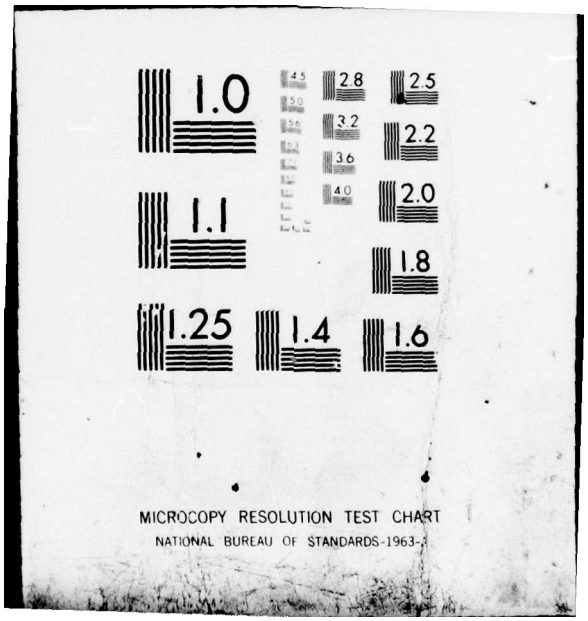
AD-A079 732 WISCONSIN UNIV-MADISON MATHEMATICS RESEARCH CENTER F/G 12/1
IMPLIED ASSUMPTIONS FOR SOME PROPOSED ROBUST ESTIMATORS.(U)
SEP 79 G CHEN, G E BOX DAAG29-75-C-0024
UNCLASSIFIED MRC-TSR-1997 NL

END
DATE FILMED
2-80
DDC


1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0
4.5
5.0
5.6
MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

(3)

AD A079732

MRC Technical Summary Report #1997

IMPLIED ASSUMPTIONS FOR SOME PROPOSED ROBUST ESTIMATORS

Gina Chen and George E. P. Box

LEVEL II

Mathematics Research Center
University of Wisconsin—Madison
610 Walnut Street
Madison, Wisconsin 53706

September 1979

(Received July 27, 1979)

Approved for public release
Distribution unlimited

Sponsored by

U.S. Army Research Office
P.O. Box 12211
Research Triangle Park
North Carolina 27709

UNIVERSITY OF WISCONSIN-MADISON
MATHEMATICS RESEARCH CENTER

IMPLIED ASSUMPTIONS FOR SOME PROPOSED ROBUST ESTIMATORS

Gina Chen and George E. P. Box

Technical Summary Report # 1997
September 1979

ABSTRACT

Assumptions which could motivate various L-estimators and M-estimators are discussed. In particular, for samples of size ten a distribution in the contaminated exponential power family is found whose posterior mean approximates each of a number of proposed L-estimators. Also distributions in this family are found whose posterior modes approximate suggested M-estimators.

AMS (MOS) Subject Classification - 62G35

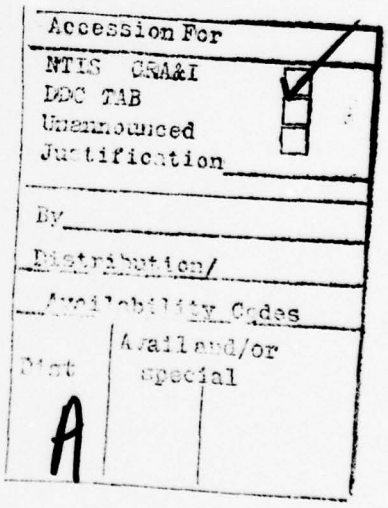
Key Words - Robust estimators, L-estimators, M-estimators, Contaminated normal model, Exponential power distribution, Bayesian approach

Work Unit Number 4 - Probability, Statistics, and Combinatorics

Sponsored by the United States Army under Contract No. DAAG29-75-C-0024.

SIGNIFICANCE AND EXPLANATION

In recent years attempts have been made to obtain "robust estimators"; that is, functions of the data which produce estimates which are less sensitive to non-normality of the error distribution and to occasional "bad" observations. Some of the methods proposed are highly empirical. In this paper an attempt is made to link the variously proposed robust estimators to particular models. It is then possible to say of a given estimator that it would be applicable if some specified model were believed roughly to represent the process of data generation.



The responsibility for the wording and views expressed in this descriptive summary lies with MRC, and not with the authors of this report.

IMPLIED ASSUMPTIONS FOR SOME PROPOSED ROBUST ESTIMATORS

Gina Chen and George E. P. Box

1. Introduction

Statistical inferences should depend on the parent distribution from which the data are assumed to come. The sample mean, for example, is a good estimator for the location parameter of the normal distribution, but is not necessarily good for the double exponential or rectangular distribution. Therefore, in doing data analysis, one always faces the difficulty of making appropriate inferences because the parent distribution of the data is never known.

The Normal distribution was introduced by Gauss as follows: (Gauss (1821), also Huber (1972))

"The author of the present treatise, who in the year 1797 first investigated this problem according to the principles of the theory of probability, soon realized that it was impossible to determine the most probable value of the unknown quantity, unless the function representing the errors is known. But since it is not, there is no other recourse than to assume such a function in a hypothetical fashion. It seemed most natural to him to take the opposite approach and to look for that function which must be taken as a base in order that for the simplest of all cases a rule is obtained which is generally accepted as a good one, namely, that the arithmetic mean of several observations of equal accuracy for one and the same quantity should be considered the most accurate value. This implied that the probability of an error x must be

Sponsored by the United States Army under Contract No. DAAG29-75-C-0024.

assumed proportional to an exponential expression of the form e^{-hhxx}, and that then just the same method which he had found by other considerations already a few years earlier, would become necessary in general."

Note that here Gauss introduced the normal distribution to suit the sample mean — an estimator which was thought to be good. Over the years, many other estimators for location have been proposed and claimed to be good on empirical basis. For each such estimator, one could ask the question "Which distribution is this estimator suitable for?" or alternatively "What type of distribution does the advocate of the estimator really have in mind?" We discuss this problem in this chapter.

We shall consider a class of distributions capable of representing a wide range of behavior. For each of a number of proposed estimators, we will then try to find a distribution in this class for which the estimator is appropriate. We will now be more specific.

2. Some Robust Estimators for Location

Robust estimators are estimators which are believed to be good for a broad class of distributions but not necessarily best for any one of them. There are different methods for constructing a robust estimator.

Suppose $x_1, x_2, \ldots, x_N$ are observations drawn from some unknown distribution, and $X_{(1)}, X_{(2)}, \ldots, X_{(N)}$ are the ordered

observations. The following are some of the robust estimators which have been proposed for the location parameter of a single sample.

(i) "Maximum likelihood type" estimators (M-estimators)

If the distribution is known to be $F(\frac{x-\theta}{\sigma})$ with density $f(\frac{x-\theta}{\sigma})$, then the maximum likelihood estimate of θ is given by $\hat{\theta}$ which maximizes the following equation:

$$\prod_{i=1}^{N} f(\frac{x_i-\theta}{\sigma})$$

or equivalently maximizes

$$\sum_{i=1}^{N} \log f(\frac{x_i-\theta}{\sigma}) .$$

$\hat{\theta}$ will then satisfy the following equation:

$$\sum_{i=1}^{N} -\frac{f'(\frac{x_i-\theta}{\sigma})}{f(\frac{x_i-\theta}{\sigma})} = 0 \quad \text{or} \quad \sum_{i=1}^{N} \psi(\frac{x_i-\theta}{\sigma}) = 0 \quad \text{with} \quad \psi = -\frac{f'}{f} .$$

An M-estimator $(\hat{\theta})$ was defined by Huber to be the solution of an equation

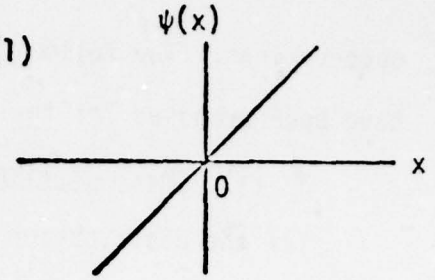
$\sum_{i=1}^{N} \psi(\frac{x_i-\hat{\theta}}{s}) = 0$, where ψ is some specifically chosen function.

In general, $\hat{\theta}$ is solved by iteration, and s is some estimate of the scale parameter. Some examples of the ψ function chosen for M estimators are:

(1) $\psi(x) = x$

This is the special case that gives as estimator the sample mean.

(1)

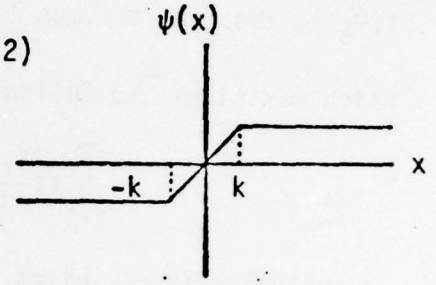


(2) Huber's (Huber 1964 , Andrews et al. 1972)

$$\psi(x,k) = \begin{cases} -k, & x < -k \\ x, & -k < x < k \\ k, & k < x \end{cases}$$

with $k = 1.5$ or 2.0.

(2)

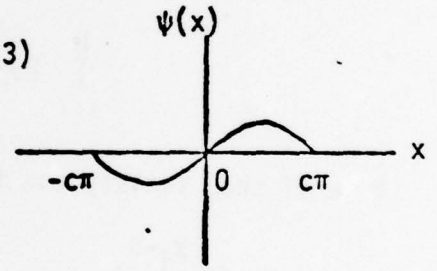


(3) Andrews' (Andrews et al. 1972)

$$\psi(x,c) = \begin{cases} \sin(x/c), & |x| \le c\pi \\ 0 & \text{o.w.} \end{cases}$$

with $c = 2.1 \times .6754$
$= 1.42.$

(3)



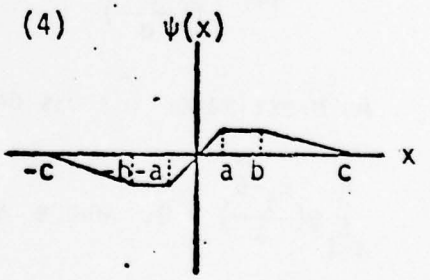
(4) Hampel's (c.f. Andrews et al. 1972)

$\psi(x,a,b,c)$

$$= \operatorname{sgn} x \begin{cases} |x|, & 0 \le |x| < a \\ a, & a \le |x| < b \\ \frac{c-|x|}{c-b}a, & b \le |x| < c \\ 0, & |x| \ge c \end{cases}$$

with $a = 1.69$ $b = 3.04$ $c = 6.42$
or $a = 1.49$ $b = 2.50$ $c = 3.98$
or $a = 1.42$ $b = 2.70$ $c = 5.54$

(4)

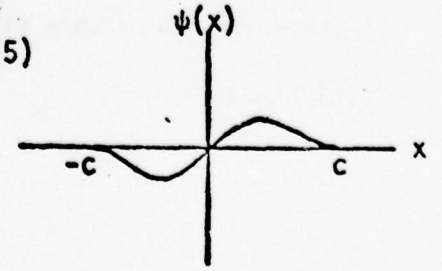


(5) Tukey's biweight (Beaton & Tukey (1974))

$$\psi(x;c) = \begin{cases} x[1-(\frac{x}{c})^2]^2, & |x| \le c \\ 0 & \text{o.w.} \end{cases}$$

with $c = 4.05$, 5.40 or 6.10.

(5)



The values given for the constants k, a, b, c, etc. are those most commonly used. In Hampel's, Andrews' and Tukey's M estimators, s is usually estimated by median of the absolute deviations from the median. Such a robust scale estimate has expection $.6754\sigma$ under normality and Huber standardizes by dividing by this constant. For comparison purpose, we have adjusted the values of a, b, and c such that in all cases the scale estimate used has expected value σ on the assumption of normality.

(ii) Linear combinations of order statistics (L-estimators)

These estimators have the form $\sum_{i=1}^{N} a_i X_{(i)}$, where $X_{(i)}$ is the i^{th} order statistics.

r is defined as $N\alpha$ and $0 \leq \alpha \leq \frac{1}{2}$ for the following estimators.

(1) Trimmed Mean (Crow & Siddiqui 1967)

$$T_N(\alpha) = (N-2r)^{-1} \sum_{i=r+1}^{N-r} X_{(i)} .$$

If r is not an integer, the following form is taken (Andrews et al,1972)

$$T_N(\alpha) = \frac{(1+[r]-r)X_{([r+1])}+X_{([r+2])}+\ldots+(1+[r]-r)X_{(N-[r])}}{N(1-2\alpha)} .$$

(2) Winsorized Mean (c.f. Crow & Siddiqui, 1967; Tukey, 1962)

When r is an integer,

$$W_N(\alpha) = \frac{1}{N}\left\{(r+1)(X_{(r+1)}+X_{(N-r)})+\sum_{i=r+2}^{N-r-1} X_{(i)}\right\}$$

if N is odd and $\alpha = \frac{(N-1)}{2N}$, then $W_N(\alpha)$ is the median.

(3) Linearly Weighted Mean (Crow & Siddiqui 1967)
(Smoothly Trimmed Mean I (Stigler 1973))

When r is an integer, $N = 2n$

$$L_N(\alpha) = \tfrac{1}{2}(n-r)^{-2}[X_{(r+1)}+X_{(N-r)}+3(X_{(r+2)}+X_{(N-r-1)}+\cdots$$

$$+(2i-2r-1)(X_{(i)}+X_{(N-i+1)})+\ldots+(2n-r-1)(X_{(n)}+X_{(n+1)})]$$

$N = 2n+1$

$$L_N(\alpha) = [(n-r)^2+(n-r+1)^2]^{-1}[X_{(r+1)}+X_{(N-r)}+3(X_{(r+2)}+X_{(N-r+1)})$$

$$+\ldots+(2i-2r-1)(X_{(i)}+X_{(N-i+1)})+\ldots+(2n-2r-1)$$

$$\times(X_{(n)}+X_{(n+2)})+(2n-2r+1)X_{(n+1)}] .$$

In particular, when $\alpha = 0$, $r = 0$

$$L_N(0) = \frac{1}{(n+1)^2}\left[\sum_{i=1}^{n} i(X_{(i)}+X_{(2n-i)})+nX_{(n)}\right] \quad N = 2n+1$$

$$\frac{1}{n(n+1)}\left[\sum_{i=1}^{n}[i(X_{(i)}+X_{(2n+1-i)})]\right] \quad N = 2n .$$

(Crow 1964)

(4) Smoothly Trimmed Mean II (Stigler 1973)

$$L_N(\alpha) = \sum_{i=1}^{N} W_i X_{(i)} \quad \text{with}$$

$$W_i = \begin{cases} (i - \frac{r}{2})c & \text{if} \quad \frac{r}{2} < i \leq r \\ (\frac{r}{2} + 1)c & \text{if} \quad r < i \leq N-r \\ (N-i+1 - \frac{r}{2})c & \text{if} \quad N+1-r \leq i < N+1 - \frac{r}{2} \\ 0 & \text{o.w.} \end{cases}$$

where c is a normalizing constant

$$c = \frac{1}{N - \frac{3}{4}r^2 - \frac{3}{2}r + \frac{N}{2}r} .$$

This is expressed differently from the form given by Stigler for consistency in notation, however, the criterion is the same.

