This is to certify that the
dissertation entitled

Trimmed and Winsorized Estimators

presented by

Mingxin Wu

has been accepted towards fulfillment
of the requirements for the

PhD degree in Statistics

Yijun Zuo
Major Professor's Signature

Date 05/04/06

TRIMMED AND WINSORIZED ESTIMATORS

By

Mingxin Wu

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Probability and Statistics Department,

2006

ABSTRACT

TRIMMED AND WINSORIZED ESTIMATORS

By

Mingxin Wu

The dissertation consists of three parts. The first part studies trimmed and winsorized means based on a scaled deviation. The influence functions of the trimmed (and winsorized) means are derived and their limiting distributions are established via asymptotic representations. The performance of these estimators with respect to various robustness and efficiency criteria is evaluated and compared with leading competitors including the ordinary Tukey trimmed (and winsorized) means. The resulting trimmed (and winsorized) means are much more robust than their predecessors. Indeed they can share the best breakdown point robustness of the sample median for any common trimming thresholds. Furthermore, for appropriate trimming thresholds they are highly efficient for light-tailed symmetric models and more efficient than their predecessors for heavy-tailed or contaminated symmetric models.

The second part of the dissertation pertains to applying the same trimming scheme to the scale setting. In this part, trimmed (and winsorized) standard deviations based on a scaled deviation are introduced and studied. The influence functions and the limiting distributions are obtained. The performance of the estimators is evaluated and compared with respect to high breakdown scale estimators. Unlike other high breakdown competitors which perform poorly for light-tailed distributions and for contaminated symmetric distributions with contamination near the center, the resulting trimmed (and winsorized) standard deviations are much more efficient than their predecessors for light-tailed distributions for suitably chosen trimming parameters and highly efficient for heavy-tailed and skewed distributions. At the

same time, they are sharing the best breakdown point robustness of the sample median absolute deviation for any common trimming thresholds.

The third part is about the multiple least square estimator. In this part we introduce the least trimmed squares estimator for multiple regression. A fast algorithm for its computation is proposed. We prove Fisher consistency for the multiple regression model with symmetric error distributions and derive the influence function. Simulation studies investigate the finite-sample efficiency of the estimator.

ACKNOWLEDGMENTS

Foremost, I am deeply grateful to my advisor, Professor Yijun Zuo for introducing me to this wonderful place–MSU. He is gratefully acknowledged for his years of encouragement, his scientific influence on me, his infinite patience, his insights in our numerous discussions, his financial support and his careful review of paper manuscripts. Without these generous assistance, this dissertation could not have come into light. To be one of his students is my great honor.

I also wish to express my gratitude to my dissertation committee, Professor Connie Page, Professor Habib Salehi, Professor Lijian Yang, for sparing their precious time to serve on my committee and giving valuable comments and suggestions.

I am grateful to Professor Connie Page and Professor Dennis Gilliland for accepting me as one of the consultants at Cstat. It's such a good experience to work with Cstat that I have had chance to encounter some very interesting topics such as survival analysis, machine learning and application aspects of statistics.

My thanks also go to Professor James Stapleton for his numerous help and constant support.

The support and encouragement of members of statistics department are greatly appreciated and acknowledged.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

Introduction and Motivation

1.1 Location

The sample mean is the most efficient location estimator for normal models. It is, however, not robust. The sample median is the most robust location estimator with the best breakdown point. It is, however, not efficient for normal models. The trimmed mean is a compromise between the two extremes. It is more robust than the mean and more efficient than the median for normal models. It also performs quite well for heavy-tailed non-normal symmetric distributions. That's one reason why we use it in Olympic rating system. We know that for some sports in the Olympics, such as diving, gymnastics (Summer Olympics) and several years ago, figure skating (Winter Olympics), the rating system is based on the trimmed means. In Olympic rating, nine different judges from nine different countries give nine scores for each athlete. They drop the highest and lowest scores, taking the average of the rest seven scores as the final score of each athlete. The Gold Medal in these games is awarded to contestants with the highest trimmed scores. And many rating systems for competitions in our life are also based on trimmed means. The high and low scores are dropped and the rest are averaged. In every Olympics year, there are a lot of controversies over the ordinary trimmed mean based scoring system used in competitions. In fact there are a lot of problems with it. The argument I make in this dissertation is that the problems can be avoided by trimmed mean/winsorized mean based on a scaled-deviation—indeed, from a statistical point of view it is statistically superior to the alternatives that have been

proposed and used.

The ordinary trimmed mean associated with the Olympic rating system has the following shortcomings which can be avoided by the trimmed mean based on a scaled-deviation

1 Ordinary trimmed mean (OTM) cannot to exclude all the outliers when outliers come from one side instead of both sides. This time outliers are from lower side, it will underestimate the athlete (center); next time when the outliers are from upper side, it will overestimate the athlete (center).

2 Another problem involved in the Olympic rating system is: Arbitrarily throwing out the high and low marks in Olympic rating is also a remarkably poor solution to the problems of national bias or human error that arise from time to time on judging panels, because if you throw away a fixed fraction of data you automatically make the assumption that two out of nine judges are necessarily mistaken and the other seven are "correct". This is an unreasonable assumption. so OTM is mechanical because it always trims a fixed fraction of data points at both ends of a data set no matter whether these data points are "good" or "bad".

These disadvantages of the Olympic rating system motivate us to consider trimmed means based on a scaled deviation. It turns out that trimmed means based on a scaled deviation is flexible and random. They may trim some or no sample points and have the power to distinguish outliers no matter which side they are coming from. Detailed comparisons with leading competitors on various robustness and efficiency aspects reveal that the scaled deviation trimmed means behave very well overall and consequently represent very favorable alternatives to the ordinary trimmed means.

Besides trimming, winsorizing is another robust method to mitigate inordinate influence of extreme values. Unlike the trimmed mean, the winsorized mean replaces the outliers with cutting-point values, rather than discarding them. The winsorized mean based on scaled-deviation has the highest breakdown point and is more efficient than the corresponding trimmed mean when the cutting parameter β(see section 2.2) is small. But when β is large, scaled-deviation trimmed mean almost has the same efficiency as winsorized mean. That is because when β is large, there is not too much information contained in both tails. When

there are "bad" points presented from either end, the winsorized mean is less efficient than trimmed mean.

1.2 Scale

A fundamental task in many statistical analyses is to characterize the spread, or variability, of a data set. Measures of scale are simply attempting to estimate this variability.

When assessing the variability of a data set, there are two key components:

1 . How spread out are the data values near center?

2 . How spread out are the tails?

Different numerical summaries will give different weight to these two elements. The choice of scale estimator is often driven by which of these components you want to emphasize.

The histogram is an effective graphical technique for showing both of these components of the spread, however it is just descriptive. There are several common numerical measures of the spread: the variance, standard deviation, average absolute deviation, median absolute deviation, interquartile range and range. The variance, standard deviation, average absolute deviation, and median absolute deviation measure both aspects of the variability, that is, the variability near the center and the variability in the tails. They differ in that the average absolute deviation and median absolute deviation do not give undue weight to the tail behavior. On the other hand, the range only uses the two most extreme points and the interquartile range only uses the middle portion of the data.

The standard deviation is an example of an estimator that is the best in terms of efficiency if the underlying distribution is normal. However, it lacks robustness validity. That is, confidence intervals based on the standard deviation tend to lack precision if the underlying distribution is in fact not normal. It has the lowest possible explosion breakdown point.

The median absolute deviation and the interquartile range are estimates of scale that have robustness of validity. However, the median absolute deviation is not particularly strong for efficiency. The median absolute deviation estimator (MAD) has a low efficiency for normal distributions (36.75%), thereby leading to rather unsatisfactory results for normal models.

The interquartile range is not particularly strong for robustness, it can not reach the highest possible breakdown point.

Rousseeuw and Croux (1993) introduced two alternative statistics more efficient than the MAD, which are defined as

$$S_n^{RC} = c_s \, med_i\{med_j|x_i - x_j|\}, \quad Q_n = d\{|x_i - x_j|;\ i < j\}_{(k)}$$

where c_s, d are consistent coefficients, $k = \binom{h}{2} \approx \binom{n}{2}/4$ with $h = [n/2] + 1$ and (k) is the k-th ordered statistics.

The S_n^{RC} has efficiency for normal distributions 58.23%, while Q_n has 82.27%. They are still not good. Especially at the situation that there are contaminating points presented close to the center, MAD, S_n^{RC} and Q_n are quite inefficient. Motivated by these facts, we introduced scaled deviation trimmed and winsorized standard deviations.

The resulting trimmed (and winsorized) standard deviatons are much more efficient than their predecessors for light-tailed distributions by suitably choosing the cutting parameter and highly efficient for heavy-tailed and skewed distributions. At the same time, they are sharing the best breakdown point robustness of the sample median absolute deviation for any common trimming thresholds. Compared with their predecessors, they can achieve the best efficiency when points around the center are contaminated. Indeed, the scaled deviation trimmed (winsorized) standard deviatons behave very well overall and consequently represent very favorable alternatives to other types of scales.

CHAPTER 2

Trimmed and winsorized means based on a scaled deviation

2.1 Introduction

Tukey trimmed (and winsorized) means are among the most popular estimators of location parameter; see, e.g., Stigler (1977). They overcome the extreme sensitivity of the mean while improving the efficiency of the median for light tailed distributions. The robustness and efficiency are two fundamentally desirable properties of any statistical procedure. They, however, do not work in tandem in general. The trimmed (and winsorized) means somehow can keep a quite good balance between the two. The Tukey trimming scheme is a symmetric one in the sense that it trims the same number of sample points at both ends of data and hence is quite efficient for symmetric distributions. It, however, becomes less efficient when there is even just a slight departure from symmetry, e.g., with one end containing outlying points. Metrical trimming, introduced in Bickel (1965), trims points based on their distance to the center – median and hence is more efficient for contaminated symmetric models. Like the ordinary trimming, it always trims a fixed fraction of sample points, no matter those points are "good" or "bad". This raises a concern as to whether there is a trimming scheme that only trims points that are "bad", which motivates us to consider in this chapter the so-called scaled deviation trimmed and winsorized means.

The main idea behind the new trimming scheme is that sample points are trimmed based

on the magnitude of their scaled (standardized) deviations to a center (say median). Only points with the scaled deviation beyond some fixed threshold are trimmed. This new trimming scheme can lead to the best possible breakdown point (see Section 5.1 for definition) robustness. The resulting estimators are also highly efficient at light-tailed symmetric models and much more efficient than the Tukey trimmed and winsorized means at models with a slight departure from symmetry or with heavy tails. Hence they represent favorable alternatives to their predecessors.

The rest of the chapter is organized as follows. Section 2 defines the scaled deviation trimmed and winsorized means and discusses some primary properties. Section 3 investigates the local robustness, the influence functions, of the estimators. The asymptotic normality of the estimators is established via their asymptotic representations in Section 4. The performance comparison of the estimators with other leading trimmed means with respect to various robustness and efficiency criteria is carried out in Section 5. Concluding remarks in Section 6 end the main body of the chapter. Proofs of main results and auxiliary lemmas are reserved for the Appendix.

