu}_{n\varepsilon}$ are tabulated.

Table 4.5.8 . The asymptotic variance of $\hat{\mu}_{NPS}$ when underlying distribution is g_ε

ε	n	A.V. of $\hat{\mu}_{NPS}$
.002	500	1.064
.001	1000	1.045

In simulations which use samples of $n = 1000$ drawn from a normal population, the variance of $\hat{\mu}_{NPS}$ is 1.013. The asymptotic theory of the Normal-like distribution slightly over estimated the asymptotic variance. But if we try various kinds of constant multiply by ε as the tail area, then we will get better approximation.

4.6. Sensitivity and Influence curves

The study of robustness involves consideration of sensitivity to outliers. The sensitivity curve of the estimator T_n is defined by

$$SC(y;y_1,y_2,\ldots,y_n,T_n) =$$

$$(n+1)\{T_{n+1}(y_1,\ldots,y_n,y) - T_n(y_1,\ldots,y_n)\} \qquad (4.6.1)$$

where the y_i's are the observations. The sensitivity curve describes the effect of an additional observation at y. An estimator with a high resistance to outliers will have a low sensitivity for outlying values of y.

For the estimator $T_n = \bar{y}_n$, $SC = y - \bar{y}_n$ which becomes large as $y \to \pm \infty$.

This curve is inconvenient to use because it depends not only on y and T but also on the observed data $y_1, y_2, \ldots, y_n$. One way of avoiding this difficulty is the use of the influence curve. If the estimator T_n can be expressed as a functional of the empirical distribution G_n, i.e. $T_n = T(G_n)$, then Hampel (1974) introduced the <u>influence curve</u>

$$IC(y;G,T) = \lim_{\varepsilon \to 0} [T\{(1-\varepsilon)G + \varepsilon\delta y\} - T\{G\}]/\varepsilon \qquad (4.6.2)$$

where δy represent the distribution which assigns probability one to the point y. For the estimator $T = \bar{y}_n$, $T(G) = \mu$ and $IC = y - \mu$. For any M-estimator it follows that (see Huber (1981))

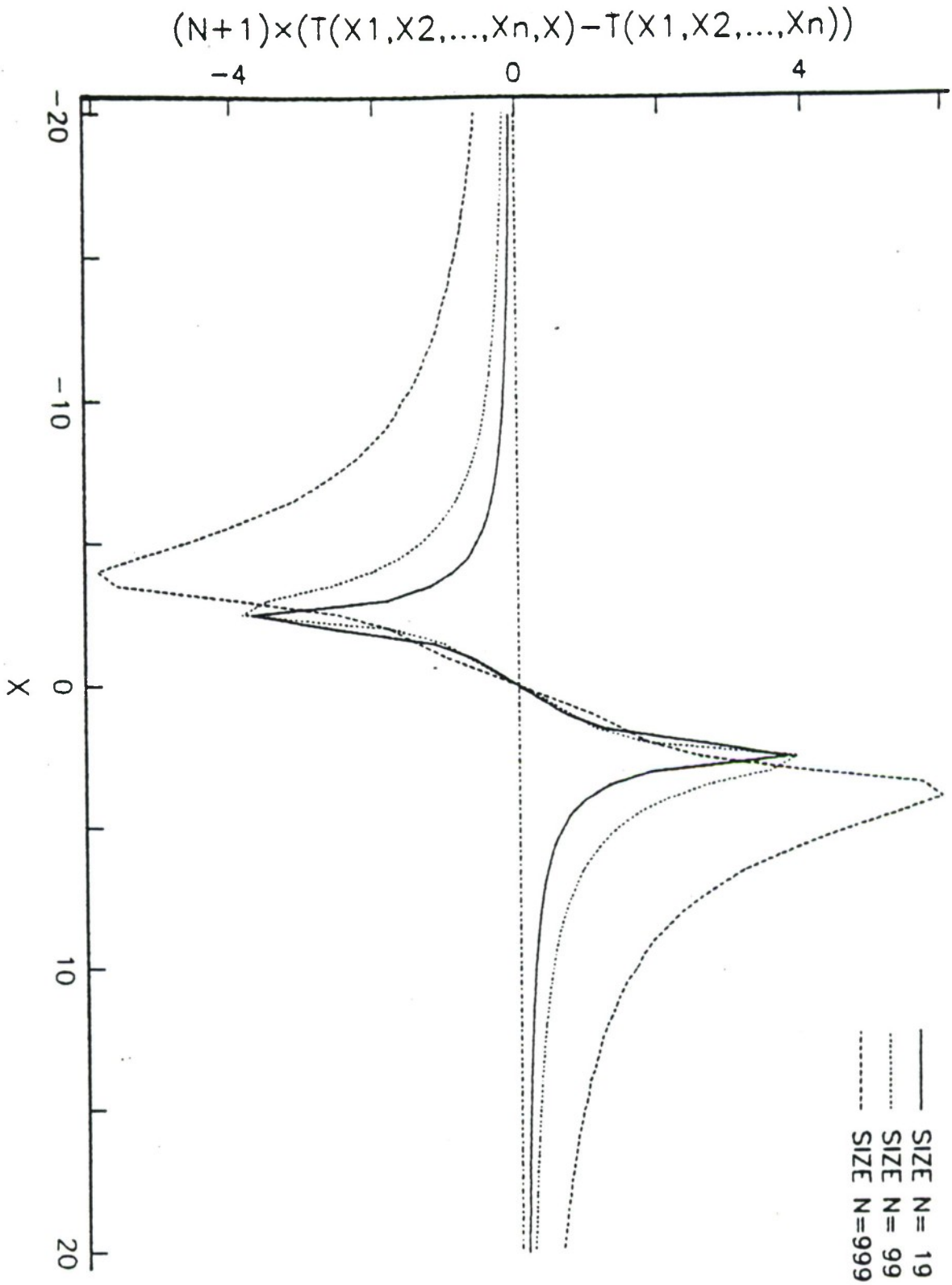
$$IC(y;\theta) = c\psi(y;\theta)$$

where c is constant and $\psi(y;\theta) = \partial\rho(y;\theta)/\partial y$. Our NPS estimator may be regarded as an adaptive M-estimator where the form of ψ is data dependent and the above result is not sufficient to make the use of the influence curve convenient for our estimator.

We choose instead to rely on the use of the synthetic example $x_{1n},\ldots,x_{nn}$ where x_{in} is the $i/(n+1)$ fractile of G. The resulting curves will be called a stylized sensitivity curve (SSC).

In Table 4.6.3, we present a qualitative description of the stylized sensitivity curve based on Gaussian G. In Figures 4.6.4-4.6.6 these curves are graphed for the Gaussian, Logistic and Slash distributions respectively. In each of these cases the SCC is bounded. This is an anticipated consequence of the following heuristic argument. First, as $y \to \pm \infty$, the estimate $\hat{\gamma}$ of the tail thick parameter or $\hat{\tau}$ the scale parameter must get large. But $\hat{\tau}$ is pretty much constrained by the implicit requirement that most of the observations should lie between $\tilde{\mu} \pm \hat{\tau}$. Explicitly the term $-\log \tau$ which occurs n+1 times in the likelihood based on the sample $y, x_{1n},\ldots,x_{nn}$, keeps τ from growing too fast. If now we treat the NPS estimator as an M-estimator with $\rho(y-\mu)$ replaced by $-\log \hat{\tau}^{-1} f(y-\mu;0,\hat{\tau},\hat{\gamma})$ as though $\hat{\tau}$ and $\hat{\gamma}$ were fixed, the sensitivity would be reflected by the corresponding

$$\psi(y) = \frac{(1+\hat{\gamma})\,\mathrm{sgn}\,(y)}{\hat{c}\hat{\tau}\{1 + \frac{\hat{\gamma}}{\hat{c}}(\frac{y}{\hat{\tau}} - 1)\}} \quad \text{for} \quad |y| > \hat{\tau}$$



SENSITIVITY CURVE (NORMAL CASE)