(5) Squared Weighted Mean (Crow 1964)

$$S_N = \frac{3}{n(2n^2+1)} \left[\sum_{i=1}^{n-1} i^2(X_{(i)}+X_{(2n-i)}) + n^2 X_{(n)} \right] \qquad N = 2n-1$$

$$\frac{3}{n(n+1)(2n+1)} \left[\sum_{i=1}^{n} i^2(X_{(i)}+X_{(2n+1-i)}) \right] \qquad N = 2n$$

(6) Cubic Weighted Mean

$$C_N = \frac{2}{n^2(n^2+1)} \left[\sum_{i=1}^{n-1} i^3(X_{(i)}+X_{(2n-i)}) + n^3 X_{(n)} \right] \qquad N = 2n-1$$

$$= \frac{2}{n^2(n+1)^2} \sum_{i=1}^{n} [i^3(X_{(i)}+X_{(2n+1-i)})] \qquad N = 2n .$$

(iii) Estimator derived from rank tests (R-estimators)

Consider a 2-sample rank test for shift. $y_1,\ldots,y_n$ and $z_1,\ldots,z_n$ are two independent samples with distribution $F(x)$ and $F(x-\Delta)$ respectively. To test $\Delta = 0$ against $\Delta > 0$, the following test statistic can be used

$$W(y_1,\ldots,y_n;z_1,\ldots,z_n) = \sum_{i \leq 2n} J(\frac{i}{2n+1})V_i$$

where $V_i = 1$ if the i^{th} smallest entry in the combined sample is a y, and $V_i = 0$ otherwise. And J is some function defined on $[0,1]$ such that $J(1-t) = -J(t)$.

An estimate for the location parameter can be derived from such tests. Determine estimate $T_n(x_1,x_2,\ldots,x_n)$ such that

$$W(x_1-T_n,\ldots,x_n-T_n;-(x_1-T_n),-(x_2-T_n),\ldots,-(x_n-T_n)) = 0 .$$

(iv) Adaptive estimators (c.f. Hogg 1974)

This is a class of estimates which will adapt to the data. For example, selecting the trimming proportion of a trimmed mean after the sample is drawn.

We shall consider only L-estimators and M-estimators in this study, but as we have seen even in these classes there are a confusingly large number of candidates. One way of better understanding their relative merits is as follows.

The originators of these estimates seem to believe that the parent distributions of the real world are likely to be non-normal and in particular to be heavy tailed. It would be valuable if these implied beliefs could be brought out into the open, examined and perhaps compared with practical reality. In order to do this we need first to be able to parameterize non-normality.

3. Distributions

In past studies of robust estimators, for example Crow (1964), Crow and Siddiqui (1967), Gastwirth and Cohen (1970), Andrets et al. (1972), distributions employed have varied from light-tailed distributions such as the rectangular to heavy-tailed distributions such as the double exponential, Cauchy and contaminated normal distributions. A convenient class which includes both light and heavy tailed distributions which we will employ in our study is the exponential power family (c.f. Box and Tiao, 1973), the density functions for which are given by:

$$P(y|\theta,\sigma,\beta) = \omega(\beta)\sigma^{-1}\exp\left[-c(\beta)\left|\frac{y-\theta}{\sigma}\right|^{2/(1+\beta)}\right] \quad -\infty < y < \infty$$

where $$c(\beta) = \left\{\frac{\Gamma[\frac{3}{2}(1+\beta)]}{\Gamma[\frac{1}{2}(1+\beta)]}\right\}^{1/(1+\beta)}$$ and $$\omega(\beta) = \frac{\{\Gamma[\frac{3}{2}(1+\beta)]\}^{1/2}}{(1+\beta)\{\Gamma[\frac{1}{2}(1+\beta)]\}^{3/2}}$$

$$\sigma > 0 \ , \quad -\infty < \theta < \infty \ , \quad -1 < \beta \leq 1 \ .$$

In the above expression, θ is a location parameter and σ is a scale parameter. The parameter β can be regarded as a measure of kurtosis indicating the extent of 'non-normality' of the distribution. As β goes from -1 to 1, the distribution goes from platykurtic to leptokurtic. In the limiting case $\beta \to -1$, the distribution tends to the rectangular, for $\beta = 0$ it is normal and for $\beta = 1$ it is double exponential distribution.

4. What do estimators imply about $P(\beta)$

Suppose it could be assumed that the parent distribution was a member of the exponential power family and that a prior distribution could be written down for $P(\beta)$ which represented the probability of occurrence of different values of β in the particular experimental situation in the study. Then after putting a non-informative prior on θ and σ and following the Bayesian procedure and integrating β and σ out, the posterior distribution of θ would be obtained. A point estimate of θ would then be given by the posterior mean which minimizes squared error loss. Minimization of other loss functions could be achieved by using other features of the posterior distribution, but we shall suppose throughout this study that the squared error loss function is appropriate.

Now if for a particular ad hoc estimator of θ, we could find a prior distribution $P(\beta)$ such that the resulting posterior

mean closely approximates this estimator, then we could say that in using this estimator the statistician behaved as if he believed that $P(\beta)$ represented this distribution of distributions in the real world. This $P(\beta)$ could then be considered for its reasonableness and possibly for its concordance with distributions that actually occurred in the real world. In this case we shall say that we have found "A Bayesian formulation associated with the estimator."

Contaminated Exponential Power Distribution

Greater flexibility in the assumed form of distribution could be obtained by allowing for contamination in the following way.

Consider a distribution with density

$$P_c(y|\theta,\sigma,\beta,\alpha,k) = (1-\alpha)P(y|\theta,\sigma,\beta) + \alpha P(y|\theta,k\sigma,\beta) \quad .$$

That is, with probability $(1-\alpha)$ the observation comes from an exponential power distribution with kurtosis parameter β, mean θ and variance σ^2 and with probability α the observation comes from an exponential power distribution with the same parameter β, mean θ but a much larger scale parameter $k\sigma$.

Posterior Distribution of θ

Consider the distribution $P_c(y|\theta,\sigma,\beta,\alpha,k)$ with prior distribution $P(\beta)$ on β. Assume that β is distributed

independently of θ and σ, so that the prior distribution for θ, σ, β is $P(\theta,\sigma,\beta) = P(\beta)P(\theta,\sigma)$. Adopt as noninformative prior for (θ,σ); $P(\theta,\sigma) \propto \sigma^{-1}$. Then the joint posterior distribution of (θ,σ,β) for a chosen α and k is

$$P(\theta,\sigma,\beta|Y,\alpha,k) = \frac{\sigma^{-1}P(\beta)\pi_i P_c(y_i|\theta,\sigma,\beta,\alpha,k)}{\iiint \sigma^{-1}P(\beta)\pi_i P_c(y_i|\theta,\sigma,\beta,\alpha,k)d\theta d\sigma d\beta} .$$

Also the marginal posterior distribution of θ given α and k is

$$P(\theta|Y,\alpha,k) = \iint \frac{\sigma^{-1}P(\beta)\pi_i P_c(y_i|\theta,\sigma,\beta,\alpha,k)}{\iiint \sigma^{-1}P(\beta)\pi_i P_c(y_i|\theta,\sigma,\beta,\alpha,k)d\theta d\sigma d\beta} d\sigma d\beta$$

$$= \int \frac{P(\theta,\beta,Y|\alpha,k)}{P(Y|\alpha,k)} d\beta = \int \frac{P(\beta,Y|\alpha,k)}{P(Y|\alpha,k)} \cdot \frac{P(\theta,\beta,Y|\alpha,k)}{P(\beta,Y|\alpha,k)} d\beta$$

$$= \int P(\beta|Y,\alpha,k)P(\theta|Y,\beta,\alpha,k)d\beta .$$

Thus the posterior mean of θ with a given prior $P(\beta)$ is the weighted average of the posterior mean for each given β value with the weighting function $P(\beta|Y,\alpha,k)$. Also if the sample size is not large there will be little information about β coming from the sample and $P(\beta|Y,\alpha,k)$ can be approximated by $P(\beta)$. The results are not particularly sensitive to k (Box and Tiao (1968)) and we set it equal to a constant 3 for this part of the study. For the time being therefore the parameter k is treated as a known constant and dropped from the formulas.

Approximation of a Posterior Mean by a Linear Combination of Ordered Observations

To compare the estimates given by linear combinations of order statistics with the posterior mean derived from a particular parent distribution, we need to first approximate the posterior mean with a weighted average of the ordered observations. Suppose the parent distribution is $F(y|\theta,\sigma)$, then given n ordered observations $Y = (y_{(1)},\ldots,y_{(n)})$, we can calculate the posterior mean M_Y. If, for any sample Y from $F(y|\theta,\sigma)$, M_Y is approximately equal to $\sum_{i=1}^{n} w_i y_{(i)}$ for some fixed set of w_i, then we can say F is "a distribution associated with the estimator $\sum_{i=1}^{n} w_i y_{(i)}$."

In particular, the posterior mean for a normal distribution is always equal to the sample mean = $\sum \frac{1}{n} y_{(i)}$. Thus we can say that a distribution associated with the sample mean is the normal distribution.

To obtain an approximation more generally one can proceed as follows:

Given a distribution, take a random sample of size n from the distribution, and denote the ordered sample by $Y_1 = (y_{1(1)}, \ldots, y_{1(n)})$, calculate the posterior mean M_{Y_1}. Take a second random sample and have them ordered and denoted by $Y_2 = (y_{2(1)}, \ldots, y_{2(n)})$, then calculate the posterior mean M_{Y_2}. Repeat the

procedure m times; we get m sets of ordered sample $Y_1, Y_2, \ldots, Y_m$ and their corresponding posterior means $M_{Y_1}, M_{Y_2}, \ldots, M_{Y_m}$. The relationship we'd like to establish is $M \approx w_1 y_{(1)} + \ldots + w_n y_{(n)}$ where M is the posterior mean of the sample $(y_{(1)}, \ldots, y_{(n)})$. More specifically, we are trying to find a set of weights $(w_1, \ldots, w_n)$ such that for any sample from $F(y|\theta,\sigma)$ the posterior mean is approximately the weighted average of the ordered observations with this set of weights. Regressing $\underset{\sim}{M_Y} = \begin{pmatrix} M_{Y_1} \\ \vdots \\ M_{Y_m} \end{pmatrix}$ on $\underset{\sim}{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_m \end{pmatrix}$, we get a set of estimated weights $(\hat{w}_1, \ldots, \hat{w}_n)$. Obviously, these estimated weights $(\hat{w}_1, \ldots, \hat{w}_n)$ depend on the m sets of random samples chosen, the $\hat{w}_i$'s would be different if different samples had been taken. However, it is shown in Appendix 2.A that there is a set of limiting estimates of the weights as $m \to \infty$ and that these limiting weights are the same as those weights that give rise to the Ordered Linear Unbiased Minimum Variance Estimator (OLUMVE). These weights can thus be calculated without going through the above sampling procedures. (The method of calculation is given later.) And from now on

we should keep in mind that the posterior mean is approximately the OLUMV estimator for samples from $F(y|\theta,\sigma)$.

So for a contaminated exponential power distribution with fixed β,α, the posterior mean $M_{\beta,\alpha}$ is approximately $w_{1\beta\alpha}y_{(1)}+w_{2\beta\alpha}y_{(2)}+\ldots+w_{n\beta\alpha}y_{(n)}$ where $(w_{1\beta\alpha},\ldots,w_{n\beta\alpha})$ are the weights for the OLUMV estimator. Furthermore, if a prior $P(\beta)$ was put on β, then the posterior mean can be written as

$$\begin{aligned} M &= \int\theta P(\theta|Y,\alpha)d\theta \\ &= \int\theta(\int P(\theta|Y,\beta,\alpha)P(\beta|Y,\alpha)d\beta)d\theta \\ &= \int P(\beta|Y,\alpha)(\int\theta P(\theta|Y,\beta,\alpha)d\theta)d\beta \\ &= \int P(\beta|Y,\alpha)(w_{1\beta\alpha}y_{(1)}+w_{2\beta\alpha}y_{(2)}+\ldots+w_{n\beta\alpha}y_{(n)})d\beta \\ &= (\int P(\beta|Y,\alpha)w_{1\beta\alpha}d\beta)y_{(1)}+\ldots+(\int P(\beta|Y,\alpha)w_{n\beta\alpha}d\beta)y_{(n)} . \end{aligned}$$

When the sample size is not large, the posterior distribution $P(\beta|Y,\alpha)$ will be dominated by the prior $P(\beta)$, and $P(\beta|Y,\alpha)\approx P(\beta)$. In this case, the posterior mean is approximately $w_1y_{(1)}+\ldots+w_ny_{(n)}$ with $w_i = \int P(\beta)w_{i\beta\alpha}d\beta$. We then conclude that $w_1y_{(1)}+\ldots+w_ny_{(n)}$ is a good estimator for the contaminated exponential power distribution with prior $P(\beta)$ on β and parameter α. Conversely, if we are given an L-estimator $w_1y_{(1)}+\ldots+w_ny_{(n)}$, it is also possible to find a prior distribution $P(\beta)$ and a value α such that $w_i \approx \int P(\beta)w_{i\beta\alpha}d\beta$. This $P(\beta)$ and α will then lead

to a posterior mean close to $w_1 y_{(1)} + \ldots + w_n y_{(n)}$. Such a prior is of course not unique. For simplicity therefore we proceed by supposing that $P(\beta)$ belonged to the family of beta function priors

$$P(\beta) = \frac{\Gamma(p+q+2)}{2 \cdot \Gamma(p+1)\Gamma(q+1)} \left(\frac{1+\beta}{2}\right)^p \left(\frac{1-\beta}{2}\right)^q \qquad -1 < \beta < 1.$$

This family contains two adjustable parameters p and q. At this point in the investigation therefore the class of non-normal distributions is defined by three parameters which could be conveniently thought of as α and the mean and variance of $P(\beta)$. (Recall k was fixed at 3). A comprehensive numerical investigation however showed that the crucial factor was where the prior was centered, the variance of the prior distribution did not have much influence on the weights. In what follows, therefore, we shall employ a point prior for the distribution for β.