2.2 Scaled deviation trimmed and winsorized means

Let $\mu(F)$ and $\sigma(F)$ be some robust location and scale measures of a distribution F. For simplicity, we consider μ and σ being the median (Med) and the median absolute deviations (MAD) throughout the chapter. Assume $\sigma(F) > 0$, namely, F is not degenerate. For a given point x, we define the *scaled deviation* (generalized standardized deviation) of x to the center F by

$$D(x, F) = (x - \mu(F))/\sigma(F). \tag{2.2.1}$$

Now we trim points based on the absolute value of this scaled deviation and define the β *scaled deviations trimmed mean* at F as (c.f. Zuo (2003) for a multi-dimensional version)

$$T^{\beta}(F) = \frac{\int \mathbf{I}(|D(x,F)| \leq \beta)\, \mathrm{w}(D(x,F))x dF(x)}{\int \mathbf{I}(|D(x,F)| \leq \beta)\, \mathrm{w}(D(x,F))dF(x)}, \tag{2.2.2}$$

where $0 < \beta \leq \infty$ and w is an even bounded weight function on $[-\infty, \infty]$ so that the denominator is positive. The heuristic idea behind this definition is that one trims points

that are far ($\beta\sigma$) away from the center and then one *weights* (not just simply average) remaining points based on the robust scaled deviation with larger weights for points closer to the center. When w is a non-zero constant, T^β becomes the plain average of points after the trimming. To cover a broader class of the trimmed means, we consider general w in our treatment. Note that in the extreme case $\beta = \infty$ ($w = c \neq 0$) T^β becomes the usual mean. A concern might be that T^β throws away useful information in the tails. A remedial measure is the *Winsorization*. For the completeness of our discussion, we consider here the *β scaled deviations winsorized mean* at F, defined as

$$T_w^\beta(F) = \frac{\int (x\mathbf{I}(|D(x,F)| \leq \beta) + L(F)\mathbf{I}(x < L(F)) + U(F)\mathbf{I}(x > U(F)))\, \mathrm{w}(D(x,F))dF(x)}{\int \mathrm{w}(D(x,F))dF(x)} \tag{2.2.3}$$

where $L(F) = \mu(F) - \beta\sigma(F)$ and $U(F) = \mu(F) + \beta\sigma(F)$. In the extreme case $\beta = 0$, T_w^β degenerates into the median. For a fixed β, we sometimes suppress β in T^β and T_w^β for convenience.

Since both μ and σ are affine equivariant, i.e., $\mu(F_{aX+b}) = a\mu(F_X) + b$, $\sigma(F_{aX+b}) = |a|\sigma(F_X)$ for any scalars a and b, where F_X is the distribution of X, it is readily seen that $|D(x,F)|$ is affine invaraint and T thus is *affine equivariant* as well. For $X \sim F$ symmetric about θ (i.e. $\pm(X - \theta)$ have the same distribution), it is seen that $T(F) = \theta$, i.e., T is *Fisher consistent*. Without loss of generality, we can assume $\theta = 0$. Let F_n be the usual empirical version of F based on a random sample. It is readily seen that $T(F_n)$ is also affine equivariant. It is *unbiased* for θ if F is symmetric about θ and has an expectation. For $T_w(F)$ and $T_w(F_n)$, all these properties hold.

Two popular trimmed means in the literature are: the ordinary trimmed mean (Tukey (1948)) and the metrically trimmed mean (Bickel (1965), Kim (1992)), defined respectively as

$$T_o^\alpha(F) = \frac{1}{1-\alpha}\int_{F^{-1}(\alpha/2)}^{F^{-1}(1-\alpha/2)} x dF(x), \quad T_m^\alpha(F) = \frac{1}{1-\alpha}\int_{\mu(F)-\nu(F)}^{\mu(F)+\nu(F)} x dF(x), \tag{2.2.4}$$

where $F^{-1}(r)$ is the rth quantile of F and $F(\mu(F)+\nu(F)) - F(\mu(F)-\nu(F)) = 1-\alpha$. It is readily seen that these trimmed means are also affine equivariant and consequently Fisher consistent for symmetric F. The two trimming schemes are probability content based. The

former, however, trims equally ($50\alpha\%$) of points at each tail. This is not always the case for the latter (though total points trimmed are also $100\alpha\%$). At the sample level, $T_o^\alpha(F_n)$ trims a fixed (equal) number of sample points at each tail while $T_m^\alpha(F_n)$ trims sample points at both tails or just one tail with the same total number of points trimmed as in the former case. For performance evaluation and comparison of T^β and T_w^β in later sections, T_o^α and T_m^α will be used as benchmarks.

Note that the proportion of the trimmed points for a fixed β, $P(|D(X,F)| > \beta)$, in $T^\beta(F)$ is not fixed but F-dependent. In the sample case, the proportion of sample points trimmed is random. $T^\beta(F_n)$ may trim some or no sample points. So $T^\beta(F_n)$ is also called a *randomly trimmed mean*. The random trimming scheme here based on the scaled deviation is interconnected to the usual trimming scheme based on the probability content, nevertheless. Indeed, in the population case set β to be the $(1-\alpha)$th quantile of the scaled centered variable $|X - \mu(F)|/\sigma(F)$, then T^β is just a regular trimmed mean that trims $100\alpha\%$ of points at tails for symmetric F. For example, if one wants to trim $\alpha = 10\%$ points at tails, then simply set $\beta = \Phi^{-1}(0.95)/\Phi^{-1}(0.75) = 2.4387$ for normal F and $\beta = 6.3138$ for Cauchy F. A large β corresponds to a small α and consequently is in favor of the efficiency of T^β (and T_w^β) at light-tailed F (see Sections 5.3 and 5.4).

2.3 Influence function

We first investigate the local robustness of the functional $T^\beta(F)$ and $T_w^\beta(F)$. Here F is the assumed distribution. The actual distribution, however, may be (slightly) different from F. A simple departure from F may be due to the point mass contamination of F that results in the distribution $F(\varepsilon, \delta_x) = (1-\varepsilon)F + \varepsilon\delta_x$, where δ_x is the point mass probability distribution at a fixed point $x \in R$. It is hoped that the effect of the slight deviation from F on the underlying functional is small relative to ε. The influence function (IF) of a statistical functional M at a given point $x \in \mathrm{R}$ for a given F, defined as (see Hampel et al. (1996))

$$IF(x; M(F)) = \lim_{\varepsilon \to 0^+} (M(F(\varepsilon, \delta_x)) - M(F))/\varepsilon, \tag{2.3.1}$$

exactly measures the relative effect (influence) of an infinitesimal point mass contamination on M. It is desirable that this relative influence $IF(x; M(F))$ be bounded. This indeed is the case for $T_o^\alpha(F)$ (see, e,g, Serfling (1980)), $T_m^\alpha(F)$ (Kim (1992)), T^β (Theorem 2.3.1) and T_w^β (Theorem 2.3.3) but not for the mean functional with $x - E(X)$ as its influence function for r.v. $X \sim F$.

The integrands in $T^\beta(F)$ ($T_w^\beta(F)$) are complicated functions of F and the derivation of the influence functions thus is a bit involved. We first work out the influence functions of L and U. Assume $F' = f$ exists at μ and $\mu \pm \sigma$ with $f(\mu)$ and $f(\mu+\sigma) + f(\mu-\sigma)$ positive, where μ and σ stand for $\mu(F)$ and $\sigma(F)$. Invoking the chain rule we have the following preliminary results:

$$IF(x; L(F)) = IF(x; \mu(F)) - \beta\, IF(x; \sigma(F)), \tag{2.3.2}$$

$$IF(x; U(F)) = IF(x; \mu(F) + \beta\, IF(x; \sigma(F)), \tag{2.3.3}$$

$$IF(x;\, D(y,F)) = -(D(y,F) IF(x;\sigma(F)) + IF(x;\mu(F)))/\sigma \;\equiv\; h(x,y), \tag{2.3.4}$$

$$IF(x; \mu(F)) = \text{sign}(x-\mu)/(2f(\mu)), \tag{2.3.5}$$

$$IF(x; \sigma(F)) = \frac{\text{sign}(|x-\mu| - \sigma) - 2\, IF(x;\mu(F))(f(\mu+\sigma) - f(\mu-\sigma))}{2(f(\mu+\sigma) + f(\mu-\sigma))}, \tag{2.3.6}$$

Now assume that w is differentiable and f exists at $L(F)$ and $U(F)$. Write L and U for $L(F)$ and $U(F)$ respectively and δ for $\int \mathbf{I}(|D(x,F)| \le \beta)\, \mathrm{w}(D(x,F))dF(x)$ and define

$$\ell_1(x) = \frac{1}{\delta}\Big[(U-T)\,\mathrm{w}(\beta) f(U)\, IF(x;U(F)) - (L-T)\,\mathrm{w}(\beta) f(L)\, IF(x;L(F))\Big] \tag{2.3.7}$$

$$\ell_2(x) = \frac{1}{\delta}\Big[\int_L^U (y-T)\,\mathrm{w}^{(1)}(D(y,F)) IF(x; D(y,F)) dF(y)\Big] \tag{2.3.8}$$

$$\ell_3(x) = \frac{1}{\delta}\Big[\mathbf{I}(x \in [L,U])(x-T)\,\mathrm{w}(D(x,F))\Big] \tag{2.3.9}$$

We then have the influence function of the scaled deviation trimmed mean $T^\beta(F)$ as follows.

Theorem 2.3.1. *Assume that $F' = f$ exists at μ, $\mu \pm \sigma$, $L(F)$, and $U(F)$ with $f(\mu)$ and $f(\mu+\sigma) + f(\mu-\sigma)$ positive and is continuous in small neighborhoods of $L(F)$ and $U(F)$, and that $\mathrm{w}(\cdot)$ is continuously differentiable. Then for a given $0 < \beta < \infty$,*

$$IF(x; T^\beta(F)) = \ell_1(x) + \ell_2(x) + \ell_3(x). \tag{2.3.10}$$

The proof is given in section 5.1, chapter 5.

Under the conditions of Theorem 2.3.1, $IF(x; T^{\beta}(F))$ clearly is bounded and consequently T^{β} is locally robust. For symmetric F and $w = c \neq 0$, the influence function simplifies substantially.

Corollary 2.3.2. *Let $X \sim F$ be symmetric about the origin and* w *a non-zero constant. Under the conditions of Theorem 2.3.1, we have*

$$IF(x; T^{\beta}(F)) = \frac{x\mathbf{I}(x \in [-\beta\sigma, \beta\sigma])}{2F(\beta\sigma) - 1} + \frac{\beta\sigma f(\beta\sigma))\,\mathrm{sign}(x)}{f(0)(2F(\beta\sigma) - 1)}. \tag{2.3.11}$$

A graph of this influence function is given in Figure 2.1. The boundedness is clearly revealed.