Fig. 4.6.4

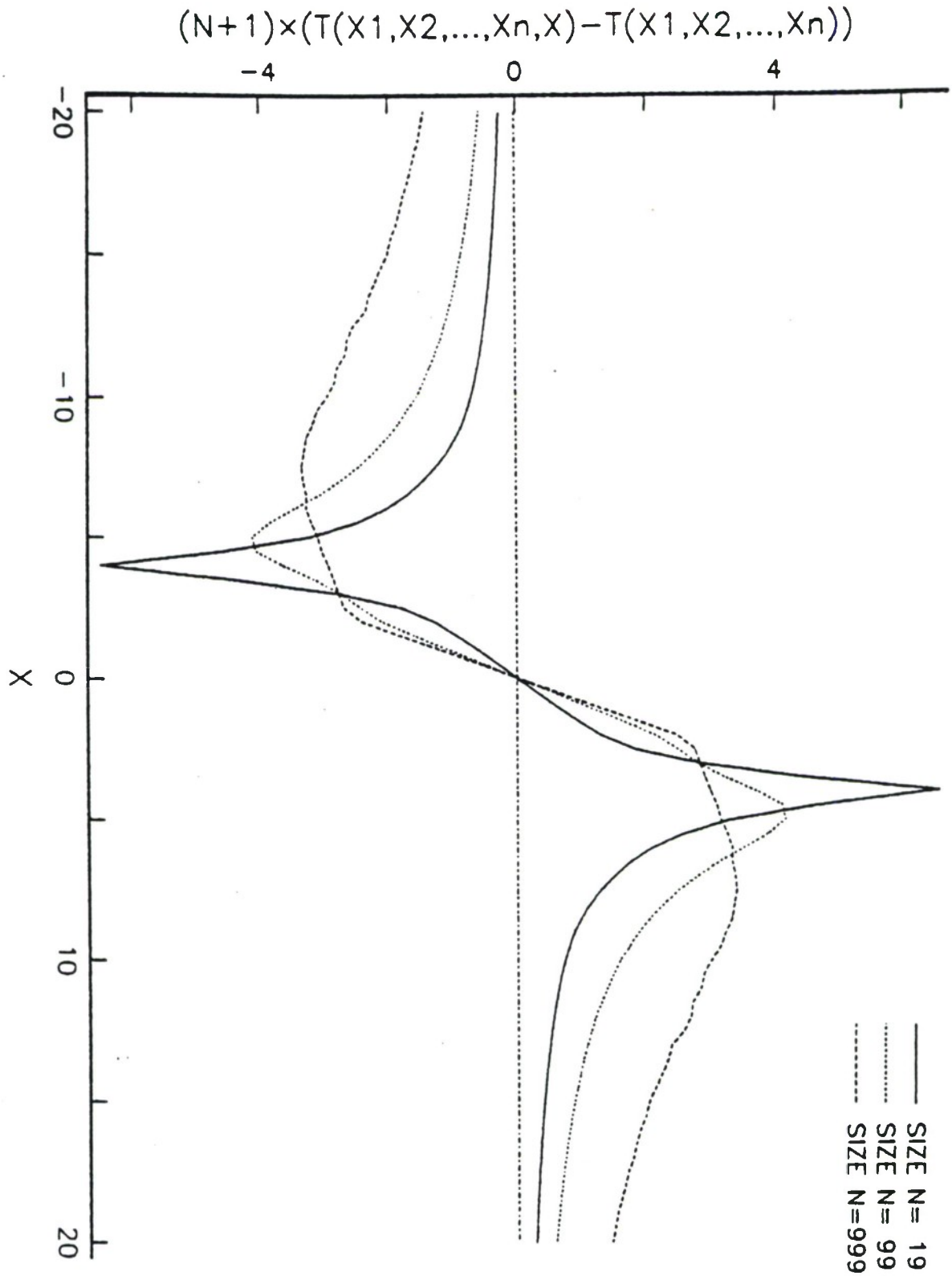
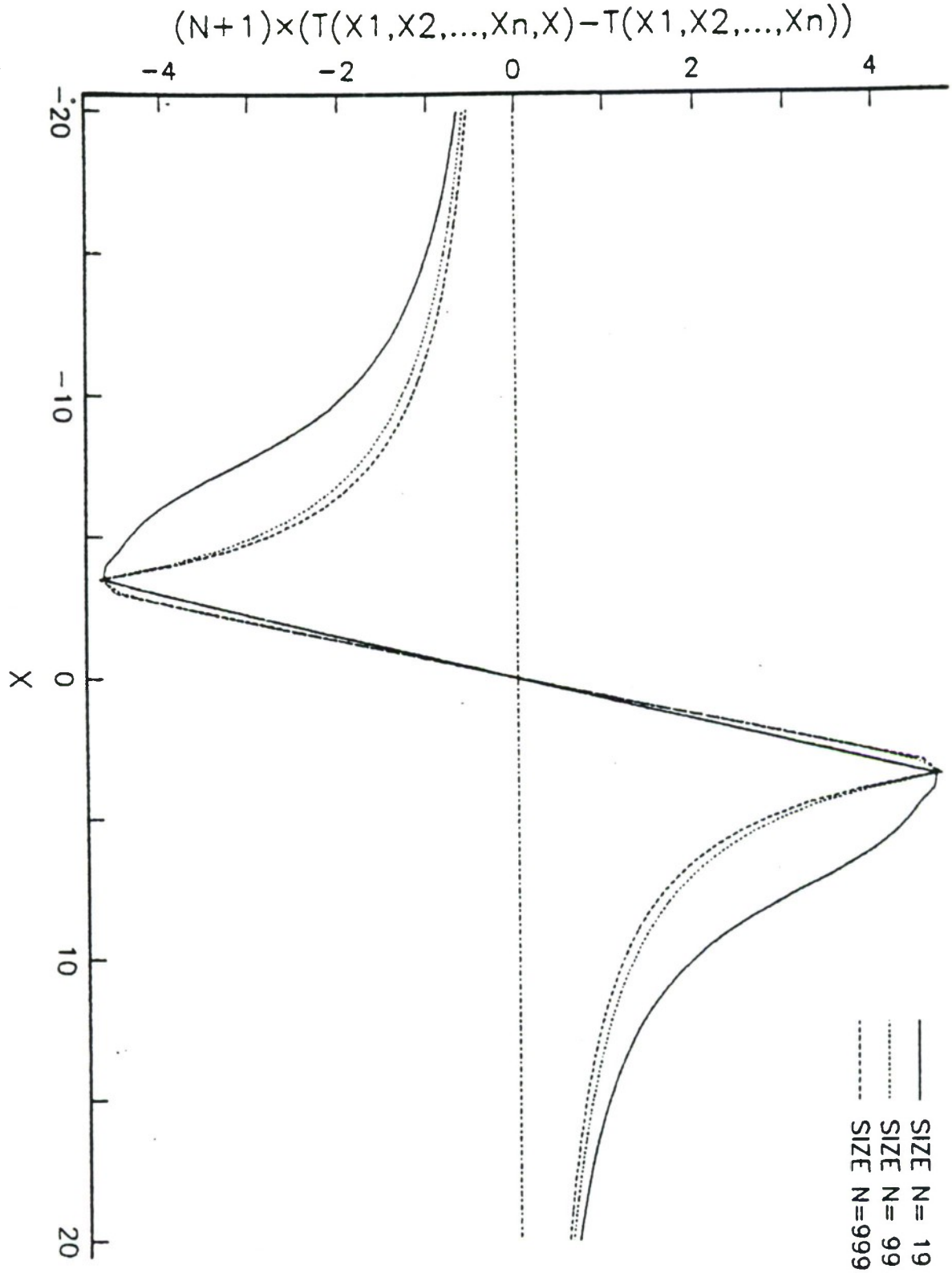

(N+1)×(T(X1,X2,...,Xn,X)−T(X1,X2,...,Xn))
−4
0
4
−20
−10
0
10
20
X
SENSITIVITY CURVE (LOGISTIC CASE)
SIZE N= 19
SIZE N = 99
SIZE N=999

Fig. 4.6.5



SENSITIVITY CURVE (SLASH CASE)

Fig. 4.6.6

as $y \to \infty$, and $\hat{\gamma} \to \infty$, this quantity behaves like $(y-\hat{\tau})^{-1}$ which approaches 0.

For the sensitivity curves in Fig. 4.6.4 - 4.6.5 (Normal or Logistic case) $\hat{\gamma}$ is negative where $y = 0$. Since $\hat{\gamma}$ for $n = 19$ reaches 0 much faster than $\hat{\gamma}$ for $n = 999$, so in the $n = 19$ case we have an earlier peak point than in the $n = 999$ case.

For the sensitivity curve in the slash case, $\hat{\gamma}$ is positive where $y = 0$, so the peak point occurs at $\hat{\tau}$. When y is large, $\hat{\gamma}$ for $n = 999$ is larger than $\hat{\gamma}$ for $n = 19$, so the sensitivity curve for $n = 999$ drops down faster than the sensitivity curve for $n = 19$.

4.7. Simulation results

Simulations were carried out to determine the sampling properties of the NPS estimator for finite samples from the Normal, Logistic and Slash distributions. We present the variance of μ (relative efficiency compared with M.L.E.) and standard deviation of variance of μ based on sampel of sizes 20, 100, 1000 in Table 4.7.1. Also for the comparison purpose, we present the results of further simulations using two adaptive trimmed means for comparisons with the NPS estimator. The two adaptively trimmed means will be denoted JBT and WHD. The JBT method

Table 4.6.3 Behavior of the S.C. for the NPS estimate where underlying distribution is Gaussian

Range	Behavior
$0 \leq y \leq \tau$	Almost linear. $\hat{a}$ does not vary much (since $\psi(y) = -2\hat{a}\cdot y$ if $\lvert y\rvert < \tau$)
$\tau < y \leq \tau^*$	Curved upward, since $\hat{\gamma} < 0$ $\psi(y)$ goes up.
$\tau^* < y \leq \tau^{**}$	Curved downwards, because $\hat{\gamma}$ increases and approaches to 0. We have range expansion for $\psi(y)$
$\tau^{**} < y$	Asymptotically goes to 0, because $\hat{\gamma} = 0$ at τ^{**} and goes up. We have infinite range for $\psi(y)$

uses either the 8% trimmed mean or the 25% trimmed mean, the choice based on whichever has the smaller estimated standard error based on variance calculations on the particular sample. The JBT estimate is proposed by John Tukey and, as described in Andrews et Al. (1972) is simple, robust and performs relatively well. The WHD method is simpler and chooses between the ordinary mean and the 25% trimmed mean based on same criteria, proposed by William DuMouchel.

Table 4.7.2 presents the results from 500 replications with various variance reduction methods which will be described in sections 6.2 and 6.3.

Although these location estimates of μ may not be affected by outliers, we also present probability plot in Fig. 4.7.3 - Fig. 4.7.5, so that we can examine the outliers.

Since $I_{\mu,\mu}$ is given by $\int (\frac{\partial \log f}{\partial x})^2 f \, dx$, for the logistic distribution,

$$I_{\mu,\mu} = \int \{\frac{\partial}{\partial x} \log \frac{e^{-x}}{(1+e^{-x})^2}\}^2 \frac{e^{-x}}{(1+e^{-x})^2} \, dx = \frac{1}{3},$$

so the Cramer Rao bound for estimating the location parameter of a logistic distribution is 3. For the slash, by numerical integration, we get 4.847 as the asymptotic variance of an efficient location estimator.