Simplified Problem

Fixing the variance parameter for $P(\beta)$ at zero, we were left with a point prior at (β,α). It was thus possible to find a single contaminated exponential power distribution for which the Bayesian posterior mean was almost the same as the L-estimator. In many cases where the likelihood function is almost symmetric, the posterior mean is essentially the same as the maximum likelihood estimate. We subsequently concentrated therefore on finding a contaminated exponential power distribution associated with each L-estimator.

5. The Contaminated Exponential Power Distribution Associated with the L-estimator

The problem now reduces to the following:

Given an L-estimator $T = w_1 y_{(1)} + \ldots + w_n y_{(n)}$, find α and β such that T is approximately OLUMV estimator for the contaminated exponential power distribution with parameters α and β.

Method of Computing the Weights for OLUMVE

Suppose $Y = (y_{(1)}, \ldots, y_{(n)})'$ is an ordered sample from a distribution $F(\frac{y-\theta}{\sigma})$. Let $u_{(i)} = \frac{y_{(i)}-\theta}{\sigma}$, then $U = (u_{(1)}, \ldots, u_{(n)})'$ is an ordered sample from $F(y)$. Set

$$\underset{\sim}{1} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}_{n\times 1} \qquad \underset{\sim}{a} = EU = \begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix}_{n\times 1} \qquad V = \mathrm{Var}(U) = (v_{ij})_{n\times n}$$

where

$$v_{ij} = \mathrm{Cov}(u_{(i)}, u_{(j)})$$

$$A = (\underset{\sim}{1}, \underset{\sim}{a}) \ .$$

Then $Y = \theta\underset{\sim}{1} + \sigma\underset{\sim}{a} + \sigma(U - \underset{\sim}{a})$

$$= A\begin{pmatrix}\theta \\ \sigma\end{pmatrix} + \underset{\sim}{\varepsilon} \qquad E(\underset{\sim}{\varepsilon}) = \underset{\sim}{0} \qquad V(\underset{\sim}{\varepsilon}) = \sigma^2 V \ .$$

By the Gauss-Markov Theorem, the best linear unbiased estimator (minimum variance) is given by

$$\begin{pmatrix} \hat{\theta} \\ \hat{\sigma} \end{pmatrix} = (A'V^{-1}A)^{-1}A'V^{-1}Y$$

in particular, when F is symmetric, $\hat{\theta} = \dfrac{\underset{\sim}{1}'V^{-1}Y}{\underset{\sim}{1}'V^{-1}\underset{\sim}{1}}$.

When the sample size is fixed, a numerical method for calculating the variance and covariance matrix (V) was given by Lund (1967). When V is known, the weights for the OLUMV estimator is then given by $\dfrac{\underset{\sim}{1}'V^{-1}}{\underset{\sim}{1}'V^{-1}\underset{\sim}{1}}$. Thus we are able to find the weights of OLUMV estimators for contaminated exponential power distributions with parameters β and α. The value of β lies between -1 and +1, and the value of α is between 0 and 1. However, α is the probability of an observation coming from some distribution with large variance, and α is expected to be small. The weights which produce OLUMV estimator when the sample size is 10 are actually calculated for all pairs of (β,α) with α = 0, .01, .025, .05, .075, .1 and β = -.99, -.75, -.5, -.25, 0, .25, .5, .75, 1.0. A plot of the first five weights $w_{1\alpha\beta},\ldots,w_{5\alpha\beta}$ ($w_{i\alpha\beta} = w_{(11-i)\alpha\beta}$ for $6 \leq i \leq 10$) and their values are shown in Figure 1. From the figure, it is clear that the weights change smoothly as (β,α) change, it is then appropriate to use polynomial interpolation to get the weights $w_{i\alpha\beta}$ for any other (β,α) with $0 \leq \alpha < .1$ and $-.99 < \beta < 1.0$.

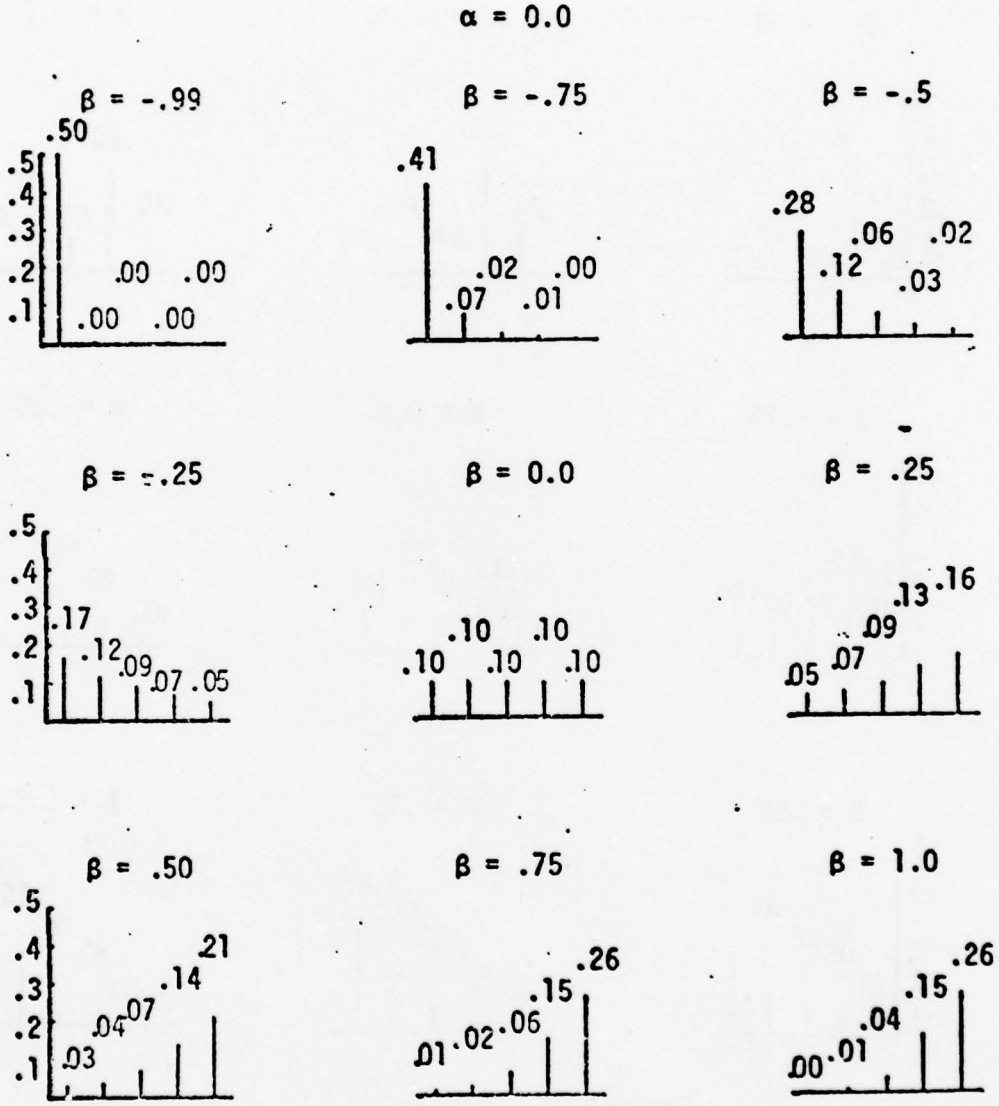
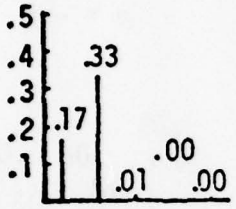


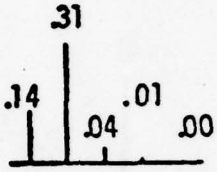
Figure 1 First five weights for OLUMV estimators when the underlying distribution is contaminated exponential power distribution with specified α and β values.

$\alpha = .01$

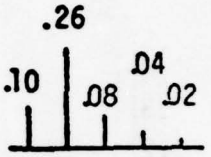
$\beta = -.99$



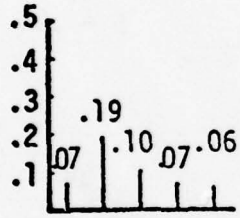
$\beta = -.75$



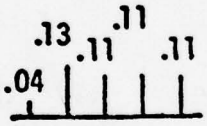
$\beta = -.5$



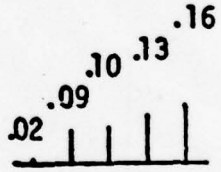
$\beta = -.25$



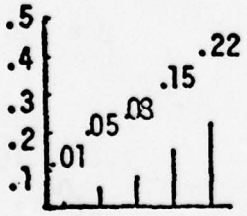
$\beta = 0.0$



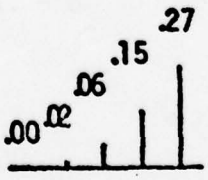
$\beta = .25$



$\beta = .50$



$\beta = .75$



$\beta = 1.0$

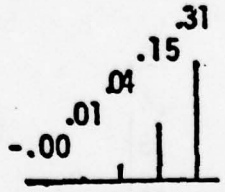
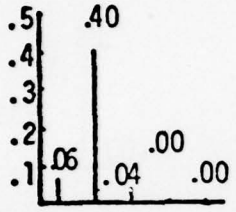


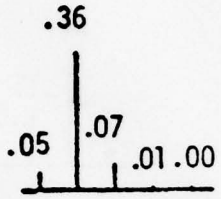
Figure 1 continued

$\alpha = .025$

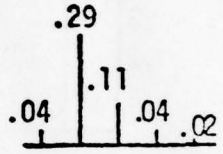
$\beta = -.99$



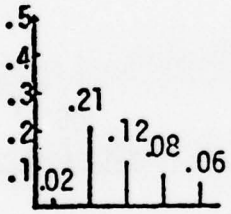
$\beta = -.75$



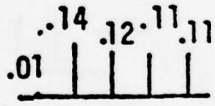
$\beta = -.50$



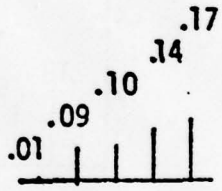
$\beta = -.25$



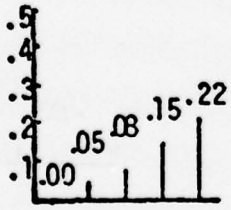
$\beta = 0.0$



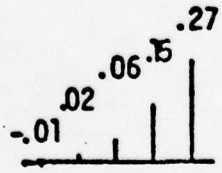
$\beta = .25$



$\beta = .50$



$\beta = .75$



$\beta = 1.0$

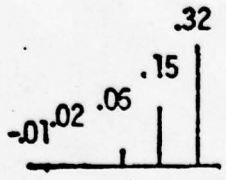
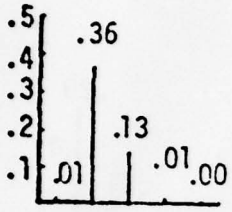


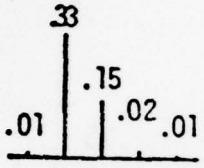
Figure 1 continued

$\alpha = .05$

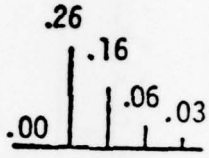
$\beta = -.99$



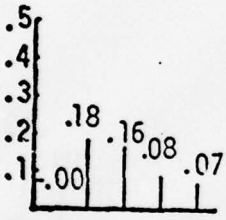
$\beta = -.75$



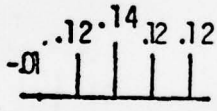
$\beta = -.50$



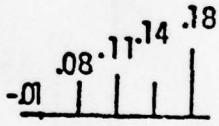
$\beta = -.25$



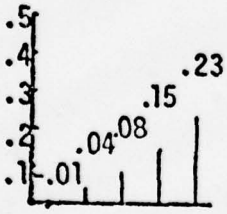
$\beta = 0.0$



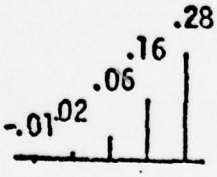
$\beta = .25$



$\beta = .50$



$\beta = .75$



$\beta = 1.0$

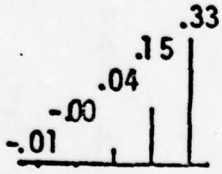
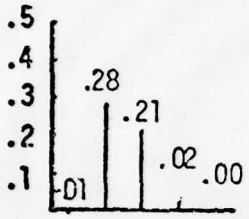


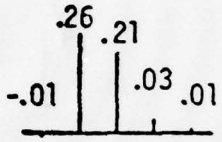
Figure 1 continued

$\alpha = .075$

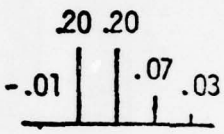
$\beta = -.99$



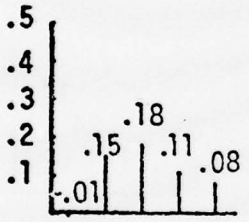
$\beta = -.75$



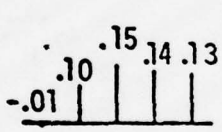
$\beta = -.50$



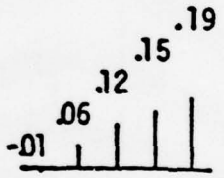
$\beta = -.25$



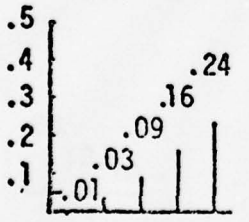
$\beta = 0.0$



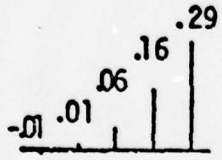
$\beta = .25$



$\beta = .50$



$\beta = .75$



$\beta = 1.0$

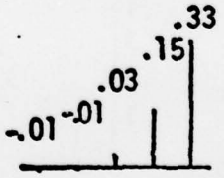
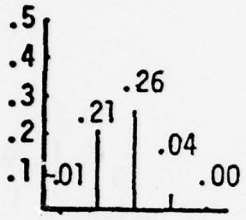


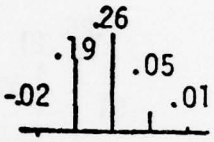
Figure 1 continued

$\alpha = .10$

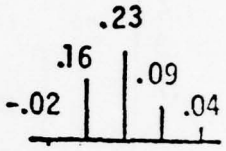
$\beta = -.99$



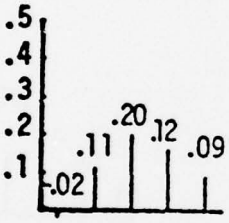
$\beta = -.75$



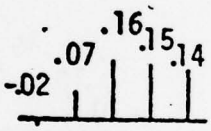
$\beta = -.50$



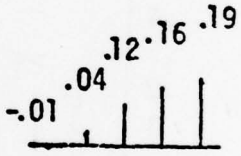
$\beta = -.25$



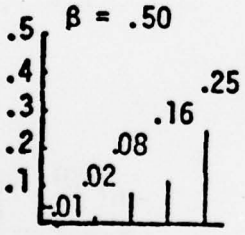
$\beta = 0.0$



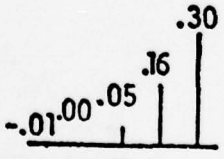
$\beta = .25$



$\beta = .50$



$\beta = .75$



$\beta = 1.0$

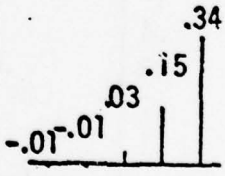


Figure 1 continued

Given an L-estimator $T = w_1 y_{(1)} + \ldots + w_n y_{(n)}$, the contaminated exponential power distribution associated with T (if there is one) would be the one with parameters (β,α) such that

$$T \approx w_{1\alpha\beta} y_{(1)} + \ldots + w_{n\alpha\beta} y_{(n)}$$

or equivalently

$$\begin{aligned} w_1 &\approx w_{1\alpha\beta} \\ &\vdots \\ w_n &\approx w_{n\alpha\beta} . \end{aligned}$$

To find (β,α), we used nonlinear regression with w_i's as the dependent variables, the coefficients of the interpolated polynomial (for predicting the weights for fixed α,β) as independent variables and α, β as the regression coefficients to be determined. The least square estimates $(\hat{\beta},\hat{\alpha})$ then specify the distribution associated with $T = w_1 y_{(1)} + \ldots + w_n y_{(n)}$.