To work out the influence function for $T_w^{\beta}(F)$, we write δ_1 for $\int \mathrm{w}(D(x,F))dF(x)$ and define

$$\ell_{w1}(x) = \frac{1}{\delta_1}\int [y\mathbf{I}(L \leq y \leq U) + L\mathbf{I}(y < L) + U\mathbf{I}(y > U) - T_w]\,\mathrm{w}^{(1)}(D(y,F))h(x,y)dF(y) \tag{2.3.12}$$

$$\ell_{w2}(x) = \frac{1}{\delta_1}\int [IF(x;L)I(y < L) + IF(x;U)\mathbf{I}(y > U)]\,\mathrm{w}(D(y,F))dF(y) \tag{2.3.13}$$

$$\ell_{w3}(x) = \frac{1}{\delta_1}[x\mathbf{I}(L \leq x \leq U) + L\mathbf{I}(x < L) + U\mathbf{I}(x > U) - T_w]\,\mathrm{w}(D(x,F)). \tag{2.3.14}$$

We then have the influence function of the scaled deviation winsorized mean T_w^{β} as follows.

Theorem 2.3.3. *Assume that $F' = f$ exists at μ, $\mu \pm \sigma$, $L(F)$, and $U(F)$ with $f(\mu)$ and $f(\mu+\sigma) + f(\mu-\sigma)$ positive and is continuous in small neighborhoods of $L(F)$ and $U(F)$, and that* w *is continuously differentiable with $rw^{(1)}(r)$ being bounded for $r \in \mathrm{R}$. Then for a given $0 < \beta < \infty$,*

$$IF(x; T_w(F)) = \ell_{w1}(x) + \ell_{w2}(x) + \ell_{w3}(x). \tag{2.3.15}$$

Under the conditions of Theorem 2.3.3, $IF(x; T_w(F))$ is readily seen to be bounded and T_w^{β} thus is locally robust. For symmetric F and constant w, the influence function simplifies greatly.

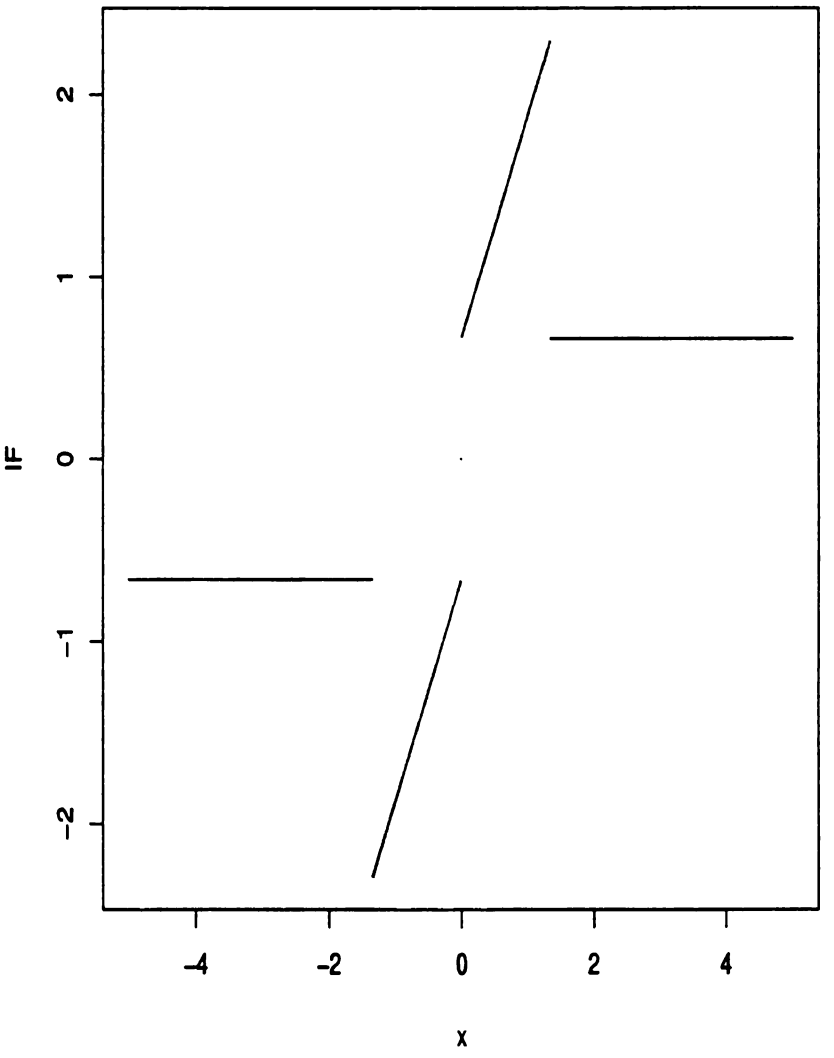

Figure 2.1. Influence function of T^2 for $N(0,1)$ with $\beta = 2$ and a constant weight.

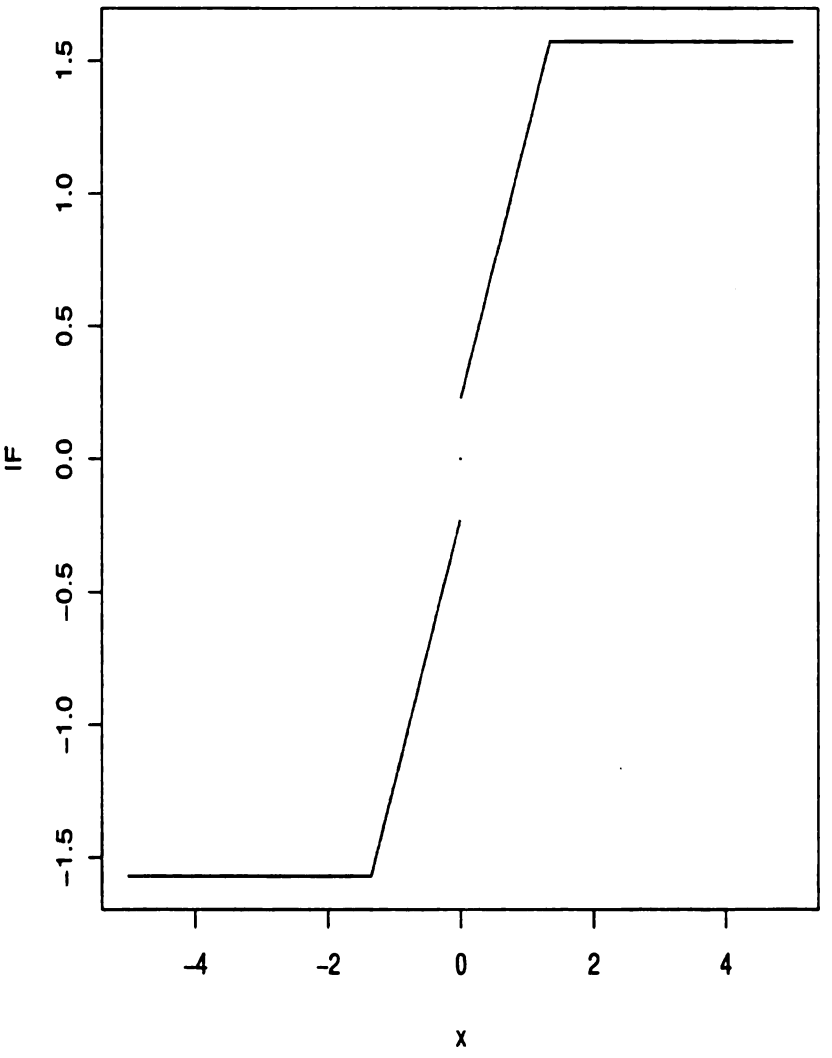

Figure 2.2. Influence functions of T_w^2 for $N(0,1)$ with $\beta = 2$ and a constant weight.

Corollary 2.3.4. *Let $X \sim F$ be symmetric about the origin and* w *a non-zero constant. Under the conditions of Theorem 2.3.3, we have*

$$IF(x; T_w(F)) = \frac{sign(x)}{f(0)} F(-\beta\sigma) + x\mathbf{I}(-\beta\sigma \le x \le \beta\sigma) - \beta\sigma\mathbf{I}(x < -\beta\sigma) + \beta\sigma\mathbf{I}(x > \beta\sigma) \tag{2.3.16}$$

The boundedness of this influence function is very clear and also shown in Figure 2.2.

In addition to being local robustness measures, the influence functions in this section are useful for establishing the limiting distribution of $T^\beta(F_n)$ and $T_w^\beta(F_n)$ in the next section.

2.4 Asymptotic representation and limiting distribution

Establishing the limiting distribution of the scaled deviation trimmed and winsorized means turns out to be a quite challenging task. One possible approach is to establish first the Hadamard differentiability of the functional involved under the supremum norm and then to employ the influence function results. This is exactly what is done for the metrically trimmed mean in Kim (1992). The treatment (proof) there, however, is not quite rigorous. Here we combine an empirical process theory argument with the influence function results obtained in the last section to fulfil the task. Asymptotic representations of the estimators are established first.

Theorem 2.4.1. *Let $F' = f$ exist at μ and be continuous in small neighborhoods of $\mu \pm \sigma$, L and U with $f(\mu)$ and $f(\mu - \sigma) + f(\mu + \sigma)$ positive. Let $w^{(1)}$ be continuous on* R. *Then for $0 < \beta < \infty$*

$$T^\beta(F_n) - T^\beta(F) = \frac{1}{n}\sum_{i=1}^{n} IF(X_i; T^\beta(F)) + o_p(\frac{1}{\sqrt{n}}), \tag{2.4.1}$$

where $IF(x; T^\beta(F))$ is given in Theorem 2.3.1. Consequently

$$\sqrt{n}(T^\beta(F_n) - T^\beta(F)) \to N(0, \tilde{\sigma}^2), \tag{2.4.2}$$

with $\tilde{\sigma}^2 = E(IF(X; T^\beta(F)))^2$.

The distribution F is usually assumed to be symmetric about a point θ in the location setting. By affine equivariance, we can let $\theta = 0$. In this case with a (non-zero) constant weight, $\tilde{\sigma}^2$ takes a much simpler form and will be evaluated at a number of distributions in the next section.

Corollary 2.4.2. *Let $X \sim F$ be symmetric about the origin and w a non-zero constant. Under the conditions of Theorem 2.4.1, we have*

$$\tilde{\sigma}^2 = \frac{\int_{-\beta\sigma}^{\beta\sigma} x^2 dF(x) + \frac{2\beta\sigma f(\beta\sigma)}{f(0)} \int_{-\beta\sigma}^{\beta\sigma} |x| dF(x) + (\frac{\beta\sigma f(\beta\sigma)}{f(0)})^2}{(2F(\beta\sigma) - 1)^2}. \tag{2.4.3}$$

For T_w^β, we can establish results similar to Theorem 2.4.1 and Corollary 2.4.2.

Theorem 2.4.3. *Assume that $F' = f$ exists at μ, $\mu \pm \sigma$, $L(F)$, and $U(F)$ with $f(\mu)$ and $f(\mu+\sigma) + f(\mu-\sigma)$ positive and is continuous in small neighborhoods of $L(F)$ and $U(F)$, and that* w *is continuously differentiable with $rw^{(1)}(r)$ being bounded for $r \in$ R. Then for a given $0 < \beta < \infty$,*

$$T_w^\beta(F_n) - T_w^\beta(F) = \frac{1}{n} \sum_{i=1}^{n} IF(X_i; T_w^\beta(F)) + o_p(\frac{1}{\sqrt{n}}), \tag{2.4.4}$$

where $IF(x; T_w^\beta(F))$ is given in Theorem 2.3.3. Consequently

$$\sqrt{n}(T_w^\beta(F_n) - T_w^\beta(F)) \to N(0, \tilde{\sigma}_w^2), \tag{2.4.5}$$

with $\tilde{\sigma}_w^2 = E(IF(X; T_w^\beta(F)))^2$.