Table 4.7.1 Product of sample size and variance of the location estimate $\hat{\mu}$, relative efficiency*, and standard deviation of variation of $\hat{\mu}$.

Distribution	Normal	Logistic	Slash
sample size (n)			
n × v 1000	1.013±.001**	3.006±.074***	5.240±.139**
efficiency*	.987	.998	.925
n × V 100	1.067±.005	3.065±.080	5.698±.196
efficiency*	.937	.979	.851
n × V 20	1.086±.006	3.330±.124	6.711±.380
efficiency*	.921	.901	.722

* compared with MLE

** from formula 6.2.1

*** from formula 6.3.4

Table 4.7.2 Product of sample size and variance of the location estimate using JBT(WHD), relative efficiency and standard deviation of variance of location estimate

Estimators compared					
JBT vs. NPS	n × V	1000	1.049±.004	3.009±.075	5.474±.161
	efficiency*		.967	.999	.957
WHD vs. NPS			(1.000±.001)	(3.316±.084)	(5.474±.161)
			(1.013)	(.907)	(.957)
JBT vs. NPS	n × V	100	1.087±.008	3.036±.078	5.698±.179
	efficiency*		.981	1.010	1.000
WHD vs. NPS			(1.051±.008)	(3.037±.077)	(5.691±.177
			(1.015)	(1.010)	(1.001)
JBT vs. NPS	n × V	20	1.111±.009	3.172±.102	6.610±.318
	efficiency*		.977	1.050	1.015
WHD vs. NPS			(1.089±.010)	(3.073±.108)	(6.258±.302)
			(.997)	(1.084)	(1.072)

*compared with NPS estimator from simulation results

We conclude that for most cases NPS estimation is better than or about as efficient as the JBT and WHD estimators when the sample size n is greater than 100. When $n = 20$, the WHD method performed slightly better in these simulations, with NPS and JBT about equal.

If we compare our simulation results for $n = 1000$ with the asymptotic variance of Huber's M-estimate (see Table 4.3.4), we can say that regardless of how Huber's trimmed constant k is chosen, in most cases NPS is better than Huber's M-estimator. If the tail behavior of the distribution generating data is far from exponential, then the NPS estimator is always more efficient than Huber's M-estimator.

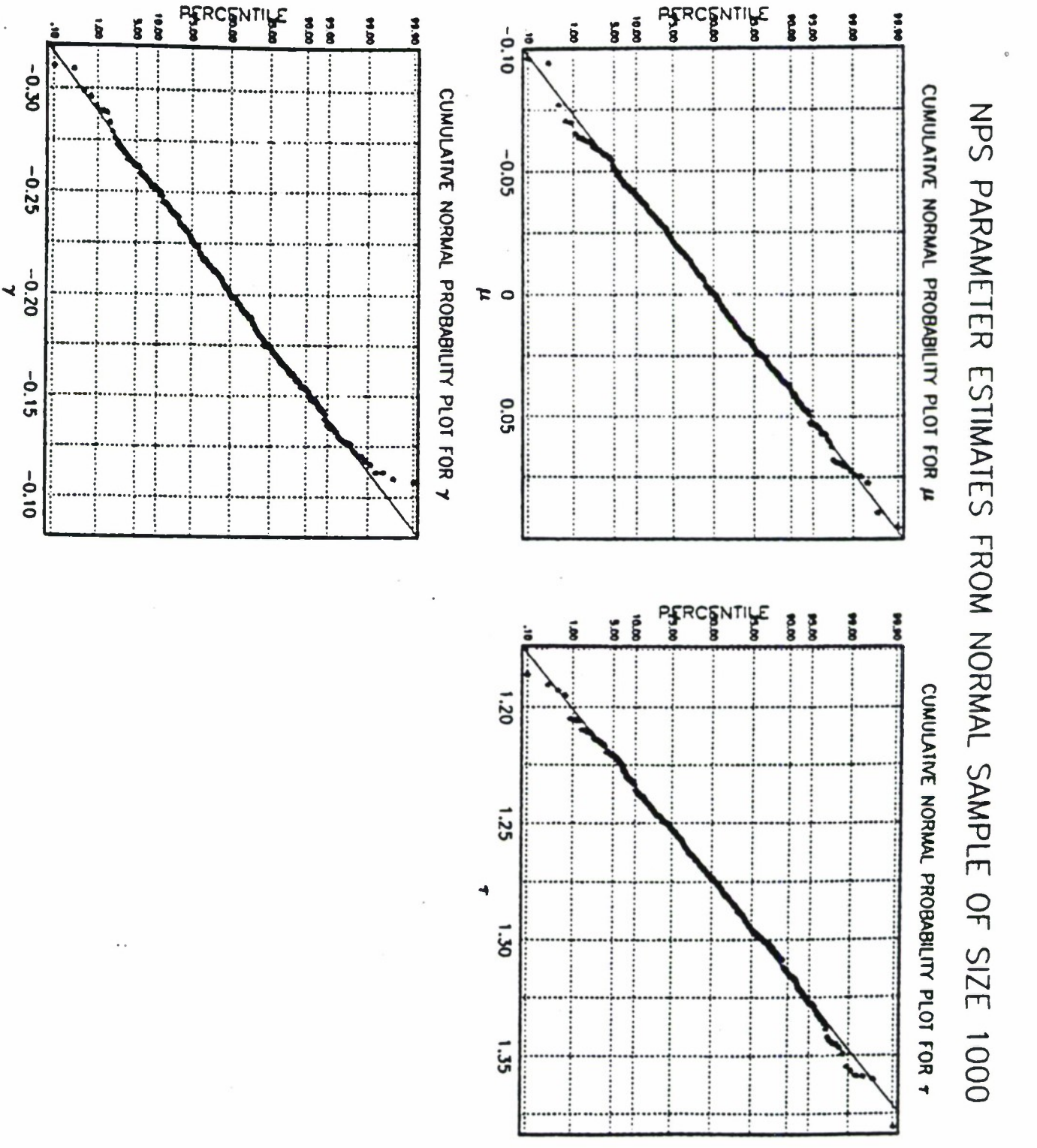

NPS PARAMETER ESTIMATES FROM NORMAL SAMPLE OF SIZE 1000
CUMULATIVE NORMAL PROBABILITY PLOT FOR μ
PERCENTILE
μ
-0.10
-0.05
0
0.05
CUMULATIVE NORMAL PROBABILITY PLOT FOR σ
PERCENTILE
σ
1.20
1.25
1.30
1.35
CUMULATIVE NORMAL PROBABILITY PLOT FOR γ
PERCENTILE
γ
-0.30
-0.25
-0.20
-0.15
-0.10