6. The Results

The following robust L-estimators were included in this study with $n = 10$:

(1) Trimmed mean: $\alpha = 1\%, 2\%, 2.5\%, 5\%, 10\%$

(2) Linearly weighted mean: $\alpha = 0, 10\%$

(3) Smoothly trimmed mean II: $\alpha = 20\%$

(4) Squared weighted mean

(5) Cubic weighted mean

(6) Winsorized mean: $\alpha = 10\%$.

For each of the above L-estimators, following the procedure in the previous section, a distribution (specified by (β,α)) is found in the family of the contaminated exponential power distributions associated with this estimator. The parameter k has been fixed at three throughout the study. Figure 2 shows the correspondence between the L-estimators and their associated distributions. Each point in the β-α space has a coordinate (β,α) and this in turn represents a distribution in the family of contaminated exponential power distributions. For example, the squared weighted mean takes coordinate (.558, .006) and it is associated with $.994P(y|\theta,\sigma,.558) + .006P(y|\theta,3\sigma,.558)$.

Table 1 shows a comparison between the weights of the robust estimators and the weights produced by the distributions associated with the estimators. In general, the two sets of weights, to our satisfaction, are very close.

Examination of Figure 2 shows that all the trimmed means (1% - 10%) and also smoothly trimmed mean have relatively small values of β. Further insight is gained by plotting the approximate confidence contours for the least square estimate $(\hat{\beta},\hat{\alpha})$ corresponding to each L-estimator (Figure 3). It is seen that for the trimmed means, smoothly trimmed mean and linearly trimmed

L-estimator	Weights for L-estimators	Value of β,α for OLUMV estimators	Weights for (β,α)-estimator
1) 1% trimmed mean	.102 .102 .091 .102 .102	β = .011 α = .00080	.103 .102 .091 .101 .103
2) 2% trimmed mean	.104 .104 .083 .104 .104	β = .020 α = .00162	.105 .104 .084 .101 .106
3) 2.5% trimmed mean	.105 .105 .078 .105 .105	β = .026 α = .00217	.107 .105 .079 .102 .107
4) 5% trimmed mean	.111 .111 .056 .111 .111	β = .050 α = .00552	.114 .109 .057 .104 .114
5) 10% trimmed mean	.126 .126 .000 .126 .126	β = .040 α = .03820	.123 .123 .002 .127 .125

Table 1 Left side: first five weights of the L-estimator
Right side: first five weights of OLUMV estimator for the contaminated exponential power distribution (with parameters (β,α) associated with the estimator.

L-estimator	Weights for L-estimators	Value of β,α for OLUMV estimators	Weights for (β,α)-estimator
6) linearly weighted mean	.033 .067 .100 .133 .167	$\beta = .298$ $\alpha = .00399$	.035 .072 .092 .132 .169
7) 10% trimmed linearly weighted mean	.000 .044 .100 .156 .211	$\beta = .382$ $\alpha = .06865$	-.012 .045 .099 .155 .211
8) 20% smoothly trimmed mean	.000 .071 .143 .143 .143	$\beta = .052$ $\alpha = .08849$	-.009 .073 .147 .146 .145
9) squared weighted mean	.009 .036 .082 .145 .227	$\beta = .558$ $\alpha = .00633$	.010 .042 .074 .146 .227
10) cubic weighted mean	.002 .018 .060 .142 .278	$\beta = .786$ $\alpha = .01080$	.000 .020 .054 .150 .275
11) 10% winsorized mean	.000 .100 .100 .100 .200	$\beta = -.149$ $\alpha = .02451$	.020 .124 .081 .094 .180

Table 1 continued

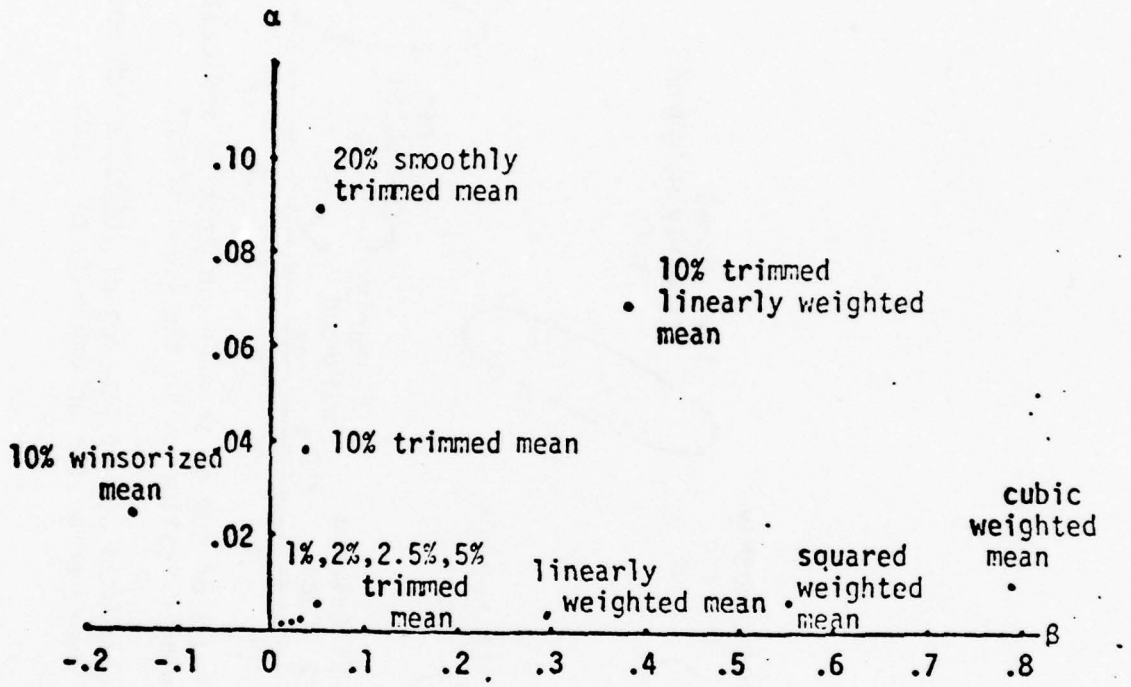


Figure 2 Values of (β,α) associated with various L-estimators

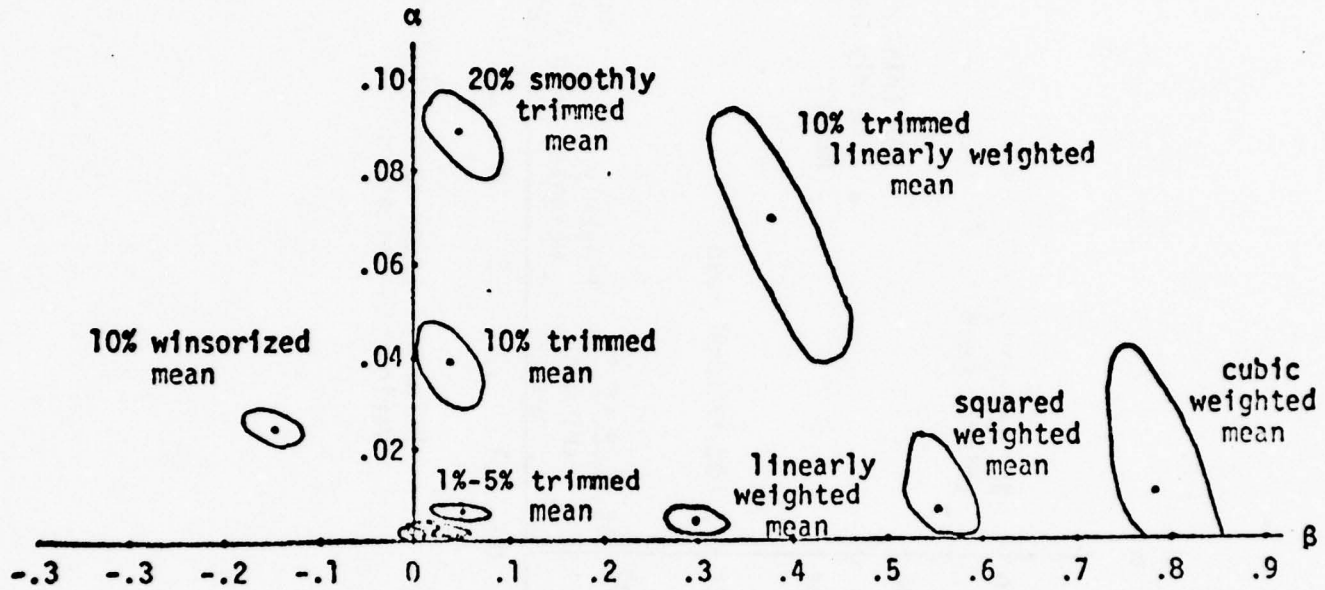


Figure 3 Typical contours of sum of squares surfaces to indicate state of conditioning of estimates in the (β,α) space*

* These correspond to an additional sum of squares of .00025 which would allow an additional root mean square error for an ordinate of .005.

mean, the contours are rather elongated in the direction of upper left to lower right, suggesting some trade-off between α and β. This is especially true when trimming is involved. Therefore, it might be possible to fix β at zero and find an α such that the distribution corresponding to $(0,\alpha)$ would approximately produce the trimmed mean as the posterior mean. In other words, for each trimmed mean we may be able to find a contaminated normal distribution for which the use of such trimmed mean is appropriate.

Restricting $\beta = 0$ and repeating the procedure used, the contaminated normal distribution associated with each trimmed mean and smoothly trimmed mean were found and similar to Figure 2 and Table 1, we now have Figure 4 and Table 2 Table 2 compares the trimmed means and the estimators produced by the contaminated normal distributions. The closeness of the two sets of weights shows that estimators from contaminated normal distributions can closely approximate trimmed means. Figure 4 gives the position each trimmed mean takes on the α-axis. Figure 5 combines the results of Figure 2 and Figure 4 with all the trimmed means restricted to the α-axis, and with the other estimators allowed to take any position in the space.

7. Interim Summary

For each robust L-estimator considered, a distribution model has been found for which the Bayesian mean approximates the estimator. A number of conclusions may be drawn, as follows.

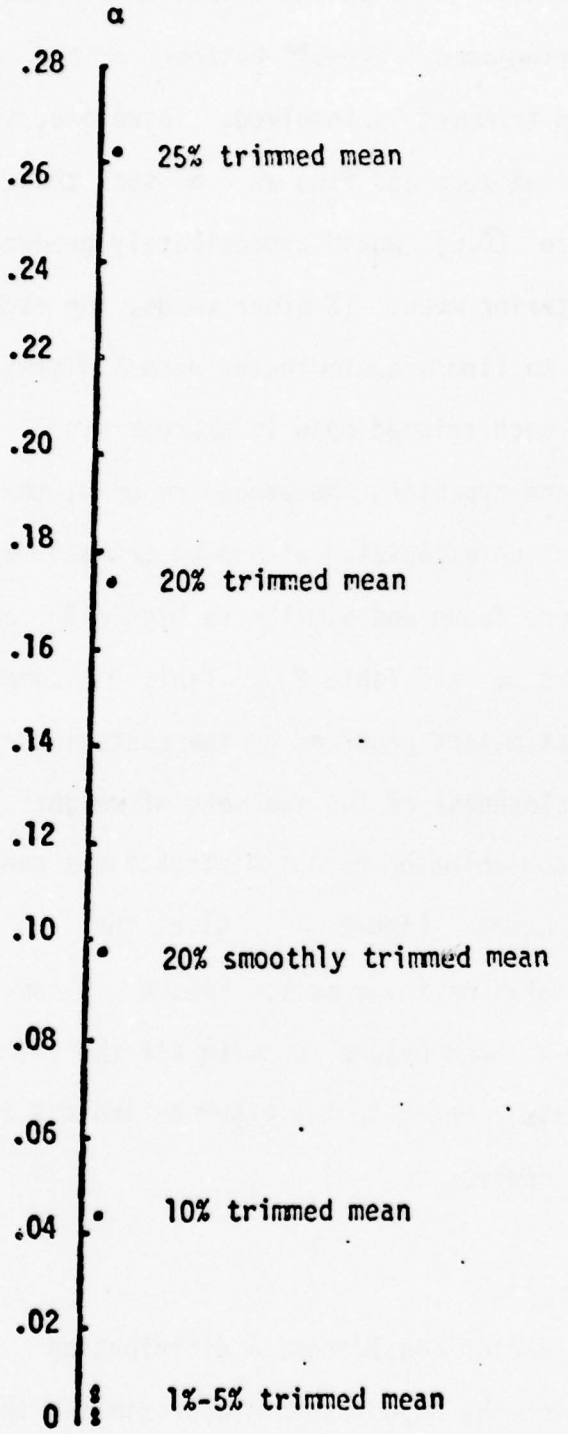


Figure 4 Values of α associated with various L-estimators, β is restricted to be zero.

L-Estimator	Weights for L-estimators	Value of α for OLUMV estimator	Weights of α-estimator
1) 1% trimmed mean	.091 .102 .102 .102 .102	α = .000893	.093 .101 .101 .101 .105
2) 2% trimmed mean	.083 .104 .104 .104 .104	α = .00179	.086 .102 .101 .101 .110
3) 2.5% trimmed mean	.078 .105 .105 .105 .105	α = .00524	.082 .102 .102 .102 .112
4) 5% trimmed mean	.056 .111 .111 .111 .111	α = .00584	.062 .105 .104 .104 .125
5) 10% trimmed mean	.000 .126 .126 .126 .126	α = .04380	-.002 .134 .119 .119 .128
6) 20% trimmed mean	.000 .167 .167 .167 .000	α = .17454	-.017 .147 .169 .177 .024
7) 25% trimmed mean	.000 .100 .200 .200 .000	α = .26307	-.014 .110 .210 .198 -.000
8) 20% smoothly trimmed mean	.000 .071 .143 .143 .143	α = .09871	-.015 .074 .145 .138 .157

Table 2 Left side: first five weights of the L-estimators
Right side: β restricted to be zero, first five weights of OLUMV estimator for the contaminated normal distribution (with parameter α) associated with the estimator on the left.