The proof is given in section 5.1, chapter 5.

For F symmetric (about 0) and w non-zero constant, we have a simple specific form for $\tilde{\sigma}_w^2$, which will also be evaluated at a variety of distributions in the next section.

Corollary 2.4.4. *Let $X \sim F$ be symmetric about the origin and w a non-zero constant. Under the conditions of Theorem 2.4.3, we have*

$$\tilde{\sigma}_w^2 = \Big(\frac{1}{f^2(0)} + \frac{4\beta\sigma}{f(0)}\Big) F^2(-\beta\sigma) + 2(\beta\sigma)^2 F(-\beta\sigma) + \int_{-\beta\sigma}^{\beta\sigma} \Big(\frac{2F(-\beta\sigma)}{f(0)}|x| + x^2\Big) dF(x). \tag{2.4.6}$$

With the results obtained in this and last sections, we are in the position to evaluate the performance of the scaled deviation trimmed and winsorized means T^β and T_w^β.

2.5 Performance comparison

We now compare the performance of the scaled deviation trimmed and winsorized means with the trimmed means in (2.2.4), the mean, and the median with respect to robustness (breakdown point and influence function) as well as efficiency (asymptotic and finite sample one) criteria.

2.5.1 Breakdown point

The finite sample breakdown point, a notion introduced by Donoho and Huber (1983), is the most popular measure of the global robustness of an estimator. Roughly speaking, the breakdown point of a location estimator is the minimum fraction of 'bad' (or contaminated) data points in a data set that can render the estimator beyond any bound. More precisely, the *finite sample breakdown point* of a location estimator T at a random sample $X^n = \{X_1, \ldots, X_n\}$ is defined as

$$BP(T, X^n) = \min\{\frac{m}{n} : \sup_{X_m^n} |T(X_m^n) - T(X^n)|\}, \tag{2.5.1}$$

where X_m^n are contaminated data resulting from replacing m points of X^n with arbitrary m points. The *asymptotic breakdown point* (ABP) of T is defined as $\lim_{n\to\infty} BP(T, X^n)$.

Since one bad point can ruin the sample mean ($\bar{X}_n$), the breakdown point of $\bar{X}_n$ thus is $1/n$, the lowest possible value. On the other hand, to break down the sample median, 50% of original points must be contaminated (moved to ∞). Thus the sample median has a breakdown point $\lfloor (n+1)/2 \rfloor / n$, the best among all affine equivariant location estimators. It is readily seen that the regular α trimmed mean in (2.2.4) has a breakdown point $(\lfloor \alpha n/2 \rfloor + 1)/n$ whereas the α metrically trimmed mean in (2.2.4) has a breakdown point $(\lfloor \alpha n \rfloor + 1)/n$ if $0 \leq \alpha \leq 1/2 - 3/(2n)$ or $\lfloor (n+1)/2 \rfloor / n$ otherwise. The breakdown point of the scaled deviation trimmed mean is shown (see Zuo (2003)) to be the same as the median, as long as $w(r)$ is defined on $[0, \beta]$ and $1 \leq \beta < \infty$. Likewise, one can show the scaled deviation winsorized mean has the same breakdown point.

Table 2.1. Breakdown points of mean, trimmed (winsorized) means, and median

	$\bar{X}_n$	T_o^α	T_m^α	T^β	T_w^β	Med
BP	$\frac{1}{n}$	$\frac{\lfloor(\alpha n+2)/2\rfloor}{n}$	$\frac{\lfloor\alpha n+2\rfloor}{n} \wedge \frac{\lfloor(n+1)/2\rfloor}{n}$	$\frac{\lfloor(n+1)/2\rfloor}{n}$	$\frac{\lfloor(n+1)/2\rfloor}{n}$	$\frac{\lfloor(n+1)/2\rfloor}{n}$
ABP	0	$\alpha/2$	$\alpha \wedge 1/2$	1/2	1/2	1/2

The regular trimmed mean T_o^α thus has the lowest breakdown point among the three trimmed means. The metrically trimmed mean T_m^α has a higher breakdown point (twice as high as that of the regular one) when $\alpha \leq 1/2 - 3/(2n)$ and can attain the best breakdown value if α is higher. The scaled deviation trimmed mean T^β always has the best breakdown value as long as $1 \leq \beta < \infty$. The difference in the breakdown points of the trimmed means is due to the difference in trimming schemes. The regular and metrically trimmed means trim always a fixed $100\alpha\%$ points with the former trimming based on the rank of X_i and the latter based on the rank of $|X_i - \mu(X^n)|$. The scaled deviation trimming is based on the value of $|X_i - \mu(X^n)|$ and it trims only points with "large" deviations. The breakdown points of the trimmed means are listed in Table 2.1 ($0 \leq \alpha < 1$, $1 \leq \beta < \infty$). The asymptotic ones are shown in Figure 2.3.

It is noteworthy that the scaled deviation trimming (or Winsorizing) can lead to the best breakdown robustness, while metrically trimming gains breakdown robustness over the ordinary trimming. All trimming schemes improve the breakdown robustness of the sample mean.

2.5.2 Influence function and gross error sensitivity

The breakdown point measures only the global robustness while the influence function can capture the local robustness of an estimator. The two together can provide a more complete picture of robustness. We now look at the influence functions of the trimmed (and winsorized) means.

The boundedness of its influence function is the fundamental concern for a functional being locally robust. The mean functional has an unbounded influence function. The ordinarily and the metrically trimmed means are known to have bounded influence functions; see, e.g.,

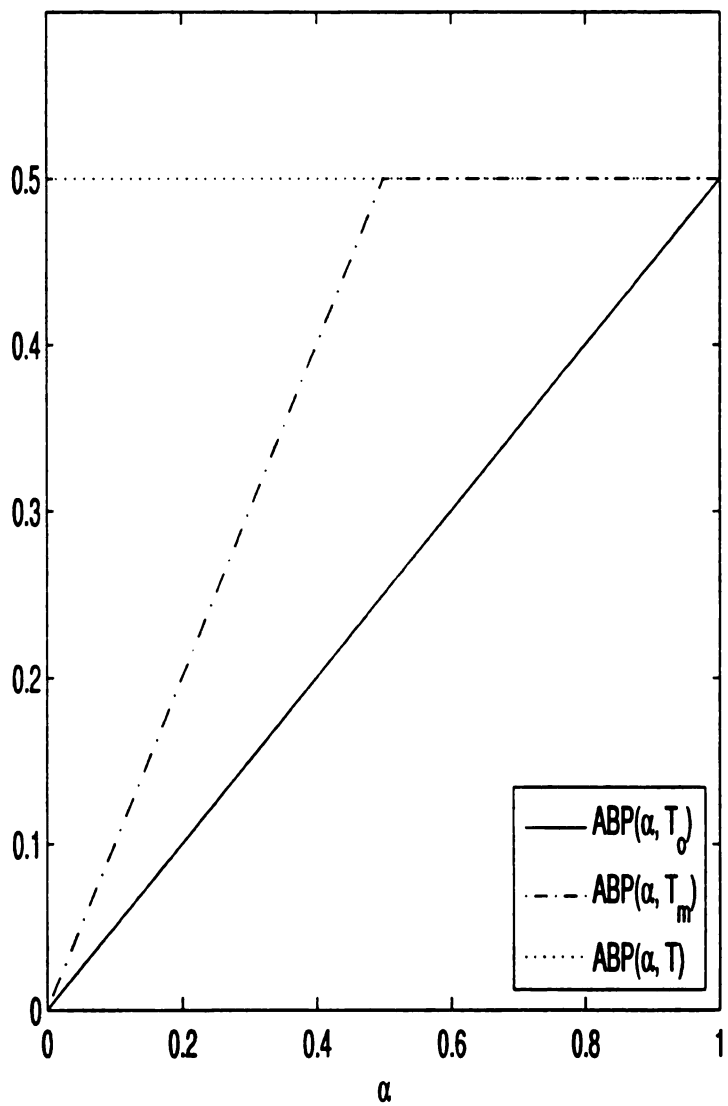

Figure 2.3. Asymptotic breakdown points of trimmed means.

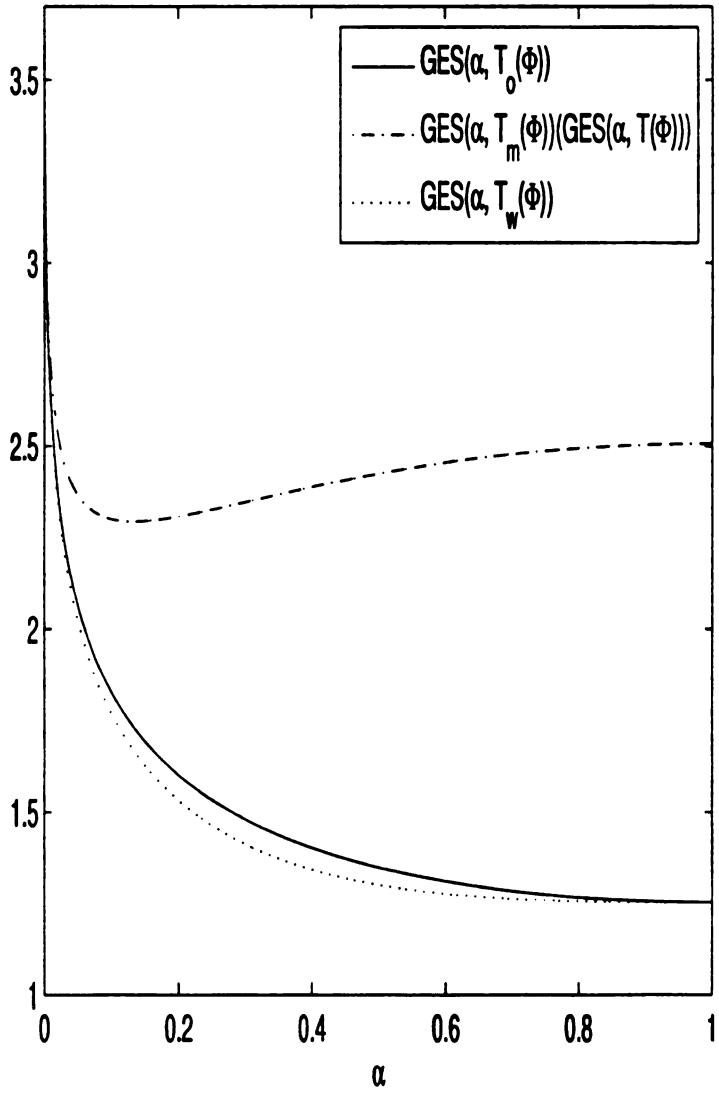

Figure 2.4. Gross error sensitivity of trimmed means.