Fig. 4.7.3

NPS PARAMETER ESTIMATES FROM LOGISTIC SAMPLE OF SIZE 1000

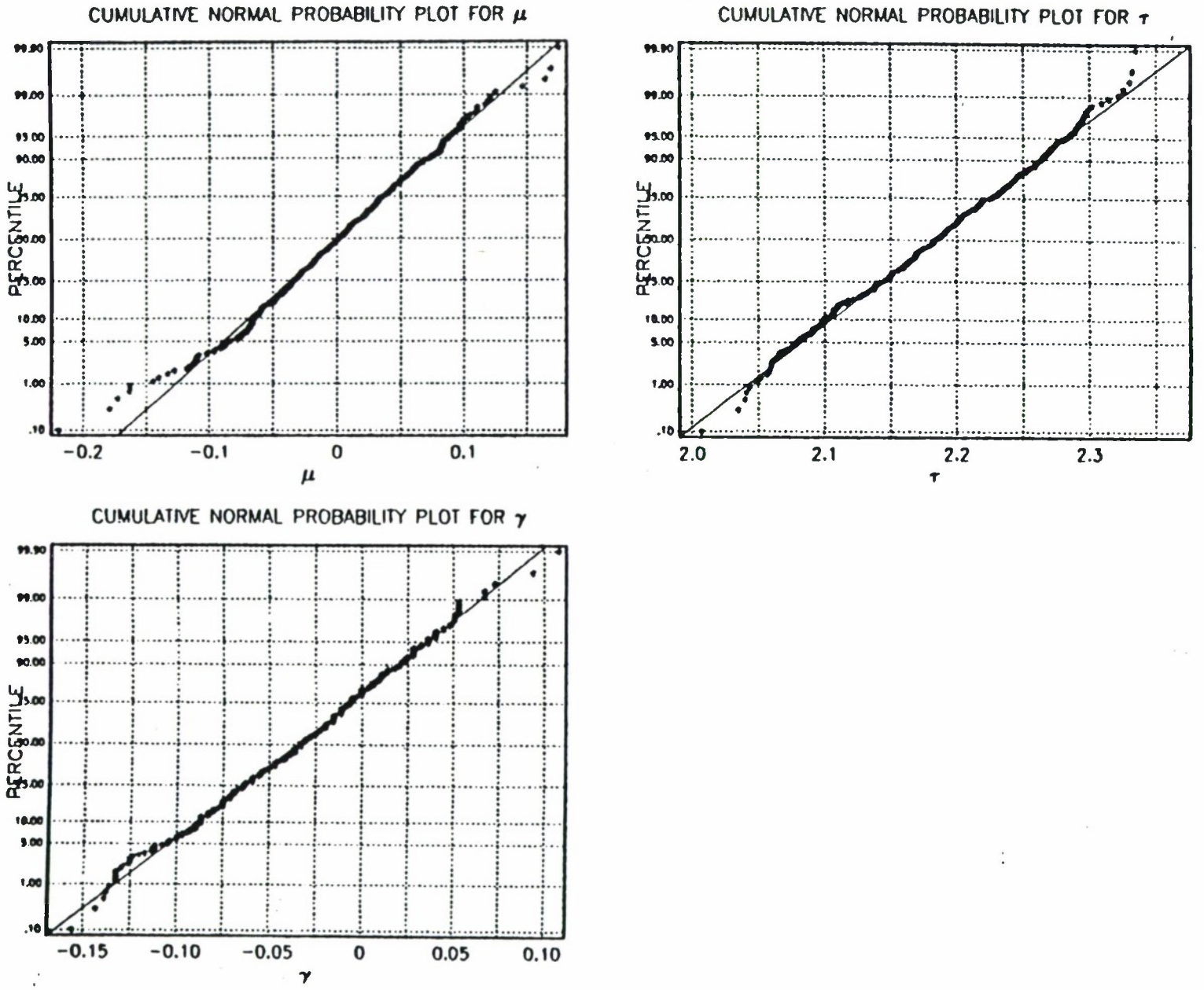


Fig. 4.7.4

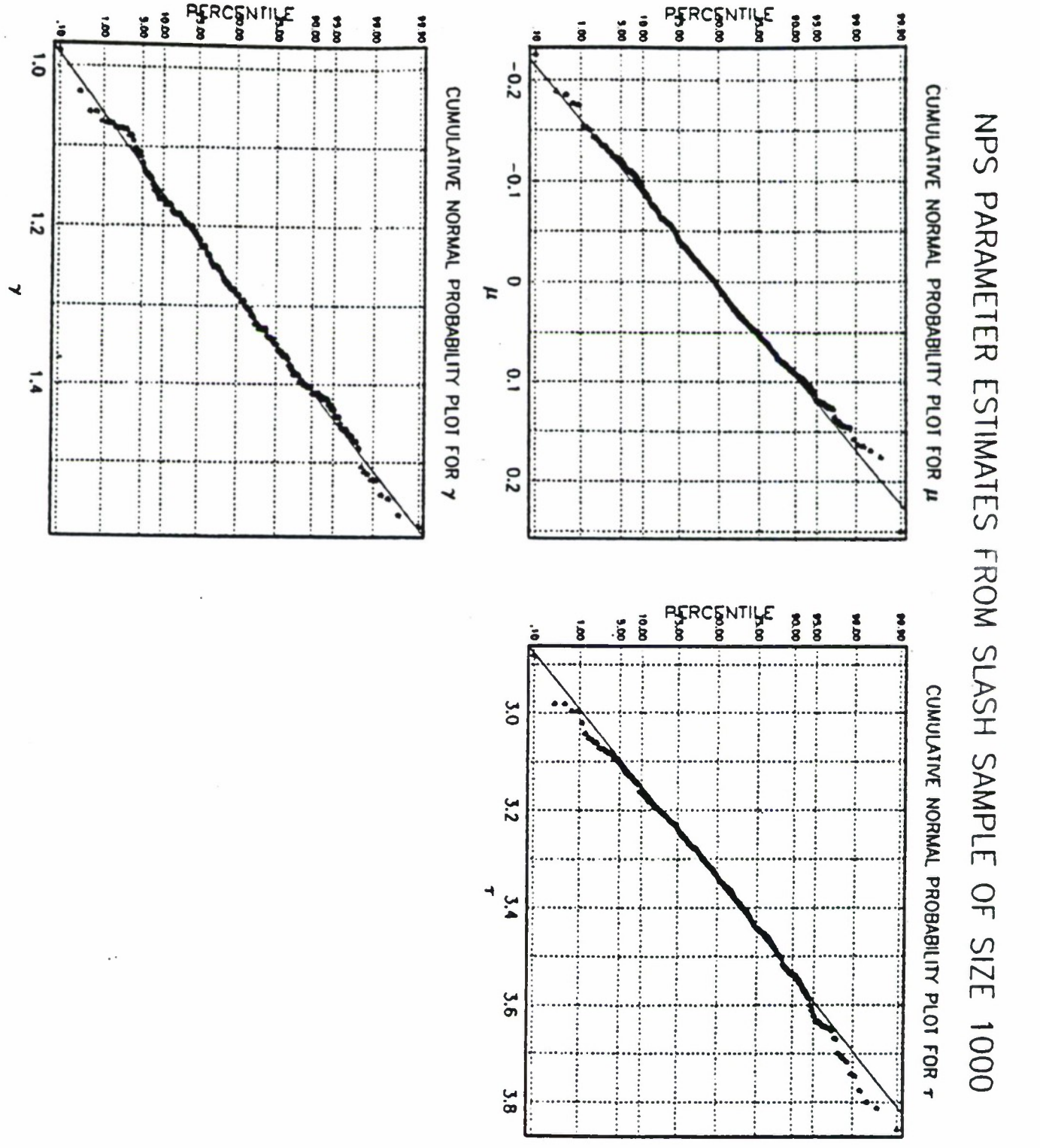

NPS PARAMETER ESTIMATES FROM SLASH SAMPLE OF SIZE 1000
CUMULATIVE NORMAL PROBABILITY PLOT FOR μ
PERCENTILE
μ
CUMULATIVE NORMAL PROBABILITY PLOT FOR τ
PERCENTILE
τ
CUMULATIVE NORMAL PROBABILITY PLOT FOR γ
PERCENTILE
γ

Fig. 4.7.5

Chapter 5

The Two Sample Location Problem and the Asymmetric Model.

5.1. Problem description

One of the fundamental problems of statistics, often encountered in applications, is the two sample location problem. Let $G(x)$ be a symmetric distribution and let $y_1, y_2, \ldots, y_m$ and $z_1, z_2, \ldots, z_n$ be independent random samples from $G(y-\mu_1)$ and $G(y-\mu_2)$ respectively. It is desired to estimate $\delta = \mu_1 - \mu_2$. One way to proceed is to estimate δ by the difference of two separate NPS estimations of μ_1 and μ_2, thereby ignoring the fact that Y and Z have common distribution except for location. A natural alternative of course, is to extend the concept of NPS estimation to this problem by applying the method of maximum likelihood to the model where the Y's and Z's have common scale and shape parameters. That is, we act as though the Y_i come from $NPS(\mu_1, \tau, \gamma)$ and the Z_j from $NPS(\mu_2, \tau, \gamma)$.

Another alternative is to pool the estimate of γ but to permit the use of separate τ's. This would be most appropriate for problems where one anticipates the possibility of different scale parameters but common tail behavior, i.e. distribution of the form $G(\frac{y-\mu_1}{\sigma_1})$ and

$G(\frac{y-\mu_2}{\sigma_2})$. Then we apply MLE to $NPS(\mu_1,\tau_1,\gamma)$ for Y_i and $NPS(\mu_2,\tau_2,\gamma)$ for Z_j.

5.2. The computation for the two sample model with possibly different variances.

We present an iterative method of estimating the parameters for the two sample problem with common γ and possibly different scale parameters.

STEP 1 Calculate the NPS estimator for each sample separately, yielding $\hat{\theta}_y = (\hat{\mu}_y,\hat{\tau}_y,\hat{\gamma}_y)'$ and $\hat{\theta}_z = (\hat{\mu}_z,\hat{\tau}_z,\hat{\gamma}_z)'$ respectively.

STEP 2 Let

$$v_i = \frac{y_i - \hat{\mu}_y}{\hat{\tau}_y} \qquad i = 1,2,\ldots,m$$

and

$$v_{m+j} = \frac{z_j - \hat{\mu}_z}{\hat{\tau}_z} \qquad j = 1,2,\ldots,n$$

Based on the sample $\{v_i : 1 \leq i \leq m+n\}$ apply MLE to the model $NPS(0,1,\gamma_v)$ to estimate $\hat{\gamma}_v$.

STEP 3 Based on the sample $\{y_i : 1 \leq i \leq m\}$, apply MLE to the model $MPS(\mu,\tau,\hat{\gamma}_v)$ with $\hat{\gamma}_v$ determined in

STEP 2 to revise $\hat{\mu}_y, \hat{\tau}_y$. Revise $\hat{\mu}_z, \hat{\tau}_z$ similarly. If $\hat{\mu}_y, \hat{\tau}_y, \hat{\mu}_z$ and $\hat{\tau}_z$ are all sufficiently close to the previous values, stop the iteration. Otherwise return to STEP 2.

Proposition 5.2.1 The iterative method leads to a sequence of estimates which converge to a local maximum of the MPS likelihood function.

To prove this proposition we introduce some notation. Let $L(\underline{y};\theta)$ be the NPS likelihood based on the sample $\underline{y} = (y_1, y_2, \ldots, y_m)$. Let $\underline{1}$ represent a vector all of whose elements are unity. Denote the i-th estimate of the parameters $\mu_y, \ldots, \gamma_v$ by $\mu_{iy}, \ldots, \gamma_{iv}$, with corresponding values for $\underline{v}_i = (\tau_{iy}^{-1}(\underline{y}-\mu_{iy}\underline{1}), \tau_{iz}^{-1}(\underline{z}-\mu_{iz}\underline{1}))$.
First, we observe that

$$L(\underline{y};\mu,\tau,\gamma) = \tau^{-m} L(\tau^{-1}\underline{y};\tau^{-1}\mu,1,\gamma)$$

$$L(\underline{y};\mu,\tau,\gamma) = L(\underline{y}-\mu\underline{1};0,\tau,\gamma)$$

and consequently

$$L(\underline{y};\mu,\tau,\gamma) = \tau^{-m} L(\tau^{-1}(\underline{y}-\mu\underline{1});0,1,\gamma).$$

After completing STEP 2, our combined likelihood is

given by

$$L_1 = \tau_{1y}^{-m}\tau_{1z}^{-n}L(\underline{v}_1;0,1,\gamma_{1v}) = L(\underline{y};\mu_{1y},\tau_{1y},\gamma_{1v})\cdot L(\underline{z};\mu_{1z},\tau_{1z},\gamma_{1v})$$

$$\leq L(\underline{y};\mu_{2y},\tau_{2y},\gamma_{1v})\cdot L(\underline{z};\mu_{2z},\tau_{2z},\gamma_{1v}) = \tau_{2y}^{-m}\cdot\tau_{2z}^{-n}L(\underline{v}_2;0,1,\gamma_{1v})$$

$$\leq \tau_{2y}^{-m}\tau_{2z}^{-n}L(\underline{v}_2;0,1,\gamma_{2v}) = L(\underline{y};\mu_{2y},\tau_{2y},\gamma_{2v})\cdot L(\underline{z};\mu_{2z},\tau_{2z},\gamma_{2v})$$

$$= L_2$$

Thus the likelihood function is nondecreasing after each pair of STEP 3 and STEP 2. If the parameters change at a step only if the likelihood increases, this procedure leads to a monotonic increasing likel ihood unless the process stops. The process can not yield a limit point which is not a local maximum for the gradient is not zero at such a point, and one of the two steps will lead to a substantial increase in the likelihood, once we are in the neighborhood of such a point. Thus the process must lead to a local maximum or an unbounded sequence. But it is easy to see that the likelihood approaches 0 for a sequence which is unbounded in the parameter space.