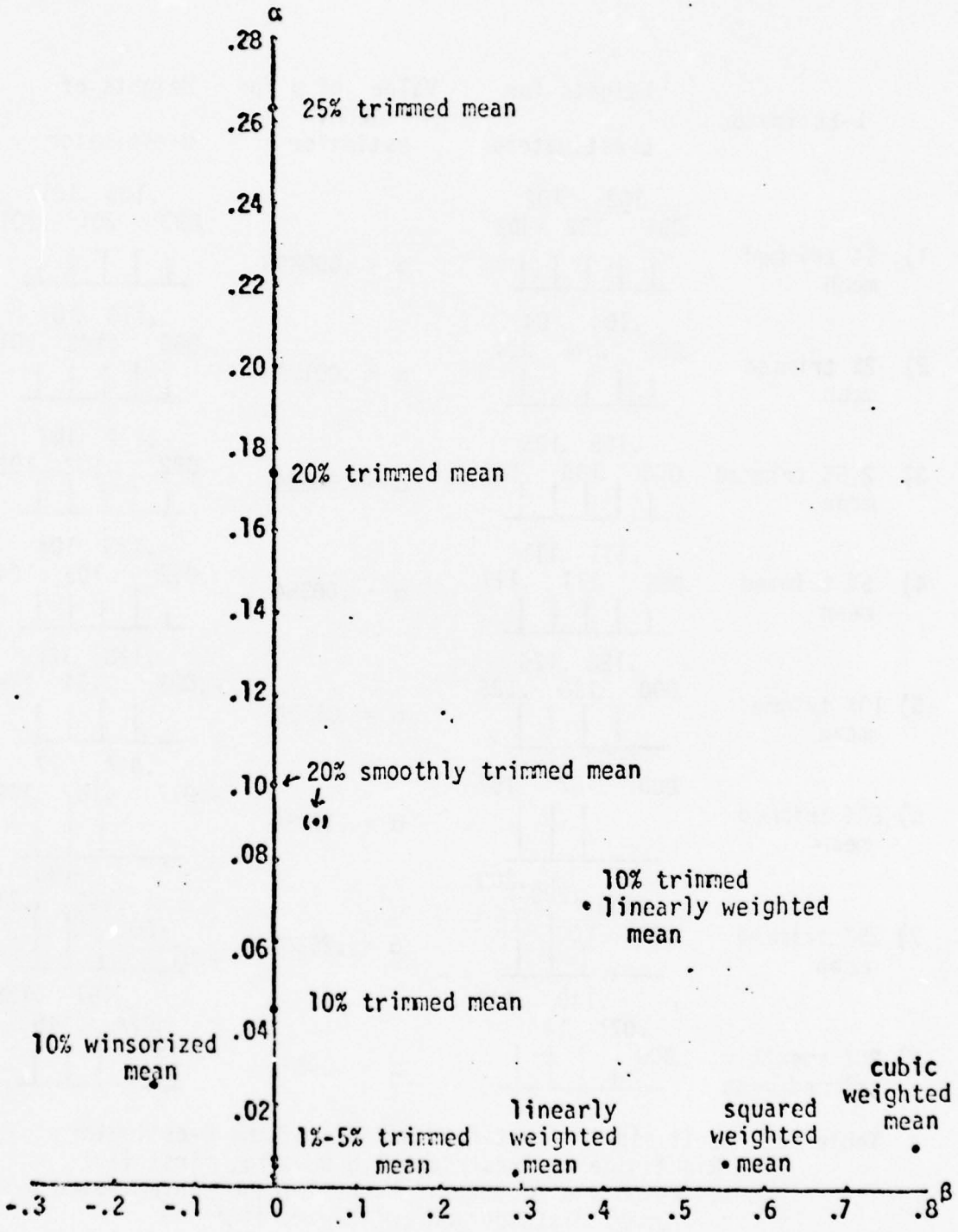


Figure 5 Combining results of Figure 2.2 and Figure 2.4.

1) The study has been effective in bringing into the open implied model assumptions inherent in the estimator. Such implied assumptions can thus be studied, criticized and, where appropriate applied to other problems. For example if trimmed means perform well for single samples, this implies the appropriateness of the contaminated normal distribution. This model may therefore with equal logic be used for more general models which traditionally employ the normal assumption.

2) It becomes clear that the Winsorized mean is appropriate for a contaminated distribution which is slightly light-tailed. Therefore unless one had confidence that such were the case, the Winsorized mean would not be appropriate since it puts heavy weight on the next to most extreme observations.

3) Linearly-weighted, squared-weighted or cubic weighted means imply that distributions met in practice are heavy-tailed with perhaps a small proportion of contamination.

4) Trimmed means, however, are appropriate if with probability $1-\alpha$ observations are from a fixed normal distribution and with probability α observations are from a normal distribution with the same mean but a larger variance. The trimming proportion depends on α.

8. Implied Assumptions of the M-estimator

Although some of the L-estimators have proved to be useful in the location estimation problem, their developers have found it hard to generalize them directly to cover, for example, regression models, general linear models and non-linear models. M-estimators are more flexible in this respect. To find the assumptions that would render appropriate each proposed M-estimator, the most natural way would be to try to find a distribution f which would make the M-estimator a maximum likelihood estimator. Let $\psi(x)$ be associated with some M-estimator; then the distribution $f(x)$ for which M will be maximum likelihood must be such that

$$\psi(x) = -\frac{d \log f(x)}{dx}$$

$$\log f(x) = -\int \psi(x)dx + c$$

$$f(x) \propto e^{-\int \psi(x)dx} .$$

If $\int_{-\infty}^{\infty} e^{-\int \psi(x)} dx < \infty$, we can properly standardize f and this would then be the precise assumption which would make the M-estimator a maximum likelihood estimator. As we will see, however, this integral condition is not satisfied by all proposed ψ functions.

The f's associated with different ψ's

(i) $\psi(x) = x \implies f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$

which is the normal distribution.

(ii) Huber's

$$\psi(x,k) = \begin{cases} -k & x < -k \\ x & -k < x < k \\ k & k < x \end{cases}$$

$$\implies f(x) = \begin{cases} \frac{c_1}{(2\pi)^{1/2}} e^{-\frac{x^2}{2}} & \text{for } |x| < k \\ \frac{c_1}{(2\pi)^{1/2}} e^{-k|x| + \frac{1}{2}k^2} & \text{for } |x| \geq k \end{cases}$$

c_1 is a constant such that $\int_{-\infty}^{\infty} f(x)dx = 1$.

(iii) Andrews'

$$\psi(x,c) = \begin{cases} \sin(x/c) & |x| \leq c\pi \\ 0 & \text{o.w.} \end{cases}$$

$$\implies f(x) = \begin{cases} ae^{+c}\cos(x/c) & |x| \leq c\pi \\ ae^{-c} & \text{otherwise} \end{cases}$$

a is some constant.

Note that $\int_{-\infty}^{\infty} f(x)dx = \infty$ so $f(x)$ is not a proper density.

(iv) Hampel's

$$\psi(x,a,b,c) = \operatorname{sgn} x \begin{cases} |x| & 0 \le |x| < a \\ a & a \le |x| < b \\ \frac{c-|x|}{c-b}\, a & b \le |x| < c \\ 0 & |x| \ge c \end{cases}$$

$$\Longrightarrow f(x) = \begin{cases} de^{-\frac{x^2}{2}} & 0 \le |x| < a \\ de^{-a|x| + \frac{1}{2}a^2} & a \le |x| < b \\ de^{-\frac{ac}{c-b}|x| + \frac{a}{2(c-b)}x^2 + \frac{ab}{2(c-b)} + \frac{1}{2}a^2} & b \le |x| < c \\ de^{-\frac{ac^2}{2(c-b)} + \frac{ab^2}{2(c-b)} + \frac{1}{2}a^2} & |x| \ge c \end{cases}$$

d is some constant.

Since $\int_{-\infty}^{\infty} f(x)dx = \infty$, $f(x)$ is not a proper density.

(v) Tukey's biweight

$$\psi(x,c) = \begin{cases} x[1-(\frac{x}{c})^2]^2 & |x| \le c \\ 0 & \text{o.w.} \end{cases}$$

$$\Longrightarrow f(x) = \begin{cases} de^{-\frac{x^2}{2} + \frac{1}{2c^2}x^4 - \frac{1}{6c^4}x^6} & |x| \le c \\ de^{-\frac{c^2}{6}} & \text{otherwise} \end{cases}$$

Since $\int_{-\infty}^{\infty} f(x)dx = \infty$, $f(x)$ is not a proper density.

A plot of these f's is given in Figure 6. For Andrews', Hampel's and Tukey's, the tails are constant which make them improper. But notice that their constant parts have very small values and one possible way to make the densities proper is to truncate the functions for very large deviations.

To allow comparison with our previous results, we attempt to obtain contaminated exponential power distributions which yield maximum likelihood estimators that are as close as possible to the M-estimators. One way to proceed is to look for a $\psi = -\frac{f'}{f}$ within the contaminated exponential power family which approximates the ψ function of the estimator of interest. To obtain better approximation, we shall in addition allow k to vary.

The density for the contaminated exponential power distribution is

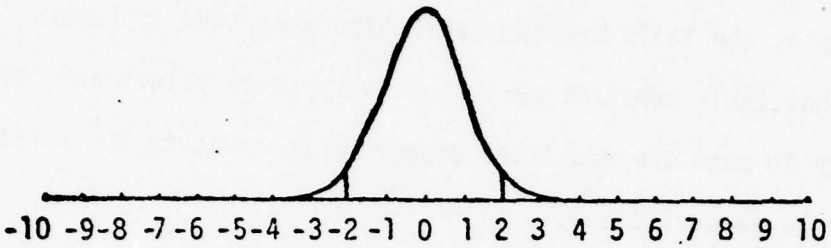
$$f(y) = (1-\alpha)P(y|\theta,\sigma,\beta) + \alpha P(y|\theta,k\sigma,\beta)$$

where $$P(y|\theta,\sigma,\beta) = w(\beta)\sigma^{-1}\exp\left[-c(\beta)\left|\frac{y-\theta}{\sigma}\right|^{2/(1+\beta)}\right] \quad -\infty < y < \infty$$

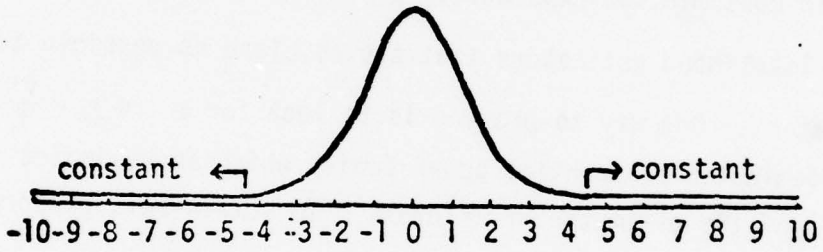
$$c(\beta) = \frac{\Gamma[\frac{3}{2}(1+\beta)]}{\Gamma[\frac{1}{2}(1+\beta)]}^{1/(1+\beta)} \quad \text{and} \quad w(\beta) = \frac{\{\Gamma[\frac{3}{2}(1+\beta)]\}^{1/2}}{(1+\beta)\{\Gamma[\frac{1}{2}(1+\beta)]\}^{3/2}} .$$

Without loss of generality, set $\theta = 0$, $\sigma = 1$

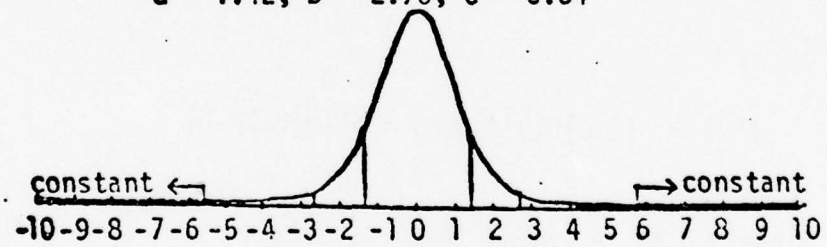
(i) Huber's k = 2.0



(ii) Andrews' c = 1.42



(iii) Hampel's
a = 1.42, b = 2.70, c = 5.54



(iv) Tukey's c = 5.4

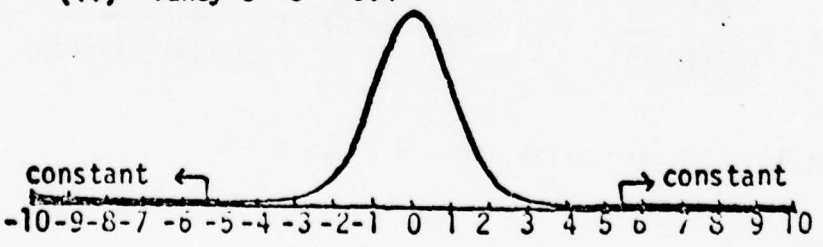


Figure 6 Function f corresponds to each M-estimator.

$$f(x) = (1-\alpha)w(\beta)\exp\left[-c(\beta)|x|^{\frac{2}{1+\beta}}\right] + \alpha w(\beta)/k \exp\left[-c(\beta)\left|\frac{x}{k}\right|^{\frac{2}{1+\beta}}\right]$$

$$-f'(x) = (1-\alpha)w(\beta)c(\beta)\exp\left(-c(\beta)|x|^{\frac{2}{1+\beta}}\right)\left(\frac{2}{1+\beta}|x|^{\frac{1-\beta}{1+\beta}}\right)$$

$$+\alpha w(\beta)/k \cdot c(\beta)/k \exp\left[-c(\beta)\left|\frac{x}{k}\right|^{\frac{2}{1+\beta}}\right]\left(\frac{2}{1+\beta}\left|\frac{x}{k}\right|^{\frac{2}{1+\beta}}\right) \quad \text{for} \quad x > 0$$

$$\psi(x) =$$

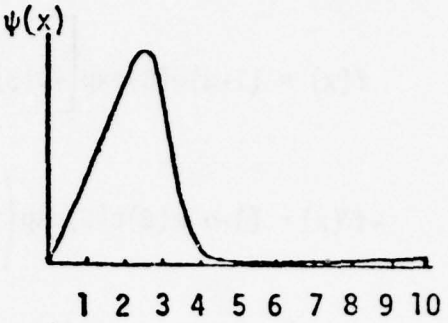
$$\frac{(1-\alpha)c(\beta)\left(\frac{2}{1+\beta}|x|^{\frac{1-\beta}{1+\beta}}\right)\exp\left(-c(\beta)|x|^{\frac{2}{1+\beta}}\right) + \frac{\alpha c(\beta)}{k^2}\left(\frac{2}{1+\beta}\left|\frac{x}{k}\right|^{\frac{2}{1+\beta}}\right)\exp\left[-c(\beta)\left|\frac{x}{k}\right|^{\frac{2}{1+\beta}}\right]}{(1-\alpha)\exp\left[-c(\beta)|x|^{\frac{2}{1+\beta}}\right] + \alpha/k \exp\left[-c(\beta)\left|\frac{x}{k}\right|^{\frac{2}{1+\beta}}\right]}$$

$$x > 0$$

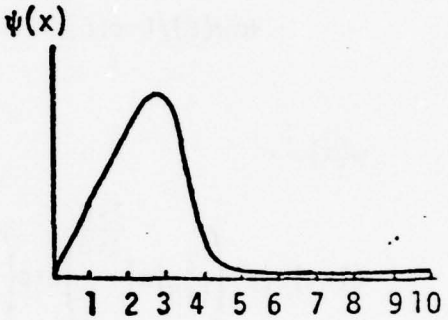
and the ψ function is antisymmetric about $x = 0$.