Serfling (1980) and Kim (1992). In the light of Theorems 2.3.1 and 2.3.3, T^β and T_w^β have bounded influence functions for suitable w and β. Figure 2.8, which plots their influence functions at normal and t (with 3 degrees of freedom) models with $\alpha = 0.1$, confirms this. Here we set $\beta = \beta(F, \alpha)$ so that $100\alpha\%$ of points are trimmed in each of the trimming cases (see Section 2). For convenience, we also set $\mathrm{w} = c \neq 0$ in T^β and T_w^β. Note that T_o^α and T^β and their influence functions are the same under this setting. Indeed, the influence function in Theorem 2.3.1 becomes

$$IF(x; T^\beta(F)) = \frac{Uf(U)\, IF(x; U^*(F)) - Lf(L)\, IF(x; L^*(F)) + x\mathbf{I}(x \in [L, U]) - T^\beta}{1 - \alpha}, \tag{2.5.2}$$

where we have for $\lambda = \beta(F)\sigma(F)$

$$IF(x; U^*(F)) = IF(x; \mu(F)) + IF(x; \lambda(F)); IF(x; L^*(F)) = IF(x; \mu(F)) - IF(x; \lambda(F))$$

$$IF(x; \lambda(F)) = \frac{(1-\alpha) - \mathbf{I}(\mu - \lambda \leq x \leq \mu + \lambda) - IF(x; \mu(F))(f(\mu + \lambda) - f(\mu - \lambda))}{f(\mu + \lambda) + f(\mu - \lambda)}.$$

Thus $IF(x; T^\beta(F))$ is the same as that of T_m^α in Kim (1992). Since a pure normal model is rare in practice, we thus consider contaminated normal models. With the same α and $\beta = \beta(F, \alpha)$ as above, the influence functions of T_m^α and T^β in the contaminated normal models, plotted also in Figure 2.8, become different (but all are still bounded). In terms of the bounded influence function criterion, we conclude that all the trimmed and winsorized means are *equally* robust (locally).

Besides boundedness, one can also look at the magnitude of the supremum of $|IF(x; T(F))|$, the so-called the *gross error sensitivity* (GES) of T at F (Hampel et al. (1986))

$$GES(T(F)) = \sup_{x \in R} |IF(x; T(F))|, \tag{2.5.3}$$

which measures the worst case effect on T of an infinitesimal point mass contamination. Generally speaking, a smaller GES is more desirable. For T^β and T_w^β, it is readily seen that their GES depends on the values of β (or α if $\beta = \beta(F, \alpha)$) and the weight function w. As a

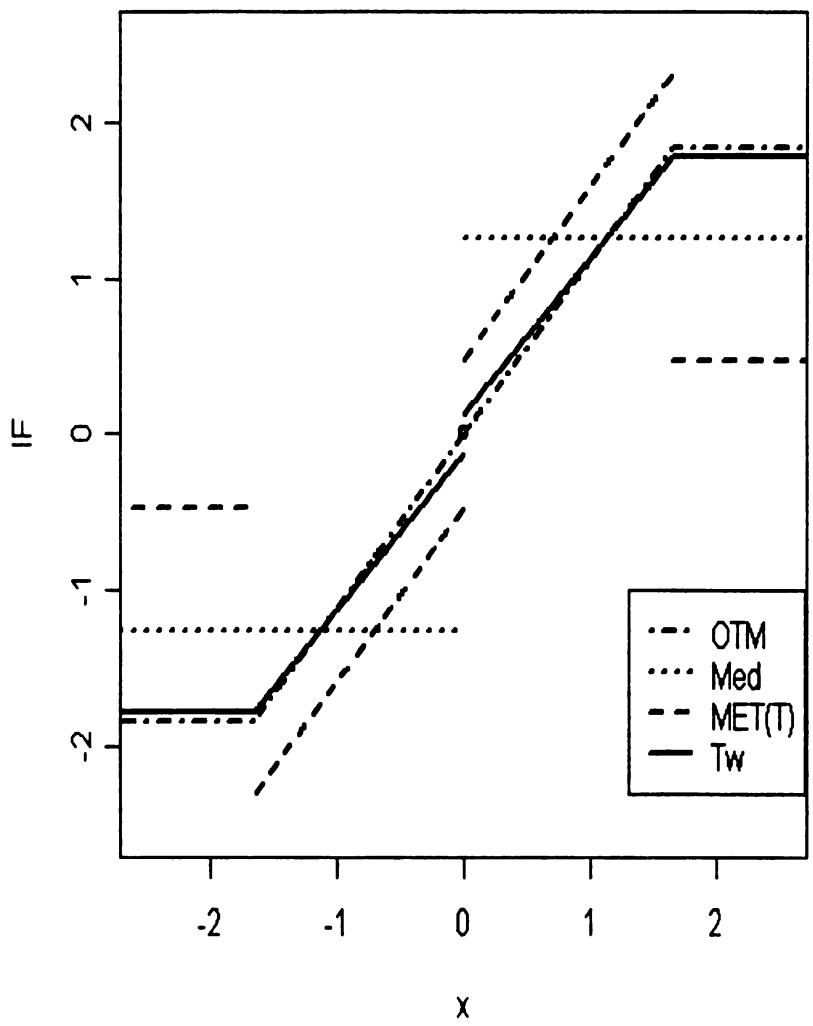

Figure 2.5. Influence functions of the trimmed and winsorized means at normal.

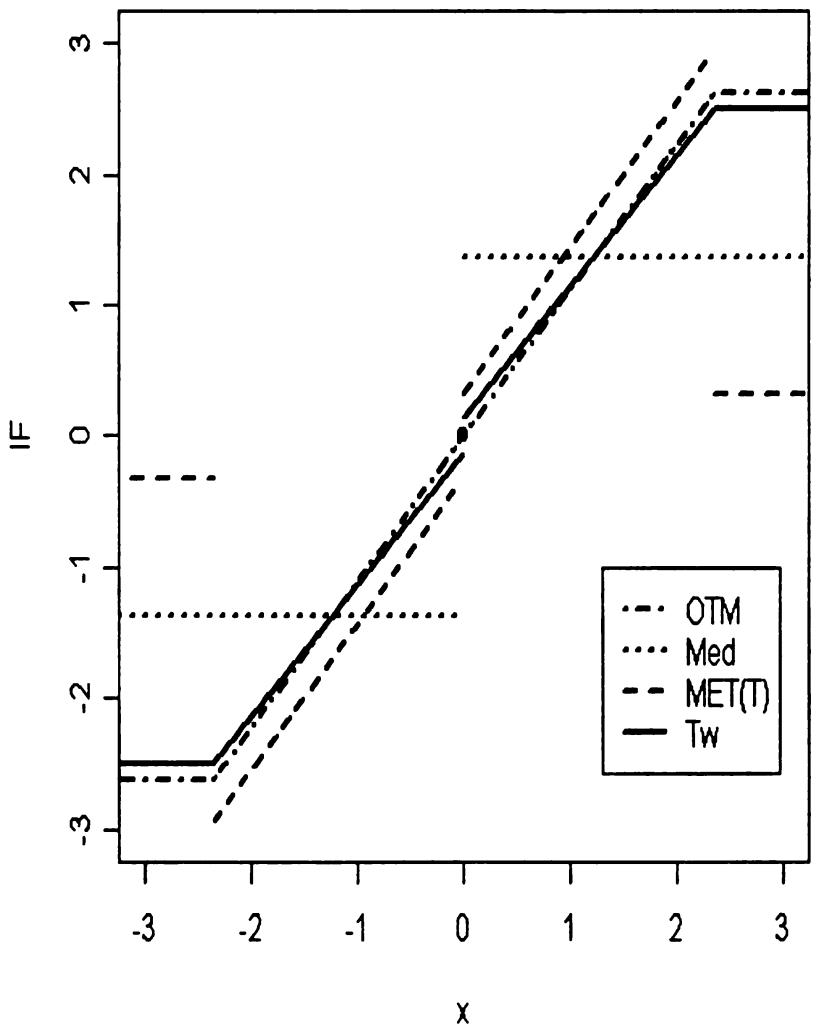

Figure 2.6. Influence functions of the trimmed and winsorized means for $t(3)$ with $\alpha = 0.1$.

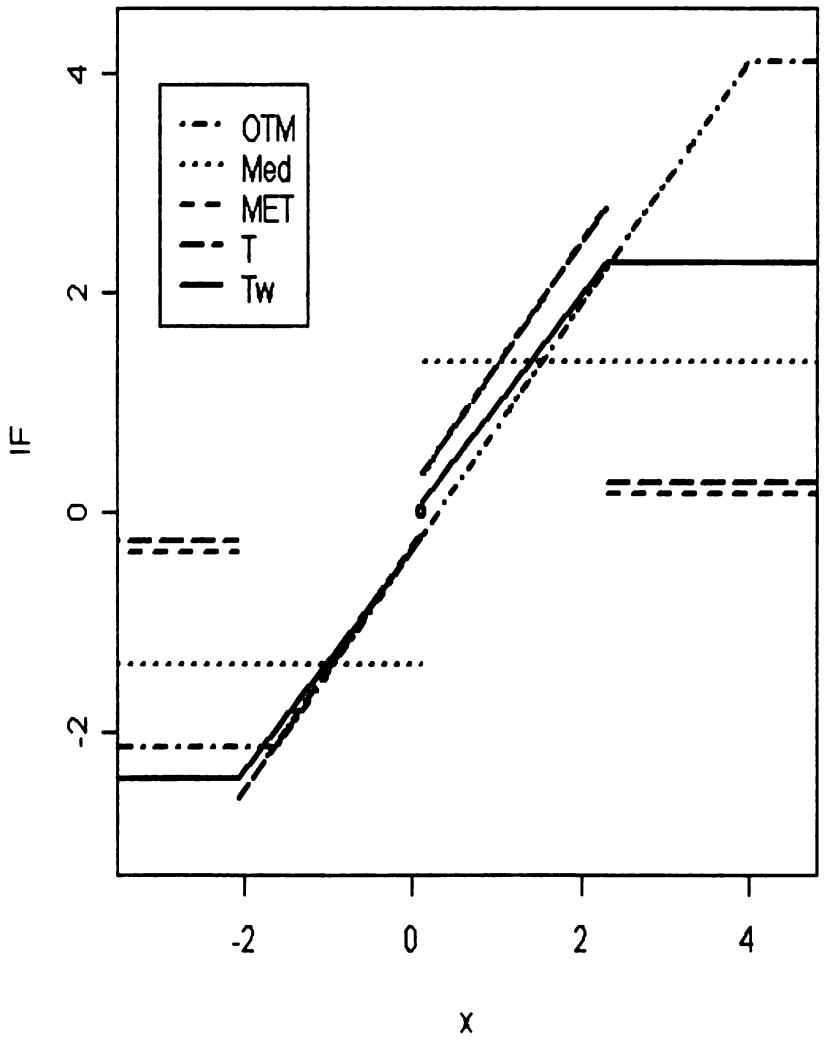

Figure 2.7. Influence functions of the trimmed and winsorized means for $0.9N(0,1) + 0.1N(4,9)$ with $\alpha = 0.1$.