5.3. The asymmetric model.

We have confined attention to the location parameter for symmetric distributions. The estimation of location for an unspecified asymmetric distribution is not a well defined problem. On the other hand, from the point of view of estimating densities, we may pose the problem of using an asymmetric version of the NPS model to approximate unimodal distributions and to estimate these distributions by estimating the parameters of that model. We shall say that a random variable X has the <u>standard asymmetric NPS distribution</u> with skewness parameter s, left tail parameter γ_L, and right tail parameter γ_R, if it has a density of the form

$$f_0(x,s,\gamma_L,\gamma_R) = \begin{cases} \frac{1}{10d_L}\{1 + \frac{\gamma_L}{d_L}(-\frac{x}{s} - 1)\}^{-\frac{1}{\gamma_L}-1} & A_1<x<-s,\ \gamma_L \neq 0 \\ e^{a_L(\frac{x}{s})^2+b_L(\frac{x}{s})+c} & -s\leq x<0 \\ e^{a_R x^2+b_R x+c} & 0\leq x<1 \\ \frac{1}{10d_R}\{1 + \frac{\gamma_R}{d_R}(x-1)\}^{-\frac{1}{\gamma_R}-1} & 1<X<A_2,\ \gamma_R\neq 0 \end{cases} \qquad (5.3.1)$$

If $\gamma_L > 0$ then $A_1 = \infty$, and if $\gamma_L < 0$ then $A_1 = s(\frac{d_L}{\gamma_L} - 1)$.

If $\gamma_R > 0$ then $A_2 = \infty$, and if $\gamma_R < 0$ then $A_2 = 1 - \frac{d_R}{\gamma_R}$.

If $\gamma_L = 0$ and $x < -s$ then $A_1 = \infty$ and

$$f_0(x,s,0,\gamma_R) = \frac{1}{10d_L} e^{-\frac{1}{d_L}(-\frac{x}{s} - 1)} \tag{5.3.1'}$$

If $\gamma_R = 0$ and $x > 1$ then $A_2 = \infty$ and

$$f_0(x,s,\gamma_L,0) = \frac{1}{10d_R} e^{-\frac{1}{d_R}(x-1)} \tag{5.3.1"}$$

The parameters d_L, d_R, a_L, a_R, b_L, b_R and c depend on s, γ_L and γ_R, and are determined by the requirement that $\Pr\{-s \leq x < 0\} = \Pr\{0 \leq x < 1\} = 0.4$, and the spline constraints so that the density and the first derivative of the density are continuous everywhere.

The parameters d_L, d_R, a_L, a_R, b_L, b_R and c satisfy the following spline equations:

$$f_0(-s^+) = f_0(-s^-) \tag{5.3.2}$$

$$f_0'(-s^+) = f_0'(-s^-) \tag{5.3.3}$$

$$f_0'(0^+) = f_0'(0^-) \tag{5.3.4}$$

$$f_0(1^+) = f_0(1^-) \tag{5.3.5}$$

$$f_0'(1^+) = f_0'(1^-) \tag{5.3.6}$$

$$\int_{-s}^{0} e^{a_L(\frac{x}{s})^2 + b_L(\frac{x}{s}) + c} dx = 0.4 \tag{5.3.7}$$

$$\int_{0}^{1} e^{a_R x^2 + b_R x + c} dx = 0.4 \tag{5.3.8}$$

In addition, we also consider the family of the variables $Y = \mu + \tau X$ where X has the standard asymmetric NPS distribution, and Y has the asymmetric NPS distribution $ANPS(\mu,\tau,s,\gamma_L,\gamma_R)$, with median at μ and interdecile range equal to $\tau(1+s)$.

(See Fig. 5.3.9)

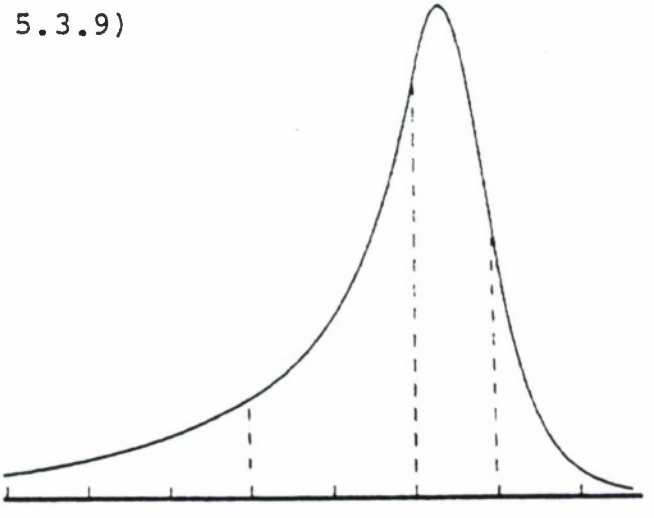


Fig. 5.3.9. Schematic asymmetric NPS model density shown has $s = 2$, $\gamma_L = 0$, $\gamma_R = 0$

where $\mu-s\tau$: 10^{th} percentile

μ : 50^{th} percentile (median)

$\mu+\tau$: 90^{th} percentile

This is a 5-parameter family. The five primary parameters are μ, the location parameter; τ, the scale parameter; s, the skewness parameter; γ_L, the left tail shape parameter; and γ_R the right tail shape parameter. The other parameters d_L, d_R, a_L, a_R, b_L, b_R and c are defined implicitly by s, γ_L and γ_R,

Fig. 5.3.10 presents some variations of the asymmetric NPS model. In Fig. 5.3.10(a), the 3 densities are all symmetric. One has exponential tails ($\gamma_L = \gamma_R = 0$), and for comparison the other two have thicker tails ($\gamma_L = \gamma_R = .0.3$) or thinner tails ($\gamma_L = \gamma_R = -0.3$). In Fig. 5.3.10(b), all 3 asymmetric NPS densities have exponential tails. One is symmetric (s=1) and for comparison the other two are skewed to the right (s=2) or skewed to the left (s=0.5). In Fig. 5.3.10(c), all 3 asymmetric NPS densities are skewed to the right, and have exponential tails. One has standard scale parameter (τ=1) and for comparison the other two have larger scale (τ=2) or smaller scale (τ=0.5). In Fig. 5.3.10(d), all 3 asymmetric NPS densities have the same scale and a skewness parameter of s = 2. One has exponential tails and for comparison the other two have a thicker right tail (γ_R=0.5) or a thinner left tail (L=-0.5).