This ψ function (only the positive half, the other half is obtained by antisymmetry) is plotted in Figure 7 for different combinations of β, α and k. Observe that as β becomes larger, the curve increases faster in the neighborhood of zero and then slows down to reach its maximum, as α increases, the curve drops sharper after reaching its maximum and as k increases the ψ function gets closer to zero and stays smaller for a longer period. Except for $\beta = 1$, all the ψ functions for contaminated exponential power distributions eventually increase to infinity.

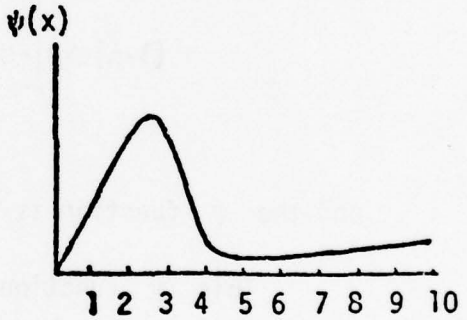
$\beta = -.1$
$\alpha = .05$
$k = 10.0$



$\beta = 0$
$\alpha = .03$
$k = 10.0$



$\beta = 0$
$\alpha = .05$
$k = 5.0$



$\beta = 0$
$\alpha = .05$
$k = 10.0$

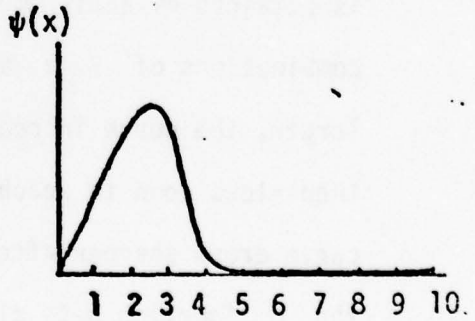
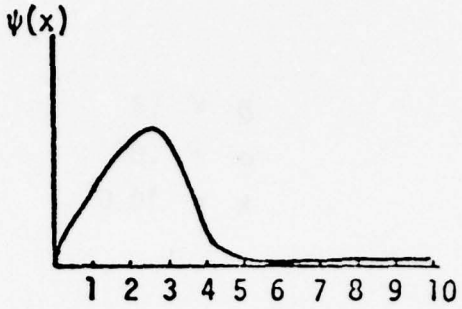
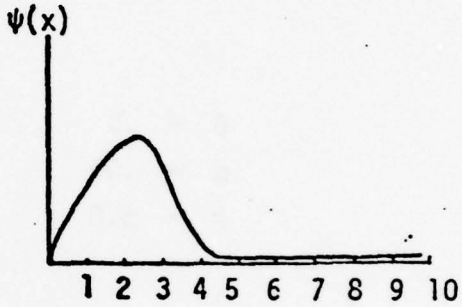


Figure 7 (a) ψ functions for contaminated exponential power distributions with different α, β and k values.

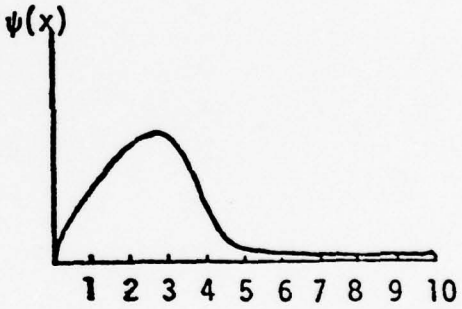
$\beta = .1$
$\alpha = .05$
$k = 10.0$



$\beta = .1$
$\alpha = .1$
$k = 10.0$



$\beta = .2$
$\alpha = .03$
$k = 10.0$



$\beta = .2$
$\alpha = .05$
$k = 8.0$

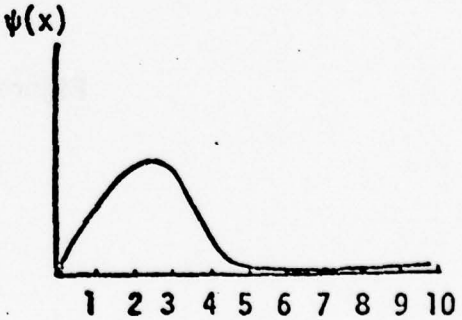
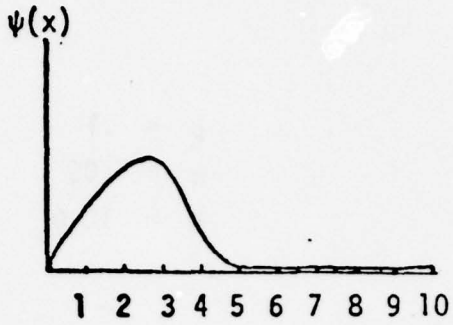


Figure 7 (a) continued

$\beta = .2$
$\alpha = .05$
$k = 10.0$



$\beta = .5$
$\alpha = .05$
$k = 5.0$

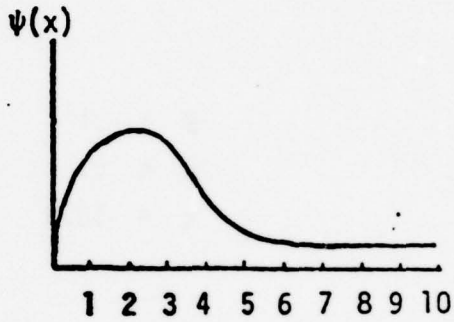
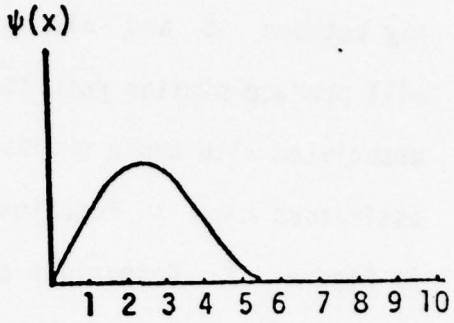
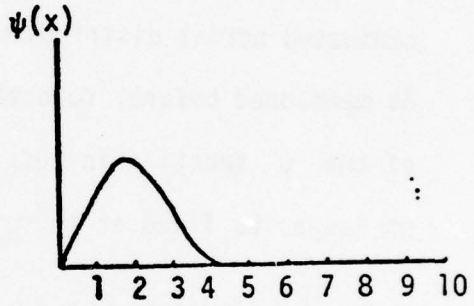


Figure 7 (a) continued

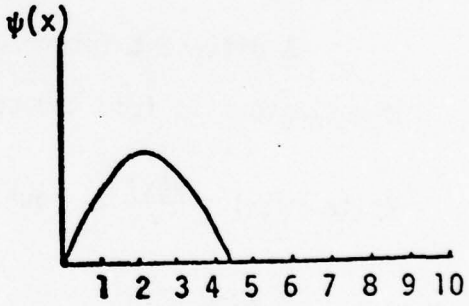
Tukey's c = 5.40



Tukey's c = 4.05



Andrews' c = 1.42



Cauchy

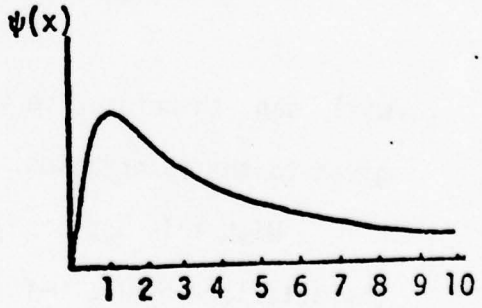


Figure 7 (b) ψ functions for various M-estimators.

Even so the ordinate can remain small within a very wide range, say between ± 5 and ± 10 and, therefore, for realistic data sets will produce similar results as those ψ functions which are associated with those proposed M-estimators. Tukey's and Andrews' estimators have ψ functions with shape somewhat similar to those in Figure 7. Comparison shows that Tukey's estimator with $c = 4.05$ is similar to the maximum likelihood estimator for a contaminated normal distribution with $\beta = .1$, $\alpha = .1$ and $k = 10.0$. As mentioned before, to obtain a better approximation in the tail of the ψ function in this study of M-estimators, the value of k will no longer be fixed at three as in the study of L-estimators.

An Alternative Approach

A different interpretation that can be attached to the M-estimators is from the point of view of the weighting function. Write $w(x) = \frac{\psi(x)}{x}$, equation $\Sigma\psi(\frac{x_i-\theta}{\sigma}) = 0$ is equivalent to

$$\Sigma(\frac{x_i-\theta}{\sigma})w(\frac{x_i-\theta}{\sigma}) = 0 \quad \text{and} \quad \theta = \frac{\Sigma x_i w(\frac{x_i-\theta}{\sigma})}{\Sigma w(\frac{x_i-\theta}{\sigma})}$$ (Beaton & Tukey 1974)

$w(x)$ can, therefore, be viewed as the weight of the estimator given to the observation.

With this approach, each M-estimator is then being characterized by the w-function. The following are some weighting

functions for the M-estimators, the plots of which are shown in Figure 9. In terms of weight functions, $w(x) = \psi(x)/x$

(i) normal $\quad w(x) = 1$

(ii) Huber's

$$w(x,k) = \begin{cases} -\frac{k}{x} & x < -k \\ 1 & -k < x < k \\ \frac{k}{x} & k < x \end{cases}$$

(iii) Andrews'

$$w(x,c) = \begin{cases} \sin(x/c)/x & |x| \leq c\pi \\ 0 & \text{o.w.} \end{cases}$$

(iv) Hampel's

$$w(x,a,b,c) = \text{sgn } x \begin{cases} |x|/x & 0 \leq |x| < a \\ a/x & a \leq |x| < b \\ \frac{c-|x|}{c-b}\, a/x & b \leq |x| < c \\ 0 & |x| \geq c \end{cases}$$

(v) Tukey's biweight

$$w(x,c) = \begin{cases} (1-(\frac{x}{c})^2)^2 & |x| \leq c \\ 0 & \text{o.w.} \end{cases}$$

Corresponding to each contaminated exponential power distribution, there is also a w-function $w(x) = \frac{\psi(x)}{x}$, and this again can be used in picking out the distribution that would lead to the robust

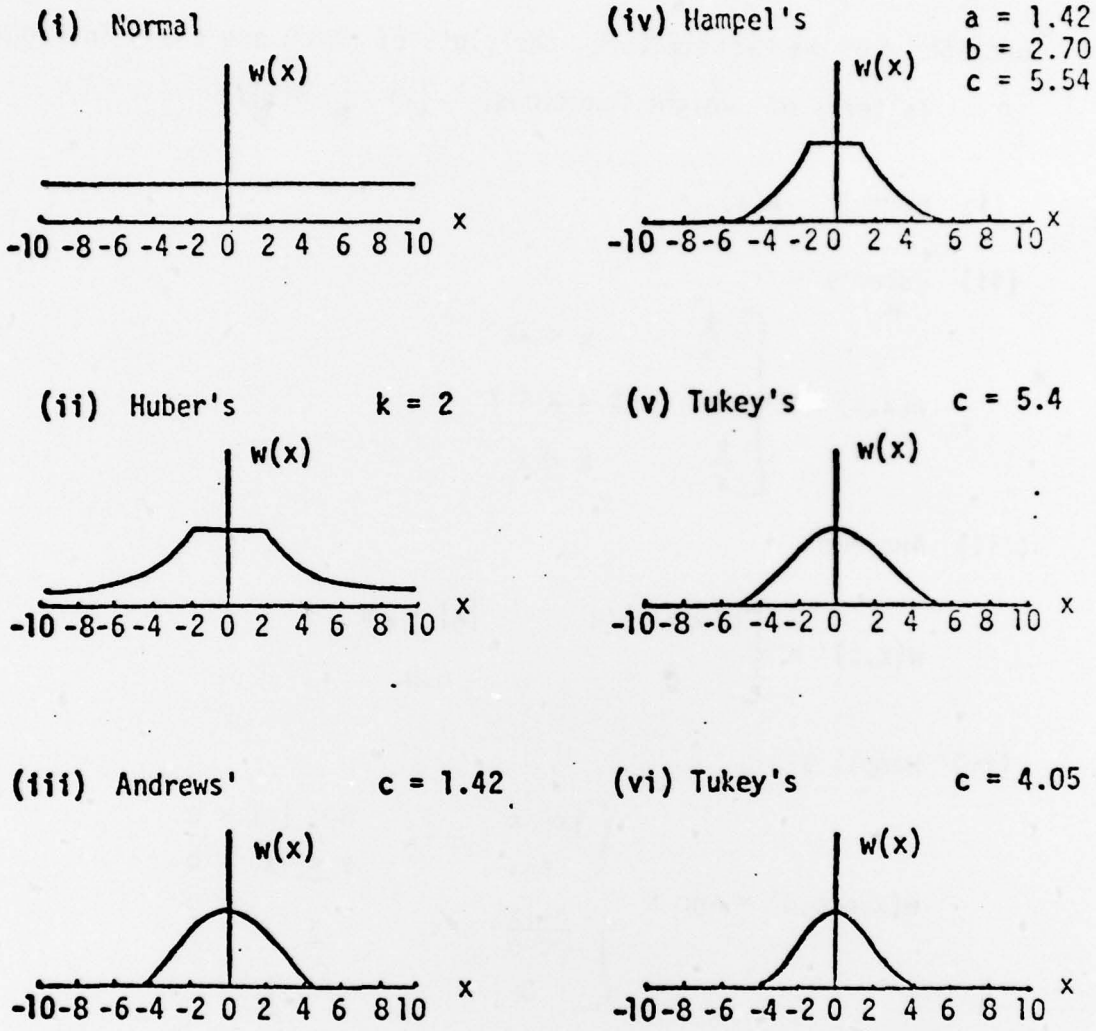


Figure 9 w-functions for various M-estimators.