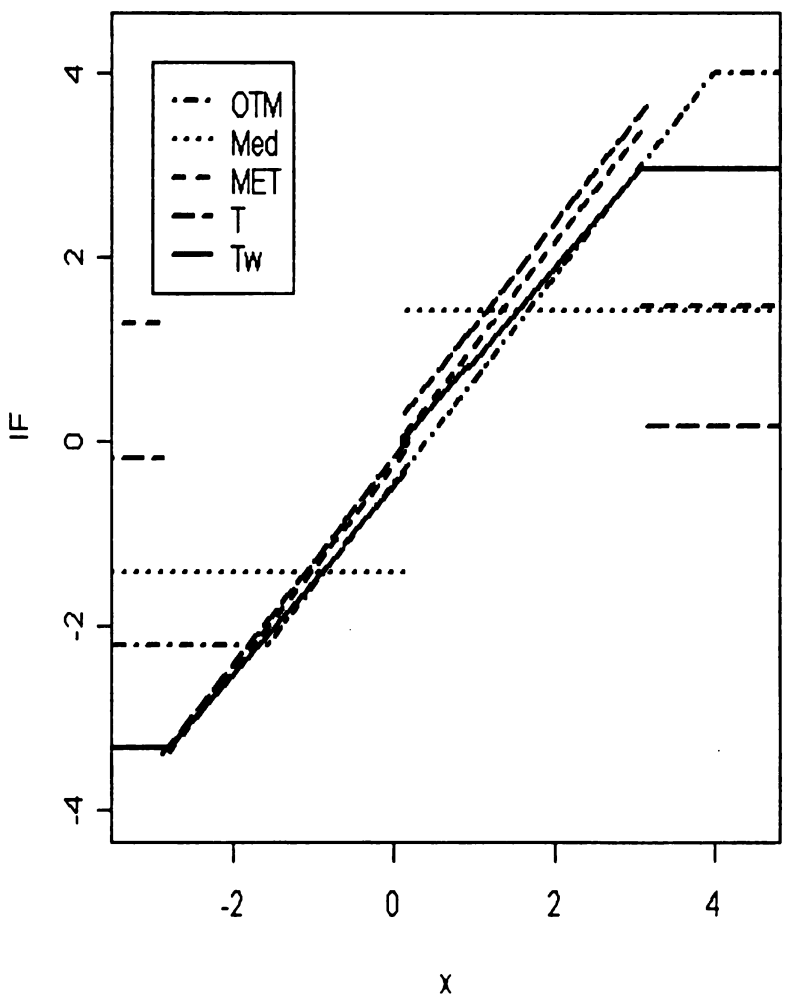

Figure 2.8. Influence functions of the trimmed and winsorized means for $0.9N(0,1)+0.1N(4,0.5)$ with $\alpha = 0.1$.

Table 2.2. GESs of mean, trimmed (winsorized) means, and median at symmetric F

Mean	T_m^α	T^β	T_o^α	T_w^β	Med
$+\infty$	$\frac{F^{-1}(1-\alpha/2)C(F,\alpha))}{(1-\alpha)}$	$\frac{F^{-1}(1-\alpha/2)C(F,\alpha))}{(1-\alpha)}$	$\frac{F^{-1}(1-\alpha/2)}{(1-\alpha)}$	$F^{-1}(1-\alpha/2)+\frac{\alpha}{2f(0)}$	$\frac{1}{2f(0)}$

Table 2.3. GESs of mean, trimmed (winsorized) means, and median at asymmetric F

	Mean	T_o^α	T_m^α	T^β	T_w^β	Med
		$F = .9N(0,1) + .1N(4,9)$				
GES	$+\infty$	4.2419	2.8563	2.7781	2.4422	1.3787
		$F = .9N(0,1) + .1N(4,.5)$				
GES	$+\infty$	3.9989	4.6025	3.5290	3.0995	1.4062

possible way to make a comparison, we again set $\beta = \beta(F,\alpha)$ (w $= c \neq 0$) in the following discussion.

Now first consider the case that F is symmetric about the origin and meets the conditions in Corollary 2.3.2. Define $C(F,\alpha) = 1 + f(F^{-1}(1-\alpha/2))/f(0)$. The GES's of the trimmed (and winsorized) means are listed in Table 2.2 for general F and illustrated in Figure 2.4 as functions of α at $F = \Phi$, the most interesting and common normal distribution used in practice.

It can be shown that for any $0 < \alpha < 1$, the GES's in Table 2.2 are increasingly small with the median having the smallest one if $F' = f$ exists and is unimodal. This is also confirmed in Figure 2.3 for $F = \Phi$. On the other hand, it is noted from Figure 2.8 that the influence function of T_o^α at symmetric F has larger absolute values than those of T_m^α, T^β and T_w^β at most values of $x \in R$, a result much favorable to the efficiency of the latter three (see Section 5.3). Again in practice data follow more often than not an asymmetric model. It is therefore sensible to consider the GESs of T^β and T_w^β at F that slightly deviates from a symmetric model. Table 2.3 lists the GES results of trimmed and winsorized means at such models with $\alpha = 0.1$ and $\beta = \beta(\Phi, 0.1)$.

The relationship between the GES's of the the trimmed (and winsorized) means for sym-

metric F is altered under just slight contamination. Table 2.3 indicates that T^{β} can have the smallest GES among three trimmed means at both contaminated models, whereas both T_o^{α} and T_m^{α} can have the largest GES. On the other hand, T_w^{β} has smaller GES than three trimmed means while the median enjoys the smallest GES under slight deviation from symmetry. The GES advantage of T^{β} and T_w^{β} over the two competitors is due to the unique trimming mechanism. With $\beta = \beta(\Phi, 0.1)$, T^{β} (T_w^{β}) trims only the "bad" points that have large scaled deviations while T_o^{α} and T_m^{α} always trim a fixed 10% percent of points even if they are "good" points.

2.5.3 Large sample relative efficiency

Now we evaluate the performance of the trimmed (and winsorized) means in terms of their efficiency behavior (relative to the sample mean). First we examine the asymptotic relative efficiency (ARE). Table 2.4 lists the ARE results of T^{β} and T_w^{β} at a number of light- and heavy-tailed *symmetric* distributions with different β values. Here we again set w $= c > 0$. The table reveals that (i) at normal model with *large* β both T^{β} and T_w^{β} can be highly efficient relative to the mean; (ii) their efficiency increases as the tail of the distribution becomes heavier and exceeds 100% at heavy tailed distributions; and (iii) T_w^{β} is more efficient than T^{β} for small β or normal distribution but when β get larger and the tail gets heavier T^{β} becomes more efficient.

To compare the efficiency behavior of T^{β} (T_w^{β}) with that of T_o^{α} and T_m^{α}, we again face the issue of the choices of the values of α and β. One possible choice again is $\beta = \beta(F, \alpha)$ as above. Such a choice is somewhat in favor of T_o^{α} and T_m^{α} since for very small α values they become quite efficient at normal and other light tailed distributions whereas their breakdown points become very low at those values but those of T^{β} and T_w^{β} are always the best. With the choice $\beta = \beta(F, \alpha)$ and $\alpha = 0.01$ (and w $= c > 0$), the AREs of the trimmed and winsorized means and the median are listed Table 2.5. Note again that $T_m^{\alpha} = T^{\beta}$ under this setting for symmetric F.

Examining Table 2.5 reveals that in terms of efficiency: (i) T^{β} performs *better* than T_o^{α} and T_w^{β} at *DE*, t_3 and t_4 (with T_o^{α} worst among the five) and *best* at t with degrees of freedom (df) 4 (to 7) and (ii) T_o^{α} performs best at normal or very light-tailed F's such as

Table 2.4. AREs of T^β and T_w^β relative to the mean

β		1	2	3	4	5	6	7
$N(0,1)$	T_w^β	0.7589	0.9262	0.9882	0.9987	0.9999	1.0000	1.0000
	T^β	0.4678	0.5630	0.7762	0.9377	0.9901	0.9991	0.9999
LG(0,1)	T_w^β	0.9644	1.0852	1.0754	1.0410	1.0190	1.0081	1.0033
	T^β	0.6346	0.7640	0.9231	1.0004	1.0167	1.0133	1.0077
DE(0, 1)	T_w^β	1.8924	1.6073	1.3656	1.2125	1.1221	1.0696	1.0394
	T	1.7060	1.5493	1.4218	1.3090	1.2159	1.1452	1.0948
t_3	T_w^β	1.8649	1.9334	1.7709	1.6059	1.4844	1.3982	1.3358
	T^β	1.3202	1.5952	1.7613	1.7500	1.6700	1.5826	1.5061

Table 2.5. AREs of trimmed and winsorized means and median with $\alpha = 0.01$ $\beta = \beta(F, \alpha)$

	T_o^α	T_m^α	T^β	T_w^β	Med
$N(0,1)$	0.9982	0.9179	0.9179	0.9981	0.6366
$LG(0,1)$	1.0192	1.0161	1.0161	1.0220	0.8225
$DE(0,1)$	1.0383	1.1107	1.1107	1.0483	2.0000
t_3	1.2953	1.4645	1.4645	1.3047	1.6211
t_4	1.1168	1.2011	1.2011	1.1226	1.1250

logistic (or t with df ≥ 10) while the median is the best at $DE(0,1)$ or very heavy-tailed F's such as t with df ≤ 4.

The results in Tables 2.4 and 2.5 are in the asymptotic sense and the F's are symmetric. This raises the concern as to whether these results are valid at finite sample practice and for F's with slight departure from symmetry. We answer this question via finite sample simulation studies.

2.5.4 Finite sample relative efficiency

We now conduct Monte Carlo studies to investigate the efficiency behavior of T^β and T_w^β at finite samples for normal and contaminated normal models. Here $\theta = 0$ is regarded as

Table 2.6. REs of trimmed and winsorized means with $\beta = 7$

		T^β	T_w^β	Mean	T^β	T_w^β	Mean	T^β	T_w^β	Mean
n		$\varepsilon = 0\%$			$\varepsilon = 10\%$			$\varepsilon = 20\%$		
20	EMSE	0.05	0.05	0.05	0.11	0.18	0.25	0.34	0.60	0.77
	RE	0.99	1.00	1.00	2.31	1.40	1.00	2.30	1.29	1.00
40	EMSE	0.03	0.03	0.03	0.07	0.15	0.20	0.28	0.56	0.70
	RE	1.00	1.00	1.00	2.99	1.40	1.00	2.51	1.26	1.00
60	EMSE	0.02	0.02	0.02	0.06	0.14	0.19	0.26	0.55	0.68
	RE	1.00	1.00	1.00	3.41	1.40	1.00	2.61	1.24	1.00
80	EMSE	0.01	0.01	0.01	0.05	0.13	0.18	0.25	0.54	0.67
	RE	1.00	1.00	1.00	3.70	1.40	1.00	2.67	1.24	1.00
100	EMSE	0.01	0.01	0.01	0.05	0.13	0.18	0.24	0.54	0.66
	RE	1.00	1.00	1.00	3.95	1.40	1.00	2.73	1.24	1.00

the target parameter to be estimated. For an estimator T the empirical mean squared error (EMSE) is: EMSE $= \frac{1}{m}\sum_{j=1}^{m} |T_j - \theta|^2$, where m is the number of samples generated and T_j is the estimate based on the jth sample. The relative efficiency (RE) of T is then obtained by dividing the EMSE of the sample mean by that of T. We generated $m = 50,000$ samples from $(1-\varepsilon)N(0,1)+\varepsilon N(4,9)$ with $\varepsilon = 0, .1$ and $.2$ for different sample sizes n. Some results are listed in Table 6 with $\beta = 7$.