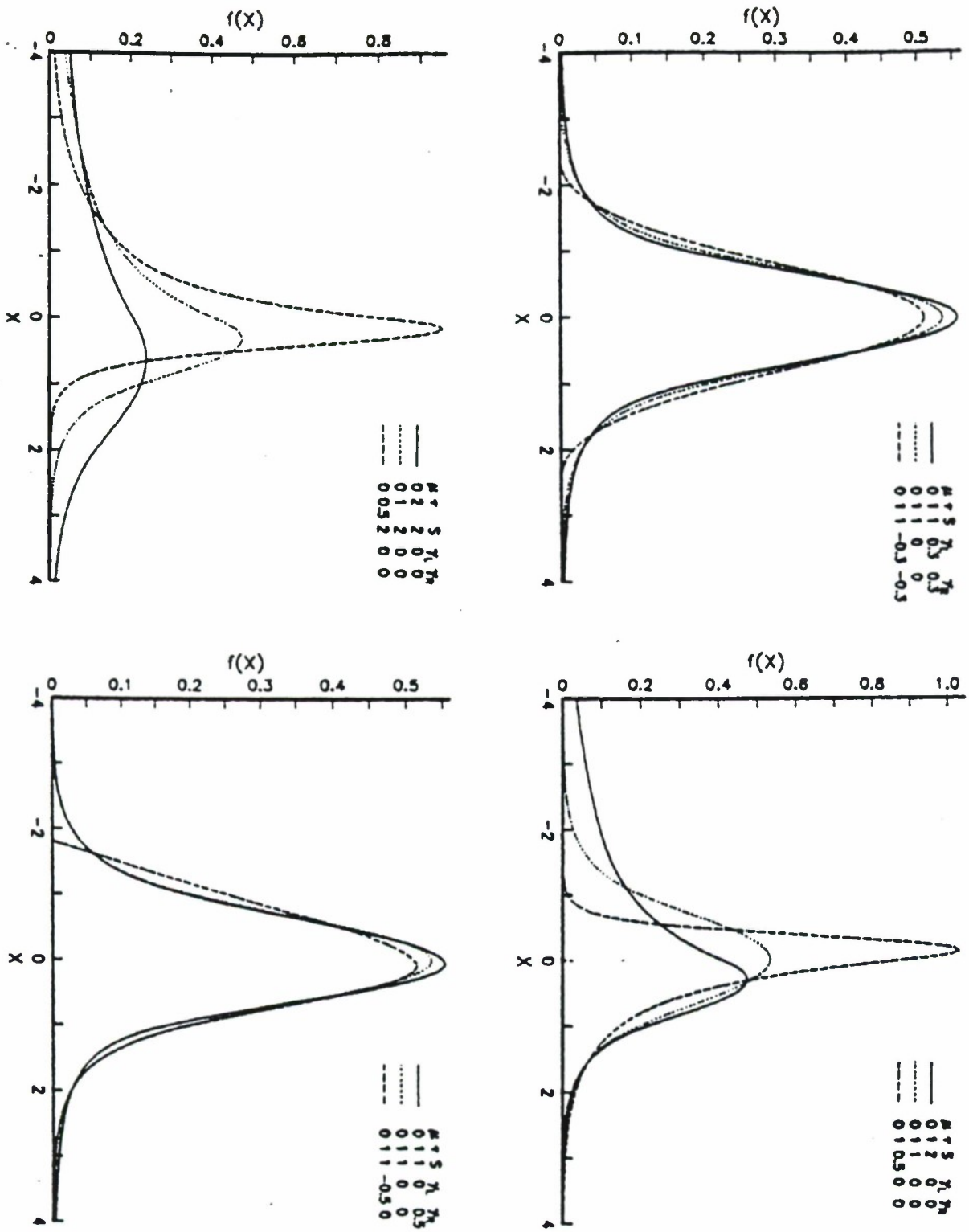


Fig. 5.3.10

Chapter 6

Computational Problems

The calculation of the NPS estimates and the simulations that were carried out required extensive computations. In this chapter we discuss three methods which were applied to reduce the computer time used or the programming difficulty.

6.1. The computation of the NPS estimates

The NPS estimates require a maximization subject to the spline constraints (3.1.2) to (3.1.4). First numerical calculations were performed to construct an accurate table representing a, b and c as functions of γ. For the later calculations, interpolation on that table was performed. A portion of that table is presented in Table 3.1.6.

The next part was that of maximizing the likelihood. For this a simplex method developed by Nelder and Mead (1965) was applied. This method is a direct search procedure and is not related to the simplex method of linear programming. It has the following advantageous properties.

(i) No assumptions are made about the surface except that it is continuous and has a unique maximum in the area of search.

(ii) This method does not involve any derivatives.

(iii) This method is effective and computationally compact.

The details of the simplex method are as follows. Suppose the problem is to find the maximum of some function $f(x_1,x_2,\ldots,x_p)$. Since f has p variables, we need to evaluate f at $(p+1)$ trial values of x, denoted here by $A_0,A_1,\ldots,A_p$. Assume that these points lie on the vertices of an irregular simplex in $(p+1)$ space and that $f(A_0) = \min f(A_i)$. In this case, a reflection is made through the point C (centroid of face opposite A_0) to a point B where $B = A_0 + 2(C-A_0)$. One version of the simplex method consists of replacing A_0 by B, relabeling the points so that $f(A_0)$ is the smallest, and repeating the process of replacing the worst point by its reflection B. A more sophisticated version of the algorithm was actually used, in which A_0 is replaced by D, where D is of the form $A_0 + d(C-A_0)$. One of four possible values of d are used, depending on the relationship of $f(B)$ to $f(A_0),\ldots,f(A_p)$. (Fig. 6.1.1-6.1.4 shows $p=2$).

Case 1. If $f(B) > \max (f(A_1),\ldots,f(A_n))$, then an extension is made where $d = 3$. (See Fig. 6.1.1)

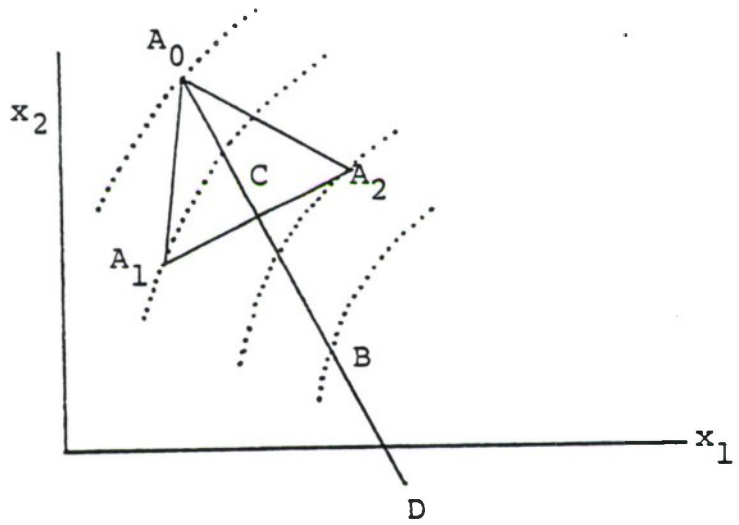


Fig. 6.1.1

(Dashed lines show contours of f)

Case 2. If $f(B) < f(A_0)$, then a contraction is made where $d = 0.5$ (See Fig. 6.1.2)

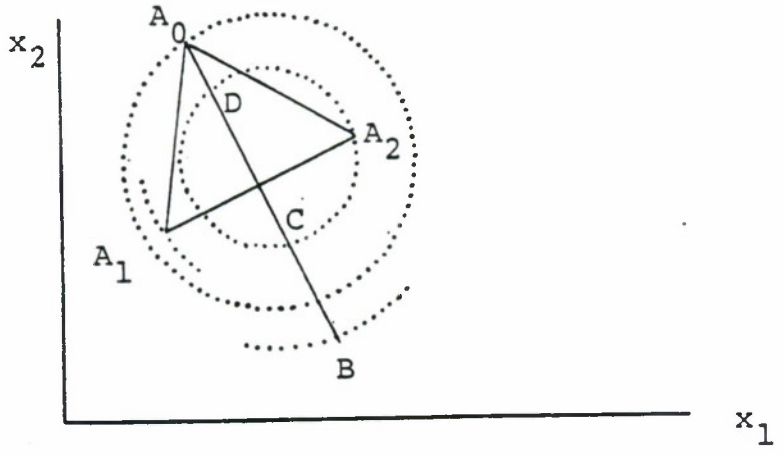


Fig. 6.1.2

Case 3. If $f(A_0) < f(B) < \min (f(A_1),\ldots,f(A_p))$, then a contraction is made where $d = 1.5$ (See Fig. 6.1.3)

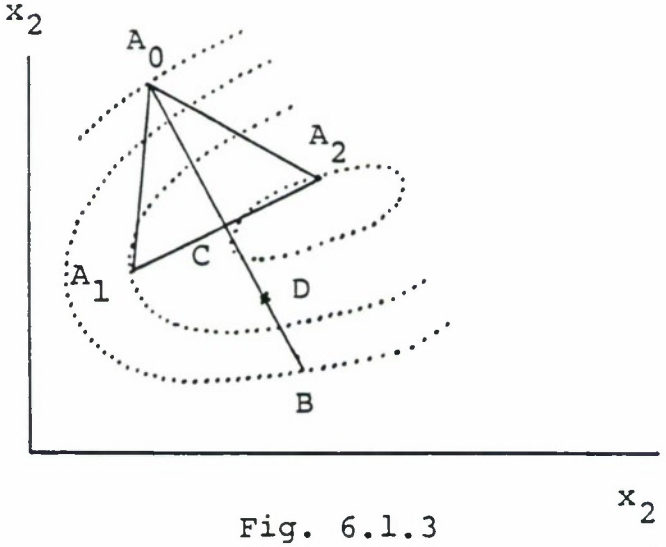


Fig. 6.1.3

Case 4. If

$\max (f(A_1),\ldots,f(A_p)) < f(B) < \min (f(A_1),\ldots,f(A_p))$,

then a reflection is made where $d = 2$ (i.e. $D = B$). (See Fig. 6.1.4)

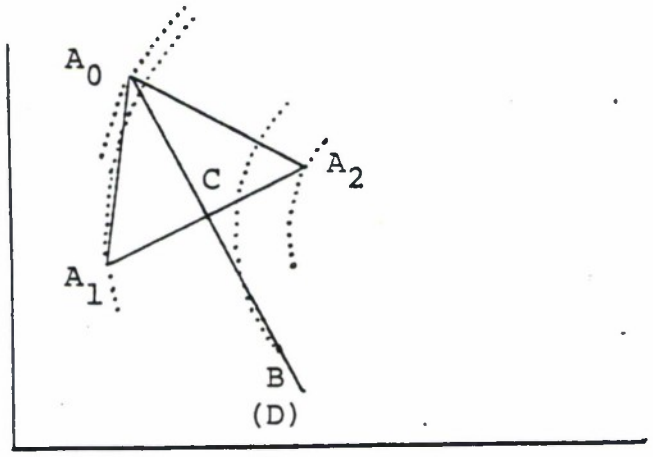


Fig. 6.1.4

6.2. Swindle

Whenever we try to do experimental sampling in a computer simulation, it is wise to try to find a restatement of the problem which reduces the amount of computation required to achieve the desired precision in results. The restatement we consider here is called a Monte Carlo swindle, (See Gross (1973)). Let $\{x_i, i = 1,2,\ldots,n\}$ be a sample of size n from some symmetric distribution G of a random variable x which has the form $x = u/v$ where u is $N(0,1)$ and v is independent and positive.

A number of important distributions (Student's t, Cauchy, double exponential, slash) are of that form. If $v \equiv 1$, G is $N(0,1)$, and if v has the Uniform $(0,1)$ distribution x has the slash distribution.