M-estimator proposed. The w-functions are plotted (Figure 10) for contaminated exponential power distribtuions with α = .025, .05, .075, .1 and β = 0, .2, .4, and also k = 1.0, 2.0, 4.0, 6.0, 8.0, 10.0. From the formula for the w-function, it is easy to see that for $\beta > 0$, w(x) is always infinity at x = 0. Although this can cause trouble when actually solving for the estimate for some data set, comparisons are still possible with the weighting function of the M-estimators.

Comparing Figure 9 with Figure 10, we can make the following matchings.

1. Huber's estimator - $\beta = 0$ $\alpha = .05$ k = 2.0
2. Hampel's estimator - $\beta = 0$ $\alpha = .10$ k = 8.0
   or maybe $\beta = 0$ $\alpha = .05$ k = 4.0
3. Andrews' estimator - $\beta = 0$ $\alpha = .10$ k = 4.0
4. Tukey's estimator (c = 4.05) - $\beta = .2$ $\alpha = .10$ k = 10.0
   or maybe $\beta = 0$ $\alpha = .10$ k = 4.0
5. Tukey's estimator (c = 5.40) - $\beta = .2$ $\alpha = .025$ k = 10.0

The correspondence is far from exact, the w-functions for these estimators simply cannot be reproduced very closely by our family of distributions. However, we do know that estimators of this kind are not very sensitive to minor discrepancies in the shape of the functions, in that similarly shaped weight functions will produce very close estimates.

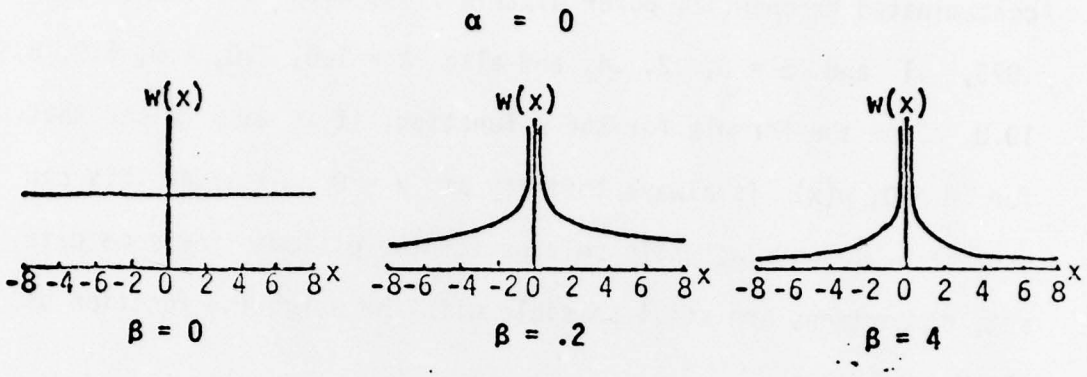


Figure 10 w functions for contaminated exponential power distributions with different α, β and k values.