The RE results at $\varepsilon = 0$ confirm at the validity at finite samples of the asymptotic result in Table 2.4 with $\beta = 7$. When there is just a 10% or 20% contamination, both T^β and T_w^β become overwhelmingly more efficient than the mean with T^β substantially more efficient than T_w^β.

To compare the efficiency of the trimmed and winsorized means, we now set $\beta = \beta(F, \alpha)$ again. Note that at this setting T_o^α and T_m^α have very low breakdown point for small α whereas T^β and T_w^β always enjoy the best breakdown point. We consider $F = \Phi$ and set $\alpha = 0.01$. Table 2.7 lists the relative efficiency results at $(1-\varepsilon)N(0,1)+\varepsilon N(4,9)$ for $\varepsilon = 0.1$ and 0.2. Note that we set $n = 200$ or larger so that T_o^α trims at least one sample point at each end of data.

Table 2.7. REs of trimmed and winsorized means with $\alpha = 0.01$ $\beta = \beta(\Phi, \alpha)$

	T_o^α	T_m^α	T^β	T_w^β	Med	T_o^α	T_m^α	T^β	T_w^β	Med
n	$\varepsilon = 10\%$					$\varepsilon = 20\%$				
200	1.158	1.579	19.70	2.996	7.988	1.075	1.269	21.79	2.384	9.015
400	1.167	1.615	30.35	3.059	9.595	1.076	1.278	25.71	2.394	9.566
600	1.171	1.630	38.34	3.102	10.48	1.079	1.281	27.33	2.397	9.780
800	1.172	1.636	43.39	3.114	10.90	1.079	1.282	28.39	2.403	9.914
1000	1.173	1.640	47.42	3.125	11.24	1.079	1.283	28.96	2.403	9.969

Our simulation results for $\varepsilon = 0.0$ (not listed in the table) confirms the validity of the asymptotic ones in Table 2.5 for $N(0,1)$. On the other hand, when there is just a 10% or 20% contamination in the distribution, all the estimators become more efficient than the sample mean with T_o^α, T_m^α, T_w^β, median, and T^β being increasingly more efficient, reflecting the robustness of these estimators. It is remarkable that T^β is overwhelmingly more efficient than all other estimators.

2.6 Remarks

Unlike the mean, all the trimmed and winsorized means discussed in the chapter have bounded influence functions for suitable distributions and weight functions and hence are locally robust. In terms of the global robustness, the scaled deviation trimmed and winsorized means T^β and T_w^β are exceptional in the sense that they can enjoy the best possible breakdown point robustness for any $\beta \geq 1$ whereas the ordinary and metrically trimmed means T^α and T_m^α have much lower breakdown point for typical choices of α and the mean has the worst.

Relative to the mean, T^β and T_w^β are highly efficient for large β's at light-tailed symmetric distributions and much more efficient at heavy-tailed ones. When β is set to be $\beta(F, \alpha)$ so that $100\alpha\%$ points are trimmed, T^β and T_w^β are less efficient than T^α at light-tailed symmetric distributions but become much more efficient at heavy-tailed or contaminated

symmetric ones. The latter models seem more popular than the light-tailed symmetric ones in practice.

The advantages of T^β and T_w^β over T^α and T_m^α on robustness and efficiency are due to the difference in the trimming schemes. The latter trim always a fixed fraction of sample points no matter they are "good" or "bad" whereas the former trim only when there are "bad" points.

A very legitimate concern for T^β and T_w^β in practice is about the choice of β value. In light of our simulation studies, we recommend a β value between 4 to 7 so that T^β and T_w^β can be very efficient at both light- and heavy-tailed distributions. Instead of a fixed value one might also adopt an adaptive data-driven approach to determine an appropriate β value. For a given data set, one determines a value for β based on the heaviness of the tail. Generally speaking, a large value of β is selected for a light-tailed data set while a smaller value for a heavy-tailed one.

The basic idea of adaptive trimming above exists in the literature, see, e.g., Jaeckel (1971), Hogg (1974), and Jureckova et al. (1994). Furthermore, a random trimming idea also appeared in Shorack (1974), though the trimming proportion there is (asymptotically) fixed. Consequently the trimmed and winsorized means in Shorack (1974) are different in essence from the ones in this chapter. We note that $T^\beta(F_n)$ in (2.2.2) has some connection with (however is very different from) the (scaled version of) Huber-type skipped mean (see, e.g., Hampel et al. (1986)), the solution T_n of $\sum_i X_i \mathbf{I}(-\beta \le (X_i - T_n)/\sigma_n \le \beta) / \sum_i \mathbf{I}(-\beta \le (X_i - T_n)/\sigma_n \le \beta) = T_n$.

We remark that a general multi-dimensional version of (2.2.2) (but not (2.2.3)) has been thoroughly studied in Zuo (2003). Here we focus on the performance evaluation of the specific one-dimensional version T^β and provide specific and concrete results for influence function and limiting distribution as well. Multidimensional version of (2.2.3) is yet to be studied.

CHAPTER 3

Trimmed and winsorized standard deviations based on a scaled deviation

3.1 Introduction

Scale is an important parameter of interest which can tell us how spread out the data is. Although there are many robust location estimators, robust scale estimators seems far much less. Despite this fact, finding robust scale statistics with a high level efficiency is always the goal of many statisticians. Two types of trimmed standard deviations are introduced and discussed by Welsh and Morrison (1990). But unfortunately, these estimators can not reach highest breakdown point while keeping satisfactory efficiency. So this chapter will not discuss these versions of trimming. Another attempt was made by Rousseeuw and Croux (1993), in which two types of high breakdown estimators are introduced. Those estimators are built through recursive medians. Put aside computation time, the efficiency is not very good at light-tailed distribution, though they are better than the widely used robust estimator, the median absolute deviation (MAD). And these estimators perform poorly when points in the neighborhood of the center are contaminated, because they only use the middle half of the data points or "extended" data points. These situations motivate us to consider in this chapter the so-called scaled-deviation trimmed and winsorized standard deviations.

The high breakdown scale estimators–scaled-deviation trimmed and winsorized standard

deviations. Trimmed and winsorized standard deviations enjoy the highest breakdown point and bounded influence functions for a variety of distributions. They are also much more efficient at light-tailed symmetric models than their predecessors and highly efficient for heavy tailed or skewed distributions. They also have the best performance among high breakdown estimators when the points somewhere around the center are contaminated. Hence they represent favorable alternatives to their predecessors.

Section 3.2 introduces Scaled-Deviation trimmed/winsorized standard deviations; Section 3.3 is devoted to the study of the local robustness; Asymptotic representation and asymptotic normality are treated in section 3.4. Comparisons of influence functions and Gross Error Sensitivities of various high breakdown point scale estimators are undertaken in section 3.5. Proofs of main results and auxiliary lemmas are reserved for Chapter 5.

3.2 Scaled-deviation trimmed and winsorized standard deviation

Let $\mu(F)$ and $\sigma(F)$ be some robust location and scale measures of a distribution F. For simplicity, we consider μ and σ being the median (Med) and the median absolute deviation (MAD) throughout the chapter. Assume $\sigma(F) > 0$, namely, F is not degenerate. For a given point x, we define the scaled deviation (generalized standardized deviation) of x to the center F by

$$D(x, F) = (x - \mu(F))/\sigma(F). \tag{3.2.1}$$

Now one trims points based on the absolute value of this scaled deviation and define the β scaled deviations trimmed variance functional as

$$S^2(F) = c_t \frac{\int \mathbf{I}(|D(x,F)| \leq \beta) w_2(D(x,F))(x - T_1(F))^2 dF(x)}{\int \mathbf{I}(|D(x,F)| \leq \beta) w_2(D(x,F)) dF(x)} \tag{3.2.2}$$

where c_t is the consistency coefficient, $T_1(F)$ the β scaled-deviations trimmed mean which is defined through

$$T_i(F) = \frac{\int \mathbf{I}(|D(x,F)| \leq \beta) w_i(D(x,F)) x dF(x)}{\int \mathbf{I}(|D(x,F)| \leq \beta) w_i(D(x,F)) dF(x)} \qquad i = 1, 2. \tag{3.2.3}$$

where $0 < \beta \leq \infty$ and w_i $(i = 1,\ 2)$ is an even bounded weight function on $[-\infty, \infty]$ so that the denominator is positive. The heuristic idea behind this location definition is that one trims points that are a robust distance $(\beta\sigma)$ away from the robust center μ and then one obtains a robust and efficient location estimator by weighting (including simply averages) left points, which integrates the robustness of σ, μ and efficiency of mean. When w_i (i=1, 2) is a non-zero constant, T_i (i=1, 2) and S^2 is the plain average and plain variance of points after trimming. We consider general w_i (i=1, 2) in our treatment which will contain a broader class of the trimmed means. Note that in the extreme case $\beta = \infty$ $(w = c \neq 0)$ T_i and S^2 become the usual mean and usual variance and c_t becomes 1.

Another robust estimator of scale is the winsorized standard deviation. Like the trimmed standard deviation, the winsorized standard deviation eliminates the outliers if they exist. Unlike the trimmed scale, the winsorized scale replaces the outliers with cutting-point values, rather than discarding them. The definition of the β scaled deviation wisonsorized variance functional is given by

$$\begin{aligned} S_w^2(F) =& c_w\Big[\int ((x - T_{w1}(F))^2\mathbf{I}(|D(x,F)| \leq \beta) + (L(F) - T_{w1})^2\mathbf{I}(x < L(F)) \\ &+ (U(F) - T_{w1})^2\mathbf{I}(x > U(F)))\, \mathrm{w}_2(D(x,F))dF(x)\Big] / \int \mathrm{w}_2(D(x,F))dF(x) \end{aligned} \tag{3.2.4}$$

where c_w is the consistency coefficient, $T_{w1}(F)$ the β scaled deviations winsorized mean which is defined through

$$\begin{aligned} T_{wi}(F) =& \Big[\int (x\mathbf{I}(|D(x,F)| \leq \beta) + L(F)\mathbf{I}(x < L(F)) \\ &+ U(F)\mathbf{I}(x > U(F)))\, \mathrm{w}_i(D(x,F))dF(x)\Big] / \int \mathrm{w}_i(D(x,F))dF(x) \qquad i = 1,2, \end{aligned} \tag{3.2.5}$$

where $L(F) = \mu(F) - \beta\sigma(F)$ and $U(F) = \mu(F) + \beta\sigma(F)$. In the extreme case $\beta = 0$, T_{wi} degenerate into the median and S_w zero.

The β scaled deviation trimmed/Wisonsorized standard deviation at F is the square root of corresponding variance.