Our observations thus are $x_i = \frac{u_i}{v_i}$. Given v_i, $x_i \sim N(0, v_i^{-2})$. Let

$$\hat{x} = \Sigma x_i v_i^2 / \Sigma u_i^2,$$

and

$$s^2 = \frac{1}{n-1} \Sigma (x_i - \hat{x})^2 v_i^2.$$

Also, let

$$c_i = (x_i - \hat{x})/s \qquad i = 1,2,\ldots,n$$

represent the elements of the <u>configuration vector</u> $\underline{c}$. It is known that $\underline{c}$, s and $\hat{x}$ are conditionally independent given $\underline{v}$. Moreover the conditional distributions of $\hat{x}$ and s^2 are $N(0, (\Sigma v_i^2)^{-1})$ and that of $\chi^2_{n-1}/(n-1)$.

Now let T_n be a symmetric, scale and location invariant, statistics of the sample $\underline{x}$. Then

$$T_n[(\underline{x} - a\cdot\underline{1})/b] = [T_n(\underline{x}) - a]/b$$

and

$$T_n(\underline{x}) = - T_n(-\underline{x})$$

and consequently $ET_n(\underline{x}) = 0$ if the expectation exists. While the variance of T_n can be estimated directly by simulation, we shall express this variance in terms of known quantities and of the expectation of a conditional expectation which has smaller variance and can therefore be estimated more precisely. Thus

$$\begin{aligned} ET_n^2 &= E\{E(T_n^2|\underline{v},\underline{c})\} \\ &= E[E\{[\hat{x} + sT_n(\underline{c})]^2|\underline{v},\underline{c}\}] \\ &= [\frac{1}{\Sigma v_i^2} + T_n^2(\underline{c})] \end{aligned}$$

since $E[\hat{x}\cdot s|\underline{v},\underline{c}] = 0$, $E[\hat{x}^2|\underline{v},\underline{c}] = 1/\Sigma v_i^2$ and $E(s^2|\underline{v},\underline{c}) = 1$.

Now $T_n(\underline{c}) = s^{-1}[T_n(\underline{x}) - \hat{x}]$ tends to be less variable than $T_n(\underline{x})$ and its variance can be estimated directly from N simulations by

$$\hat{E}(T_n^2(\underline{c})) = N^{-1} \sum_{j=1}^{N} T_n^2(\underline{c}_j).$$

Also the variance of this estimate is estimated by

$$\{\sum_{j=1}^{N} T_n^4(\underline{c}_j) - N^{-1}[\sum_{j=1}^{N} T_n(\underline{c}_j)]^2\}/N(N-1). \qquad (6.2.1)$$

Finally, for the normal G, $E[(\Sigma v_i^2)^{-1}] = n^{-1}$. For the slash distribution, $\Sigma v_i^2 = n/3 + O_p(1)$ and $E\{(\Sigma v_i^2)^{-1}\} = 3n^{-1} + 0(n^{-2})$ and this substantial contribution to $E[T_n^2]$ need not be involved in the simulation.

6.3. Variance reduction for the logistic distribution.

The Monte Carlo Swindle is not applicable to the logistic distribution. Here we use another principle. If our statistic T is highly correlated with another statistic T' whose variance is known, the variance of T can be expressed in terms of that of T' and of a relatively small part of T^2 left over from the linear approximation to the regression of T^2 on T'^2.

Let T and T' have mean 0. We write,

$$T^2 = aT'^2 + b + u = \tilde{T}^2 + u \qquad (6.3.1)$$

where $\tilde{T}^2$ is the linear approximation to the regression and

$$a = \text{Cov}\ (T^2, T'^2)/\text{var}\ (T'^2) \qquad (6.3.2)$$

then

$$E(T^2) = E[T^2 - aT'^2] + aE(T'^2) \qquad (6.3.3)$$

and

$$\text{Var}(T^2) = \text{Var}\ (T^2 - aT'^2) + a^2\ \text{Var}\ (T'^2) \qquad (6.3.4)$$

Thus if a and $E(T'^2)$ are known, we can use the sample to estimate $E(T^2 - aT'^2)$ which has relatively small variance if the correlation of T^2 with T'^2 is high. If a is unknown, it too could be estimated from the simulation. While the precise value of a is necessary for (6.3.4), an approximate value will serve for (6.3.3) which is the essential equation to exploit.

For the logistic distribution we used the mean $\bar{X}$ for T'. Then $E(T'^2) = 3.28987$. As a simulation results, we get $a = .897, .908, .768$ and variance reductions are 89.3%, 82.5%, 71.8% where sample size $n = 1000, 100, 20$ respectively.

Chapter 7

Conclusions

Simulation studies were conducted assuming the data came from Normal, Logistic, and Slash populations with sample sizes 20, 100 and 1000. The NPS estimate seems to be more efficient compared with other adaptive estimates, such as JBT and WHD, specially for medium (100) to moderately large (1000) sample sizes. We have shown that the NPS estimate of location has lower asymptotic variance than Huber's M-estimator in most cases, regardless of Huber's choice of k.

By a sensitivity curve analysis, we show that the NPS estimate of location guarantees resistance to outliers.

For the two-sample location problem, we propose an iterative method to estimate the shift parameter when the scale parameters may be unequal. We proved that this iterative method converges to the desired M-estimate for an arbitrary scale and location family of symmetric distributions.

Finally we proposed an asymmetric family of NPS distributions which can be used to generalize many of our results to help analyze asymmetric data.

References

[1] Andrew, D. et al (1972) Robust estimation of location, survey and advances, Princeton University Press.

[2] Anscombe, F. (1961) Estimating a mixed-exponential response law, J. Amer. Stat. Assoc. 56, 493-502.

[3] Beran, R. (1974) Asymptotically efficient adaptive rank estimates in location models, Ann. Stat. 2, 67-113.

[4] Beran, R. (1978) An efficient and robust adaptive estimator of location, Ann. Stat. 6, 292-313.

[5] Chernoff, H., Gastwirth, J., Johns, M. (1967) Asymptotic distribution of linear combinations of functions of order statistics with applications to estimation, Ann. Math. Stat. 38, 52-72.

[6] DuMouchel, W. (1983) Estimating the stable index α in order to measure tail thickness, Ann. Stat. 11, 1019-1031.

[7] Van Eeden (1970) Efficiency-Robust estimation of location, Ann. Math. Stat. Vol. 41, No. 1, 172-181.

[8] Hall, P. (1982) On estimating the endpoint of a distribution, Ann. Stat. 10, 556-568.

[9] Hampel, F. (1974) The influence curve and its role in robust estimation, J. Amer. Stat. Assoc. 62, 1179-1186.

[10] Harter, J., Moore, A., Curry, T. (1979) Adaptive robust estimation of location and scale parameters of symmetric populations. Comm. Stat. Theor. Meth. A8(15), 1473-1491.

[11] Hogg, R. (1974) Adaptive robust procedures: A partial review and some suggestions for future applications and theory, J. Amer. Stat. Assoc. 69, 909-927.

[12] Huber, P. (1964) Robust estimation of a location parameter, Ann. Math. Stat. 35, 73-101.

[13] Huber, P. (1967) The behavior of maximum likelihood estimates under nonstandard conditions, in: Proc. Fifth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, University of California Press.

[14] Huber, P. (1981) Robust Statistics, John Wiley & Sons.

[15] Nelder, J. and Mead, R. (1965) A simplex method for function minimization, The Computer J. vol. 7, 308-313.

[16] Olsson, D. (1974) A sequential Simplex program for solving minimization problems, J. of Quality Tech. vol. 6, 53-57.

[17] Rubinstein, R. (1981) Simulation and the Monte Carlo method, John Wiley & Sons.

[18] Serfling, R. (1980) Approximation theorems of mathematical statistics, John Wiley & Sons.

[19] Stein, C. (1956) Efficient nonparametric testing and estimation, Proc. Third Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1, University of California Press.

[20] Stone, C. (1975) Adaptive maximum likelihood estimation of a location parameter, Ann. Stat. 3, 267-284.

[21] Takeuchi, K. (1971) A uniformly asymptotically efficient estimation of a location parameter, J. Amer. Stat. Assoc. 66, 292-301.

[22] De Wet, T., Wyk, J. (1979) Efficiency and robustness of Hogg's adaptive trimmed means, Comm. Stat. - Theor. Meth., A8(2) 117-128.