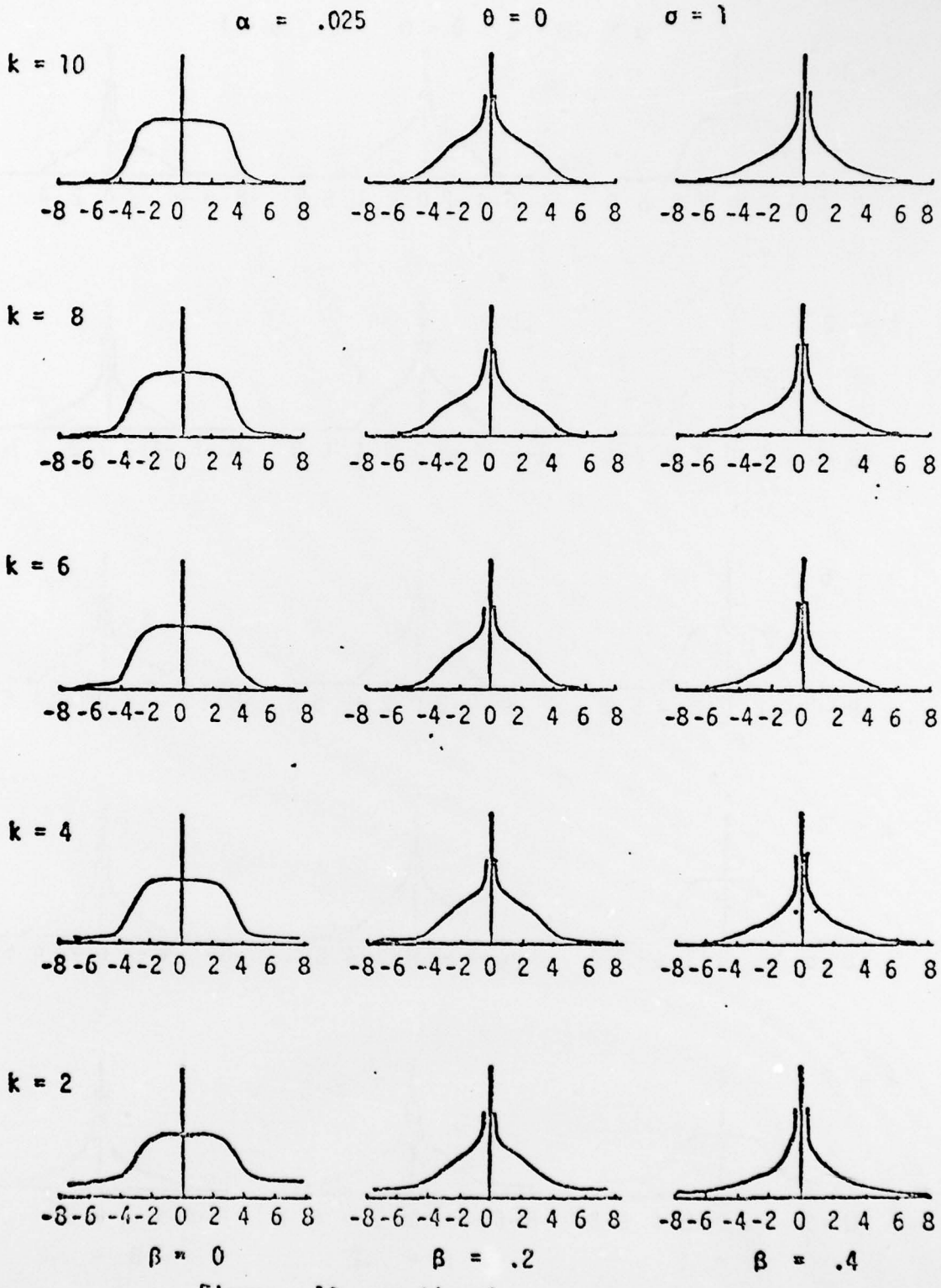


Figure 10 continued

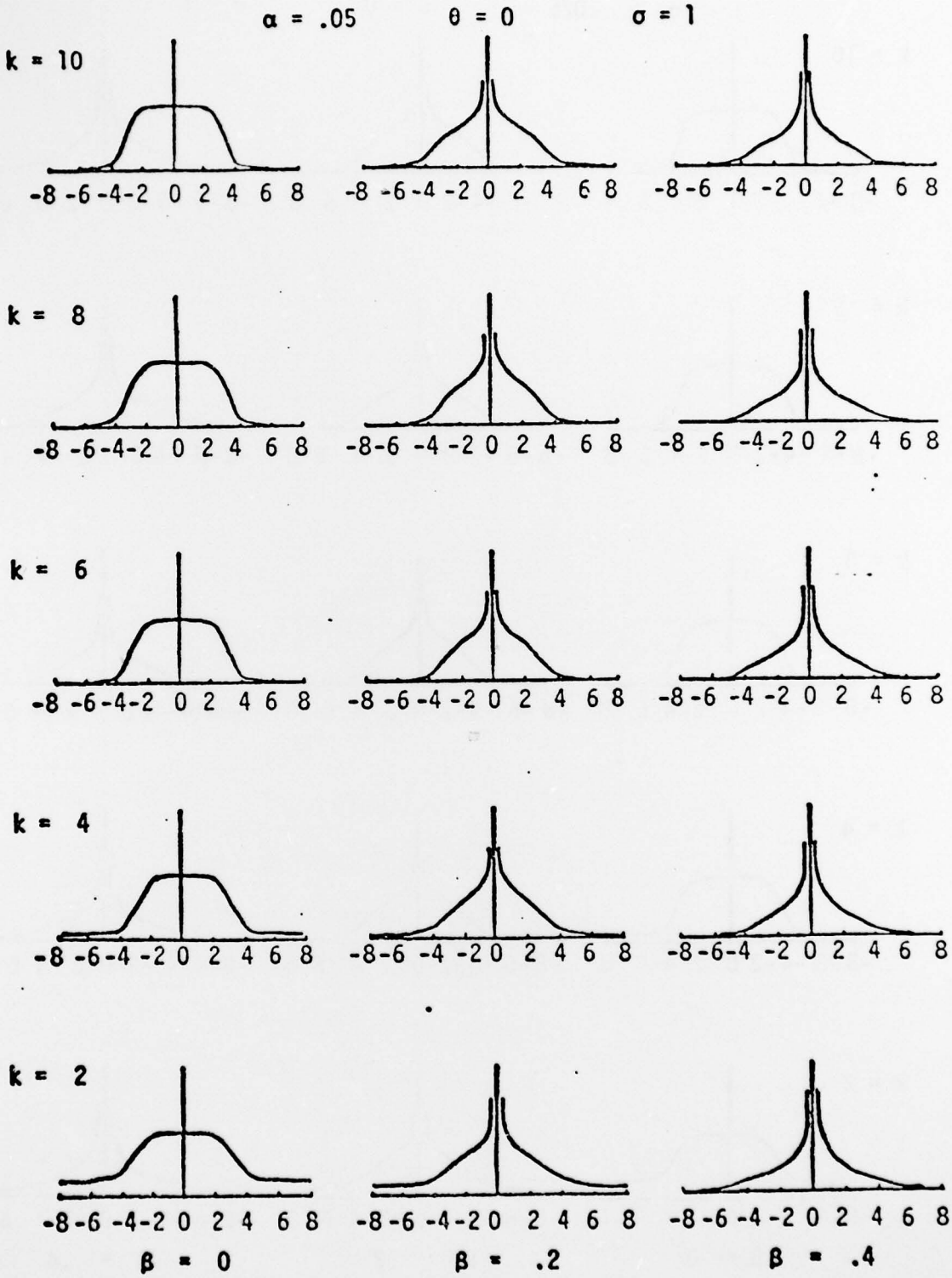


Figure 10 continued

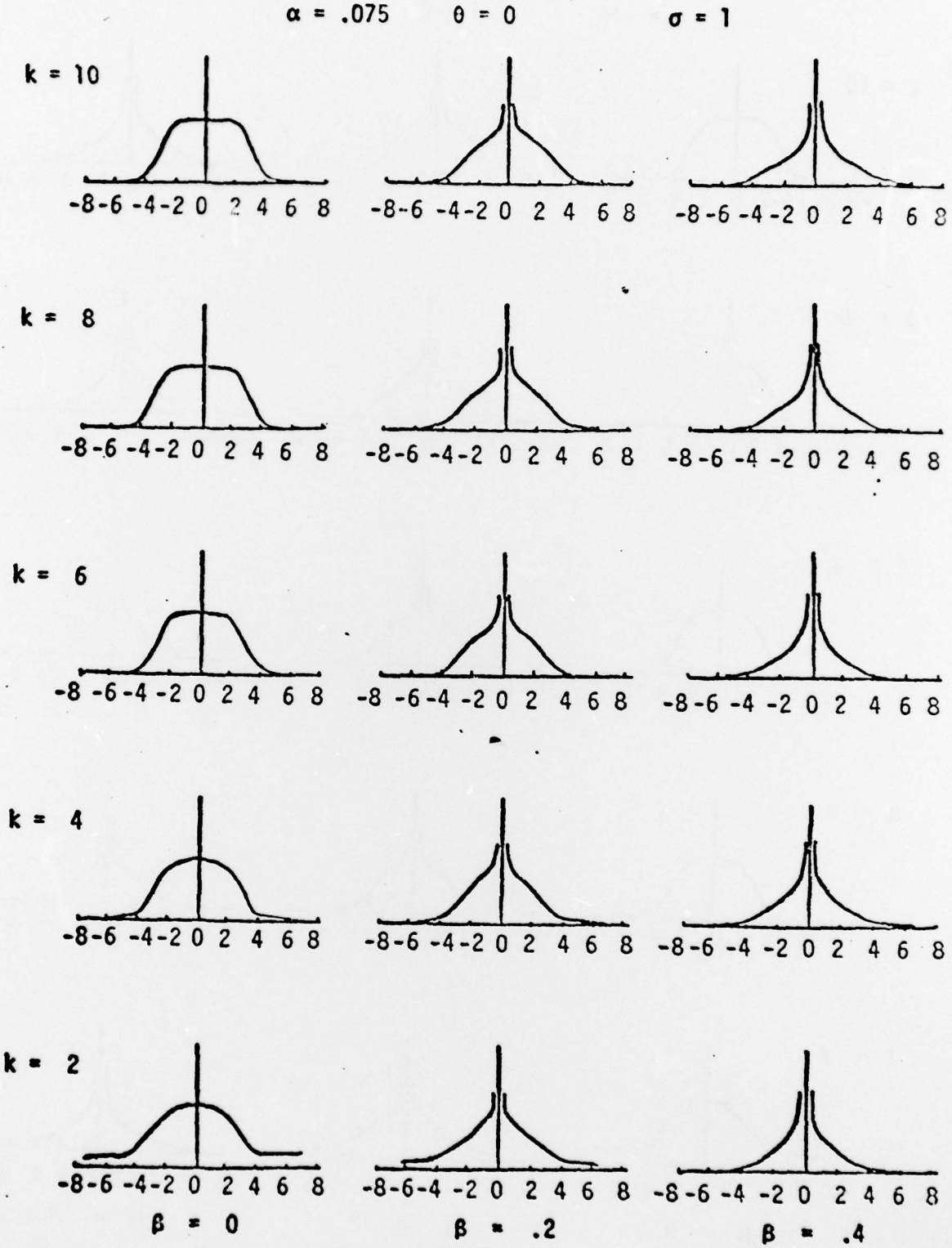


Figure 10 continued

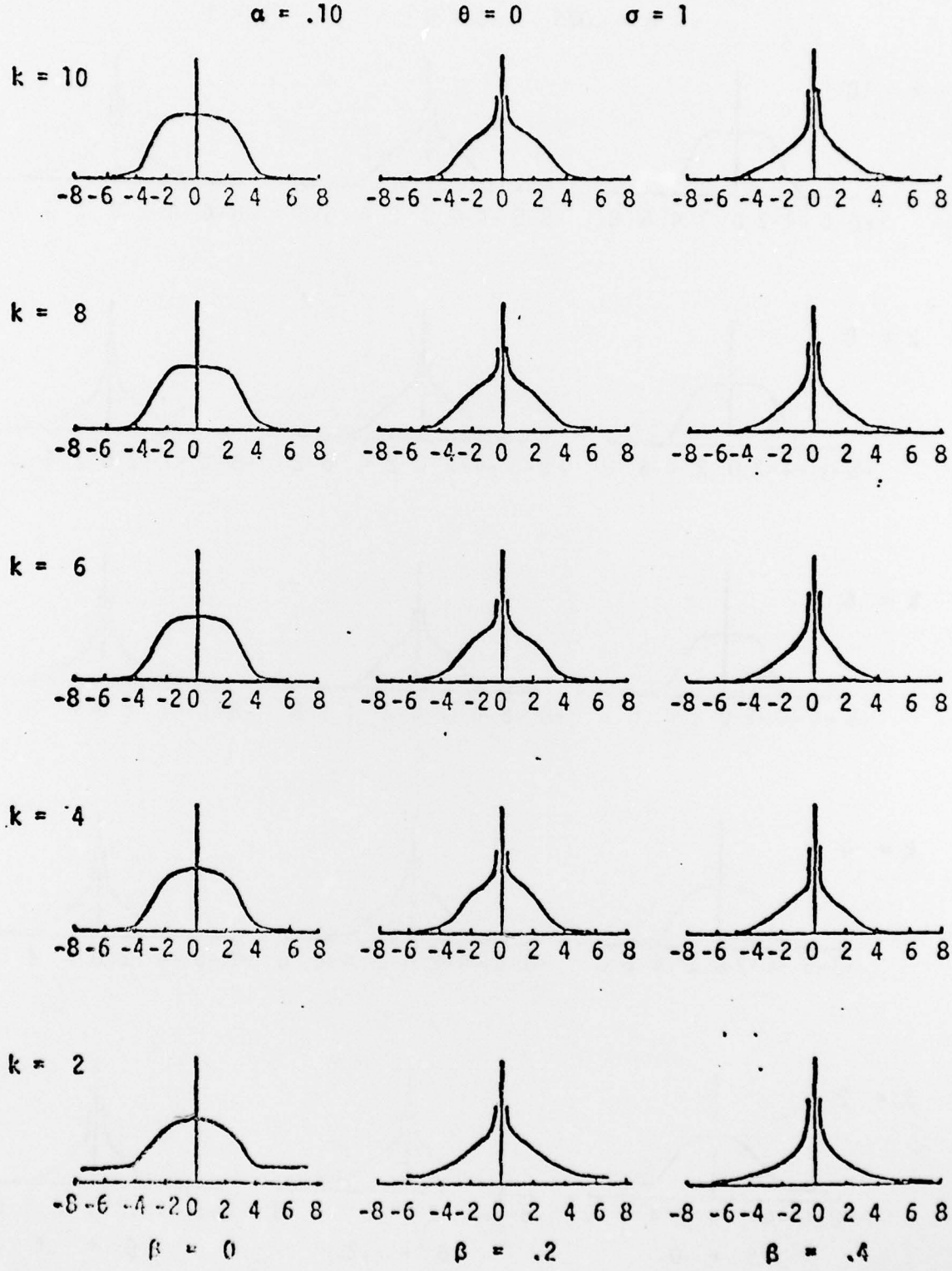


Figure 10 continued

The comparison does provide an approximate idea of the nature of these estimators in terms of the β-α space. The β-coordinates are close to zero strongly suggesting that estimators with proportions similar to the M-estimators studied above will be produced by the contaminated normal model.

9. Summary of the present chapter

Models for the (β,α) family of distributions have been found which generated estimators similar to published L-estimators and M-estimators. The contaminated normal model can produce estimates similar to trimmed means and to the M-estimates of Huber, Tukey, Hampel and Andrews. Among the L-estimators, linearly-weighted, squared-weighted and cubic-weighted means implied the need for the assumption that the parent distribution had heavy tails in addition to some contamination. The Winsorized mean implies that the parent distribution was somewhat light-tailed with contamination.

Appendix A

Proof that "the limiting weights obtained in Section 2.4 as $m \to \infty$ are the same as weights for OLUMV estimators."

Let us assume a random sample of size n is taken from a distribution $P(Y|\theta)$ and the ordered observations are $\underset{\sim}{Y} = (y_1, \ldots, y_n)$. Also assume $P(Y|\theta)$ is symmetric about θ, so that a linear combination of order statistics will be unbiased if and only if the weights are symmetric about its center.

Let $\mathcal{A}$ be the set of all symmetric weights, $\mathcal{Y}$ be the sample space, $\alpha_1 y_1 + \ldots + \alpha_n y_n$ is the OLUMV estimator, then we have

$$\int_{\mathcal{Y}} (\alpha_1 y_1 + \ldots + \alpha_n y_n - \theta)^2 P(\underset{\sim}{Y}|\theta) d\underset{\sim}{Y}$$

$$= \min_{(A_1, \ldots, A_n) \in \mathcal{A}} \int_{\mathcal{Y}} (A_1 y_1 + \ldots + A_n y_n - \theta)^2 P(\underset{\sim}{Y}|\theta) d\underset{\sim}{Y} \qquad \text{(A. 1)}$$

What we have done in Section 2.4 is to take m sets of random samples $(y_{i1}, \ldots, y_{in})$ $i = 1, \ldots, m$ and find weights that minimize $\sum_{i=1}^{m} (A_1 y_{i1} + \ldots + A_n y_{in} - M_{\underset{\sim}{Y}_i})^2$ as $m \to \infty$ or, equivalently, minimize

$$\frac{1}{m} \sum_{i=1}^{m} (A_1 y_{i1} + \ldots + A_n y_{in} - M_{\underset{\sim}{Y}_i})^2 \quad \text{as} \quad m \to \infty.$$

This is the same as finding weights that minimize

$$\int_{\mathcal{Y}} (A_1y_1+\ldots+A_1y_n-M_{\underset{\sim}{Y}})^2 P(\underset{\sim}{Y}|\theta)d\underset{\sim}{Y}$$

$$= \int_{\mathcal{Y}} (A_1y_1+\ldots+A_ny_n-\theta+\theta-M_{\underset{\sim}{Y}})^2 P(\underset{\sim}{Y}|\theta)d\underset{\sim}{Y}$$

$$= \int_{\mathcal{Y}} (A_1y_1+\ldots+A_ny_n-\theta)^2 P(\underset{\sim}{Y}|\theta)d\underset{\sim}{Y} + \int_{\mathcal{Y}} (\theta-M_{\underset{\sim}{Y}})^2 P(\underset{\sim}{Y}|\theta)d\underset{\sim}{Y}$$

$$+ 2\int_{\mathcal{Y}} (A_1y_1+\ldots+A_ny_n-\theta)(\theta-M_{\underset{\sim}{Y}})P(\underset{\sim}{Y}|\theta)d\underset{\sim}{Y} \quad . \qquad \text{(A.2.)}$$

Let $T = A_1y_1+\ldots+A_ny_n$; then the second and third terms of (A.2) become

$$\int_{\mathcal{Y}} (\theta-M_{\underset{\sim}{Y}})^2 P(\underset{\sim}{Y}|\theta)d\underset{\sim}{Y} + 2\int_{\mathcal{Y}} (T-\theta)(\theta-M_{\underset{\sim}{Y}})P(\underset{\sim}{Y}|\theta)d\underset{\sim}{Y}$$

$$= \int_{\mathcal{Y}} (\theta^2-2\theta M_{\underset{\sim}{Y}}+M^2_{\underset{\sim}{Y}}+2T\theta-2\theta^2-2TM_{\underset{\sim}{Y}}+2\theta M_{\underset{\sim}{Y}})P(\underset{\sim}{Y}|\theta)d\underset{\sim}{Y}$$

$$= \int_{\mathcal{Y}} (M^2_{\underset{\sim}{Y}}-\theta^2+2T\theta-2TM_{\underset{\sim}{Y}})P(\underset{\sim}{Y}|\theta)d\underset{\sim}{Y} \quad .$$

Since $M^2_{\underset{\sim}{Y}}-\theta^2$ does not depend on $A_1,\ldots,A_n$, we want to find weights that minimize

$$\int_{\mathcal{Y}} (A_1y_1+\ldots+A_ny_n-\theta)^2 P(\underset{\sim}{Y}|\theta)d\underset{\sim}{Y} + 2\int_{\mathcal{Y}} T\cdot(\theta-M_{\underset{\sim}{Y}})P(\underset{\sim}{Y}|\theta)d\underset{\sim}{Y} \quad .$$

If we now put a prior on θ, $P(\theta) = \frac{1}{2c}$ - $c < \theta < c$, we are looking for weights that minimize

$$\int_{-c}^{c}\int_{\mathcal{Y}} (A_1y_1+\ldots+A_ny_n-\theta)^2 P(Y|\theta)\frac{1}{2c}\,d\theta$$

$$+ \int_{-c}^{c}\int_{\mathcal{Y}} T(\theta-M_{\underset{\sim}{Y}})P(\underset{\sim}{Y}|\theta)\frac{1}{2c}\,d\underset{\sim}{Y}\,d\theta \quad . \qquad \text{(A. 3)}$$

Now if c is large enough, after proper standardization, $P(\underset{\sim}{Y}|\theta) \cdot \frac{1}{2c}$ will be very close to the posterior distribution with noninformative prior $P(\theta) \propto$ constant. Changing the order of integration in the second term of (A. 3), we have

$$\int_{-c}^{c} (\theta - M_{\underset{\sim}{Y}}) P(\underset{\sim}{Y}|\theta) \frac{1}{2c} \, d\theta \to 0 \qquad \text{as} \quad c \to \infty$$

so

$$\int_{\mathcal{Y}} \int_{-c}^{c} T(\theta - M_{\underset{\sim}{Y}}) P(\underset{\sim}{Y}|\theta) \frac{1}{2c} \, d\theta \, d\underset{\sim}{Y} \to 0 \qquad \text{as} \quad c \to \infty.$$

Thus the second term of (A.2.3) drops out as $c \to \infty$.

The limiting weights will minimize

$$\int_{-c}^{c} \int_{\mathcal{Y}} (A_1 y_1 + \ldots + A_n y_n - \theta)^2 P(\underset{\sim}{Y}|\theta) \frac{1}{2c} \, d\underset{\sim}{Y} \, d\theta$$

when c is large enough. Since $\int_{\mathcal{Y}} (A_1 y_1 + \ldots + A_n y_n - \theta)^2 P(\underset{\sim}{Y}|\theta) d\underset{\sim}{Y}$ is independent of θ, this is the same as minimizing $\int_{\mathcal{Y}} (A_1 y_1 + \ldots + A_n y_n - \theta)^2 P(\underset{\sim}{Y}|\theta) d\underset{\sim}{Y}$ which leads to the same result as (A. 1). Therefore $\alpha_1, \ldots, \alpha_n$ are our limiting weights.

References

1. Andrews, D. F., Bickel, P. J., Hampel, F. R., Huber, P. J., Rogers, W. H., and Tukey, J. W. (1972), Robust Estimates of Location: Survey and Advances. Princeton University Press.

2. Beaton, A. E. and Tukey, J. W. (1974), "The fitting of power series, meaning polynomials, illustrated on band spectroscopic data," Technometrics, 16, p 147-185.

3. Box, G. E. P., and Tiao, G. C. (1968), "A Bayesian approach to some outlier problems," Biometrika, 55, p 119-129.

4. Box, G. E. P. and Tiao, G. C. (1973), Bayesian Inference in Statistical Analysis. Addison-Wesley.

5. Crow, Edwin L. (1964), "The statistical construction of a single standard from several available standards," IEEE Transactions on Instrumentation and Measurement, I-13, p 180-185.

6. Crow, E. L., and Siddiqui, M. M. (1967), "Robust estimates of location," JASA, 62, p 353-384.

7. Gastwirth, J. L. and Cohen, M. L. (1970), "Small sample behavior of some robust linear estimators of location," JASA, 65, p 946-973.

8. Gauss, C. F. (1821), "Göttingische gelehrte Anzeigen," p 321-327. (Reprinted in Werke Bd. r, p 98).

9. Hogg, R. V. (1974), "Adaptive robust procedures: a partial review and some suggestions for future applications and theory," JASA, 69, p 909-922.

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

11. Huber, P. J. (1972), "Robust statistics: a review," Ann. Math. Statist., 43, p 1041-1067.

12. Lund, D. R. (1967), Parameter Estimation in a Class of Power Distributions. Ph.D. thesis, the University of Wisconsin-Madison.

13. Stigler, S. M. (1973), "The asymptotic distribution of the trimmed mean," Ann. Statist., 1, p 472-477.

14. Tukey, J. W. (1962), "The future of data analysis," Ann. Math. Statist., 33, p 1-67.

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

(14) MRC-TSR-1997

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: 1997

2. GOVT ACCESSION NO.

3. RECIPIENT'S CATALOG NUMBER

(9) Technical

(6) 4. TITLE (and Subtitle): IMPLIED ASSUMPTIONS FOR SOME PROPOSED ROBUST ESTIMATORS.

5. TYPE OF REPORT & PERIOD COVERED: Summary Report, no specific reporting period

6. PERFORMING ORG. REPORT NUMBER

(10) 7. AUTHOR(s): Gina Chen and George E. P. Box

(15) 8. CONTRACT OR GRANT NUMBER(s): DAAG29-75-C-0024

9. PERFORMING ORGANIZATION NAME AND ADDRESS: Mathematics Research Center, University of Wisconsin, 610 Walnut Street, Madison, Wisconsin 53706

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: #4 Probability, Statistics, and Combinatorics

11. CONTROLLING OFFICE NAME AND ADDRESS: U. S. Army Research Office, P.O. Box 12211, Research Triangle Park, North Carolina 27709

12. REPORT DATE: September 1979

13. NUMBER OF PAGES: 60

(12) 64

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)

robust estimators, L-estimators, M-estimators, contaminated normal model, exponential power distribution, Bayesian approach

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

Assumptions which could motivate various L-estimators and M-estimators are discussed. In particular, for samples of size ten a distribution in the contaminated exponential power family is found whose posterior mean approximates each of a number of proposed L-estimators. Also distributions in this family are found whose posterior modes approximate suggested M-estimators.

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

221 200

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)