Appendix

```
                                                    *** SIMPKIM **
 ∇ Z←SIMPKIM DATA;MI;SI;GI;I;KK;AA;YY;MED;TEM;NEW;VAL;DUM
⍝-------------------------------------------------------------
⍝ MLE CALCULATION FOR SYMMETRIC NPS DISTRIBUTION
⍝ USING SIMPLEX METHOD
⍝-------------------------------------------------------------
 →ERROR IF(⍴DATA)<10
 AA← 3 4 ⍴0
 YY←4⍴0
 MI←0.5 ORDERSTAT DATA
 SI←0.5×(0.9 ORDERSTAT DATA)-(0.1 ORDERSTAT DATA)
 GI←0
 DUM←(÷⍴DATA)*0.5
⍝-------- SET INITIAL 4 POINTS
 AA[;1]←MI,SI,GI
 AA[;2]←(MI+DUM),SI,GI
 AA[;3]←MI,(SI+DUM),GI
 AA[;4]←MI,SI,(GI+0.1)
⍝-------- CALCULATE LIKELIHOOD FUNCTION VALUE FOR 4 POINTS
LOOP:I←1
L1:YY[I]←DATA LIKEFUN AA[;I]
 I←I+1
 →L1 IF I≤4
⍝-------- FIND THE POINT WHICH HAS MINIMUM LIKELIHOOD VALUE
 AA←AA[;⍋YY]
 YY←YY[⍋YY]
 MED←(AA[;2]+AA[;3]+AA[;4])÷3
 TEM←(2×MED)-AA[;1]
```

```
 VAL←DATA LIKEFUN TEM
 →L2 IF VAL<YY[2]
 →L4 IF VAL>YY[4]
⍝---------- REFLECTION
 AA[;1]←TEM
 →L5
⍝---------- CONTRACTION
L2:NEW←0.5×MED+TEM
 ⍎'NEW←0.5×MED+AA[;1]' IF VAL<YY[1]
 VAL←DATA LIKEFUN NEW
 →L3 IF VAL<YY[2]
 AA[;1]←NEW
 →L5
L3:AA[;1]←0.5×AA[;1]+AA[;4]
 AA[;2]←0.5×AA[;2]+AA[;4]
 AA[;3]←0.5×AA[;3]+AA[;4]
 →L5
⍝---------- EXTENSION
L4:NEW←(2×TEM)-MED
 VAL←DATA LIKEFUN NEW
 AA[;1]←NEW
 ⍎'AA[;1]←TEM' IF VAL>YY[1]
⍝---------- CHECK FOR STOP ITERATIONS
L5:MED←(+/AA)÷4
 KK←(+/(AA[;1]-MED)*2)+(+/(AA[;2]-MED)*2)
 KK←(KK+(+/(AA[;3]-MED)*2)+(+/(AA[;4]-MED)*2))*0.5
 →LOOP IF KK>0.001
 Z←AA[;4]
 →0
ERROR:⎕←'TOO SMALL DATA NUMBERS'
```

```
 →0
 ∇
                                                        *** LIKEFUN **
 ∇ Z←DATA LIKEFUN P;L;R;M1;M2;M3;M4;PA;DUM1;DUM2;DUM3;DUM4;MID;MAD
⍝-------------------------------------------------------------------
 ⍝
⍝ LOG LIKELIHOOD FUNCTION FOR SYMMETRIC NPS DISTRIBUTION
⍝ DATA LIKEFUN (M,S,G)
⍝-------------------------------------------------------------------
 ⍝
 PA←NEWESTAB P[3]
 MID←⌊/DATA
 MAD←⌈/DATA
 →L0 IF P[3]≥0
 →L5 IF(MID<P[1]-P[2]×1-PA[1]÷P[3])
 →L5 IF(MAD>P[1]+P[2]×1-PA[1]÷P[3])
L0:L←P[1]-P[2]
 R←P[1]+P[2]
 M1←DATA IF(DATA≤L)
 M2←DATA IF((DATA>L)∧(DATA≤R))
 M3←DATA IF(DATA>R)
 →L1 IF(|P[3])≤1E¯6
 DUM1←((⍴M1)×⍟(0.1÷PA[1]×P[2]))-(1+÷P[3])×+/⍟(1+P[3]×(L-M1)÷P[2]×P
 [1])
 →L2
L1:DUM1←((⍴M1)×⍟(0.1÷PA[1]×P[2]))-+/((L-M1)÷P[2]×PA[1])
L2:DUM2←((⍴M2)×⍟÷P[2])++/((PA[2]×(M2-P[1])×(M2-P[1])÷P[2]×P[2])+PA
 3])
```

```
 →L3 IF(|P[3])≤1E¯6
 DUM3←((ρM3)×⊛(0.1÷PA[1]×P[2]))-(1+÷P[3])×+/⊛(1+P[3]×(M3-R)÷P[2]×PA
 [1])
 →L4
L3:DUM3←((ρM3)×⊛(0.1÷PA[1]×P[2]))-+/((M3-R)÷P[2]×PA[1])
L4:Z←DUM1+DUM2+DUM3
 →0
L5:Z←¯9999999999999
 →0
 ∇
```

*** NEWESTAB ***

```
 ∇ Z←NEWESTAB G;I;F;A;B;C
⍝-------------------------------------------------------
⍝ CALCULATE C,A,B USING TABLE TAB
⍝ NEWESTAB (G)
⍝-------------------------------------------------------
 →LOVER IF G≥1.9
 →LBELOW IF G<¯0.499
 F←(G+0.5)×1000
 I←⌊F
 C←TAB[I]+(TAB[I+1]-TAB[I])×(F-I)
 A←-(1+G)÷2×C
 B←-A+(⊛C)+2.302585093
 Z←C,A,B
 →0
LOVER:Z←ESTAB G,¯1.68605
 →0
LBELOW:Z←ESTAB G,¯0.64184
 →0
 ∇
```

*** ESTAB ***

```
 ∇ Z←ESTAB P;C;G;A;B;INT;M1;M2;M
⍝-----------------------------------------
⍝ CALCULATE C,A,B FOR NPS DISTRIBUTION
⍝ (C,A,B)←ESTAB (G,AI)
⍝-----------------------------------------
 G←P[1]
 INT←0.1
 A←P[2]
 →L4 IF G=¯1
L1:A←A+(-INT),0,INT
 ⍝'A←A⌊¯.000001' IF G>¯1
 ⍝'A←A⌈.000001' IF G<¯1
 B←(-A)+⍟-A÷5×1+G
 →L2 IF 0≤⌈/A
 M1←(*B)×(-3.141592654÷A)*0.5
 M2←(-2×A)*0.5
 M←M1×(NDTR M2)-(NDTR-M2)
 →L3
L2:M←(A,B) INTEG((-1),1)
L3:M←|M-0.8
 DUM←M=⌊/M
 ⍝'INT←INT÷2' IF 1=DUM[2]
 A←1↑DUM/A
 →L1 IF(⌊/M)>1E¯6
 C←-(1+G)÷2×A
 B←-(2.302585093+⍟C)+A
 →LAST
```

```
L4:A←0
 B←-0.91629
 C←1+4
LAST:Z←C,A,B
 →0
 ∇
```

*** NDTR ***

```
 ∇ P←NDTR X;T
⍝THIS PROGRAM COMPUTES THE AREA UNDER THE CURVE OF THE STANDARD NOR
 MAL DENSITY.
 T←÷1+0.2316419×|X
 P← 0.3193815 ¯0.3565638 1.781478 ¯1.821256 1.330274
 P←|(X≥0)-(0.3989423×*¯0.5×X×X)×(T∘.*⍳5)+.×P
 ∇
```

*** INTEG ***

```
 ∇ Z←P INTEG VEC1;H;M1;M2;M3;M4;FROM;TO;SUM1;SUM2;SUM3;SUM;TEMP;OLD
 ;NEW;S1;S2;NUM;STEP
⍝-------------------------------------------------------------
⍝   INTEGRATION PROGRAM FOR ESTAB (MODIFIED SIMPSON'S RULE)
⍝-------------------------------------------------------------
 XA←P[1 2 3]
 XB←P[4 5 6]
 FROM←VEC1[1]
 TO←VEC1[2]
 M1←TO-FROM
 NUM←4
 H←M1÷8
 M3←FROM+H×(1 3 5 7)
 M4←FROM+H×2×(⍳3)
```

```
 S1←50⌊((XA∘.×(M3×M3))+(XB∘.×(4ρ1)))
 SUM1←+/*S1
 S2←50⌊((XA∘.×(M4×M4))+(XB∘.×(3ρ1)))
 SUM2←+/*S2
 SUM3←(*50⌊(XA×(FROM*2))+XB)+(*50⌊(XA×(TO*2))+XB)
 SUM←SUM3+(4×SUM1)+2×SUM2
 OLD←SUM×H÷3
 STEP←0
LOOP:H←H÷2
 STEP←STEP+1
 NUM←NUM×2
 TEMP←SUM-SUM1×2
 M3←FROM+H×(2×(⍳NUM))-1
 S1←50⌊((XA∘.×(M3×M3))+(XB∘.×(NUMρ1)))
 SUM1←+/*S1
 SUM←TEMP+4×SUM1
 NEW←SUM×H÷3
 TAG←NEW≤1
 TEMP←⌈/|(NEW-OLD)×TAG
 OLD←NEW
 →LOOP IF(TEMP≥1E¯8)∧(STEP≤10)
 Z←NEW
 ∇
```

*** ORDERSTAT **

```
 ∇ Z←P ORDERSTAT X;Y;N
 N←ρX
 Y←X[⍋X]
 Z←Y[⌈N×P]
 ∇
```

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Date Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER: 37	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle): A Robust Estimator Of Location Using An Adaptive Spline Model		5. TYPE OF REPORT & PERIOD COVERED: Technical Report
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s): Dong Yoon Kim		8. CONTRACT OR GRANT NUMBER(s): N00014-75-C-0555
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Statistics Center, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: (NR-609-001)
11. CONTROLLING OFFICE NAME AND ADDRESS: Office of Naval Research, Statistics and Probability Code 436, Arlington, Virginia 22217		12. REPORT DATE: March 1985
		13. NUMBER OF PAGES: 94
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)

Pareto Tails, Spline, Outlier, Robust Adaptive Estimation

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

See reverse side.

DD FORM 1473 EDITION OF 1 NOV 65 IS OBSOLETE

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Date Entered)

ABSTRACT

This paper contains a new approach toward the robust estimation of a location parameter. We propose NPS (Normal Pareto Spline) distribution which provides rough fit to density functions for arbitrary unimodal symmetric distributions. The bases of our NPS estimation are Pareto tails and spline constraints. Pareto tails can represent a diversity of tail behavior, and spline constraints ensure the smoothness of the density function.

We show that the NPS estimate of location has lower asymptotic variance than Huber's M-estimator in most cases, regardless of how Huber's trimmed constant k is chosen.

We also show that the NPS estimate of location can guarantee resistance for outliers.

For the generalized two sample location problem, where the scale parameters are unequal, we propose an iterative method to estimate the shift parameter and also have a proof that this iterative method converges to the desired M-estimate for an arbitrary scale location family of symmetric distributions.