Two popular high breakdown scales are introduced by Rousseeuw and Croux (1993). we

denote them by S_n^{RC} and Q, which are defined as

$$S_n^{RC} = c_s \, med_i\{med_j|x_i - x_j|\}, \quad Q_n = d\{|x_i - x_j|;\, i < j\}_{(k)} \tag{3.2.6}$$

where c_s, d are consistent coefficients and $k = \binom{h}{2} \approx \binom{n}{2}/4$ with $h = [n/2] + 1$. We will compare $S(F)$, $S_w(F)$ with S^{RC}, Q. It turned out that $S(F)$ and $S_w(F)$ are more flexible, and more efficient than these two types of scales for light tailed distribution and for the situation that "bad points" are coming from the neighborhood of the center. Those estimators are built through recursive medians. For performance evaluation and comparison of S and S_w in later sections, S_n^{RC} and Q will be used as benchmarks.

The scaled-deviation trimmed/winsorized means or variances are different from usual ones. Note that the proportion of the trimmed points for a fixed β, $P(|D(X,F)| > \beta)$, in $T(F)$ (or $S(F)$) is not fixed but F-dependent. In the sample case, the proportion of sample points trimmed is not fixed but random. $T(F_n)$ (or $S(F_n)$) may trim some or no sample points. So $T(F_n)$ (or $S(F_n)$) are flexible rather than mechanical. On the other hand, there is some connections between the estimators introduced above and the usual trimming/winsorizing scheme based on the probability content. Indeed, set β to be the $(1-\alpha)th$ quantile of the scaled centered variable $|X - \mu(F)|/\sigma(F)$, then $T(F)$ (or $S(F)$) are just the regular trimmed mean and standard deviation after trimming $100\alpha\%$ of points at tails for symmetric F. For example, if one wants to trim $\alpha = 10\%$ points at tails, then simply set $\beta = \Phi^{-1}(0.95)/\Phi^{-1}(0.75) = 2.4387$ for normal F and $\beta = 6.3138$ for Cauchy F. A large β corresponds to a small trimmed proportion (α) and consequently is in favor of the efficiency of the scaled deviation trimmed mean and standard deviation at light-tailed F.

In order to keep clarity, we need to adopt some notations and write

$$S^2(F) = c_t(s(F) - \lambda(F)), \quad S_w^2(F) = c_w(s_w(F) - \lambda_w(F)),$$

$$\delta_i = \int \mathbf{I}(|D(x,F)| \le \beta) w_i(D(x,F))dF(x), \quad \delta_{wi} = \int w_i(D(x,F))dF(x) \quad i = 1,2.$$

with

$$
\begin{aligned}
s(F) &= \int \mathbf{I}(|D(x,F)| \le \beta) x^2 w_2(D(x,F)) dF(x)/\delta_2(F) \\
\lambda(F) &= 2T_1(F)T_2(F) - T_1(F)^2 \\
s_w(F) &= \Big[\int (x^2 \mathbf{I}(|D(x,F)| \le \beta) + L(F)^2 \mathbf{I}(x < L(F)) + U(F)^2 \mathbf{I}(x > U(F))) \\
&\qquad \mathrm{w}_2(D(x,F)) dF(x) \Big] / \delta_{w2}(F) \\
\lambda_w(F) &= 2T_{w1}(F)T_{w2}(F) - T_{w1}(F)^2.
\end{aligned}
$$

T, T_w, S, and S_w are affine equivariant because both μ and σ are affine equivariant, i.e., $\mu(F_{aX+b}) = a\mu(F_X) + b$, $\sigma(F_{aX+b}) = |a|\sigma(F_X)$ for any scalars a and b, where F_X is the distribution of X. For $X \sim F$ symmetric about θ(i.e. $\pm(X-\theta)/\eta(\eta > 0)$ have the same distribution F_0), it is seen that $T(F) = \theta$ and $S(F) = \eta$ ($c_t = 1/S(F_0)$), i.e., T and S^2 are Fisher consistent. Without loss of generality, we can assume $\theta = 0$ and $\eta = 1$. Let F_n be the usual empirical version of F based on a random sample. It is readily seen that $T(F_n)$ and $S^2(F_n)$ are also affine equivariant. $T(F_n)$ is unbiased for θ if F is symmetric about θ and has an expectation and $S^2(F_n)$ ($c_t = 1/(s(F_0) - \lambda(F_0))$) is unbiased for η if F has a variance. All these properties also hold for T_w and S_w^2.

3.3 Influence Function

We first investigate the local robustness of the functional $S(F)$ and $S_w(F)$ through influence functions. Here F is the assumed distribution. The actual distribution, however, may be (slightly) different from F. A simple departure from F may be due to the point mass contamination of F that results in the distribution $F(x, \delta_x) = (1-\varepsilon)F + \varepsilon\delta_x$, where δ_x is the point mass probability distribution at a fixed point $x \in R$. It is hoped that the effect of the slight deviation from F on the underlying functional is small relative to ε. The influence function (IF) of a statistical functional M at a given point $x \in R$ for a given F, defined as (see Hampel et al. (1996))

$$
IF(x; M(F)) = \lim_{x \to 0^+} (M(F(\varepsilon, \delta_x)) - M(F))/\varepsilon. \tag{3.3.1}
$$

exactly measures the relative effect (influence) of an infinitesimal point mass contamination on M. It is desirable that this relative influence $IF(x; M(F))$ be bounded. This indeed is the the case for S_n^{RC} and Q (see, e.g. Rousseeuw and Croux (1993)), but not for the standard deviation functional with $(x^2-1)/2$ as its influence function for r.v. $X \sim N(0, 1)$.

Note that the integration interval in $S^2(F)$ ($S_w^2(F)$) are a functional of F. Hence an infinitesimal point-mass contamination affects this interval. Because of this, the derivation of the influence function of the scaled-deviation trimmed/winsorized scales becomes a little bit challenging. The strategy to attack the problem is"divide then conquer". One first works out the influence functions of L and U. Assume $F' = f$ exists at μ and $\mu \pm \sigma$ with $f(\mu)$ and $f(\mu+\sigma)+f(\mu-\sigma)$ positive, where μ and σ stand for $\mu(F)$ and $\sigma(F)$. As in chapter 2, we have preliminary results (2.3.2)-(2.3.6). Now assume that w_i (i=1, 2) is differentiable and f exists at $L(F)$ and $U(F)$. Write L and U for $L(F)$ and $U(F)$ respectively and define

$$\ell_{1i}(x) = \frac{1}{\delta_i(F)}\Big[(U - T_i)w_i(D(U,F))f(U)IF(x,U(F)) - (L - T_i)\,\mathrm{w}_i(D(L,F))f(L)IF(x,L(F))\Big] \tag{3.3.2}$$

$$\ell_{2i}(x) = \frac{1}{\delta_i(F)}\Big[\int_L^U (y - T_i)w_i^{(1)}(D(y,F))h(x,y)dF(y)\Big] \tag{3.3.3}$$

$$\ell_{3i}(x) = \frac{1}{\delta_i(F)}\Big[\mathbf{I}(x \in [L,\, U])(x - T_i)w_i(D(x,F))\Big] \tag{3.3.4}$$

One can derive the influence function of the scaled deviation trimmed mean $T(F)$ as follows. The following result is in our location chapter.

Corollary 3.3.1. *Assume that $F' = f$ exists at μ, $\mu \pm \sigma$, $L(F)$ and $U(F)$ with $f(\mu)$ and $f(\mu+\sigma)+f(\mu-\sigma)$ positive and is continuous in a small neighborhood of $L(F)$ and $U(F)$, and that $w_i(\cdot)$, $(i = 1, 2.)$ are continuously differentiable. Then for a given $0 < \beta < \infty$*

$$IF(x; T_i(F)) = \ell_{1i}(x) + \ell_{2i}(x) + \ell_{3i}(x). \tag{3.3.5}$$

Furthermore, if F is symmetric about the origin and w is a non-zero constant, one has

$$IF(x; T(F)) = \frac{x\mathbf{I}(x \in [-\beta\sigma, \beta\sigma])}{2F(\beta\sigma) - 1} + \frac{\beta\sigma f(\beta\sigma)\,\mathrm{sign}(x)}{f(0)(2F(\beta\sigma) - 1)}. \tag{3.3.6}$$

In order to express the influence function of $S(F)$, we need to borrow the following

notations and write

$$\tau_1(x) = \frac{1}{\delta_2}\Big[(U^2 - s)w_2(D(U,F))f(U)IF(x,U(F)) - (L^2 - s)w_2(D(L,F))f(L)IF(x,L(F))\Big] \tag{3.3.7}$$

$$\tau_2(x) = \frac{1}{\delta_2}\Big[\int_L^U (y^2 - s)w_2^{(1)}(D(y,F))h(x,y)dF(y)\Big] \tag{3.3.8}$$

$$\tau_3(x) = \frac{1}{\delta_2}\Big[\mathbf{I}(x \in [L,U])(x^2 - s)w_2(D(x,F))\Big]. \tag{3.3.9}$$

It is ready seen that the influence function of $\lambda(F)$ and $\lambda_w(F)$ are given by

$$IF(x,\lambda(F)) = 2(T_1(F) - T_2(F))IF(x,T_1(F)) - 2T_1(F)IF(x,T_2(F));$$
$$IF(x,\lambda_w(F)) = 2(T_{w1}(F) - T_{w2}(F))IF(x,T_{w1}(F)) - 2T_{w1}(F)IF(x,T_{w2}(F)).$$

The influence function of the scaled deviation trimmed variance $S^2(F)$ is given by the following theorem.

Theorem 3.3.2. *Assume that $F' = f$ exists at μ, $\mu \pm \sigma$, $L(F)$ and $U(F)$ with $f(\mu)$ and $f(\mu+\sigma) + f(\mu-\sigma)$ positive and is continuous in a small neighborhood of $L(F)$ and $U(F)$, and that $w_i(\cdot)$, $(i = 1, 2.)$ are continuously differentiable. Then for a given $0 < \beta < \infty$*

$$IF(x; S^2(F)) = c_t\big[IF(x; s(F)) + IF(x,\lambda(F))\big] \tag{3.3.10}$$

with

$$IF(x; s(F)) = \tau_1(x) + \tau_2(x) + \tau_3(x) \tag{3.3.11}$$

The proof of theorem 3.3.2 is given in section 5.2, chapter 5.

Under the condition of Theorem 3.3.2, $IF(x; S^2(F))$ clearly is bounded and consequently $S^2(F)$ is locally robust. Use the chain rule, one easily knows the influence function of the scaled deviation trimmed standard deviation, i.e. $IF(x; S(F)) = IF(x; S^2(F))/(2\sqrt{S^2(F)})$. For symmetric F and $w = c \neq 0$, the influence function simplifies substantially.

Corollary 3.3.3. *Let f be symmetric about the origin and w_i (i=1, 2) are nonzero constants. Under the conditions of Theorem 3.3.2, we have*

$$IF(x; S^2(F)) = c_t\Big[\frac{(x^2 - s)\mathbf{I}(x \in [-\beta\sigma, \beta\sigma])}{2F(\beta\sigma) - 1} + \frac{((\beta\sigma)^2 - s)f(\beta\sigma)\beta\,\mathrm{sign}(|x| - \sigma)}{2f(\sigma)(2F(\beta\sigma) - 1)}\Big]. \tag{3.3.12}$$

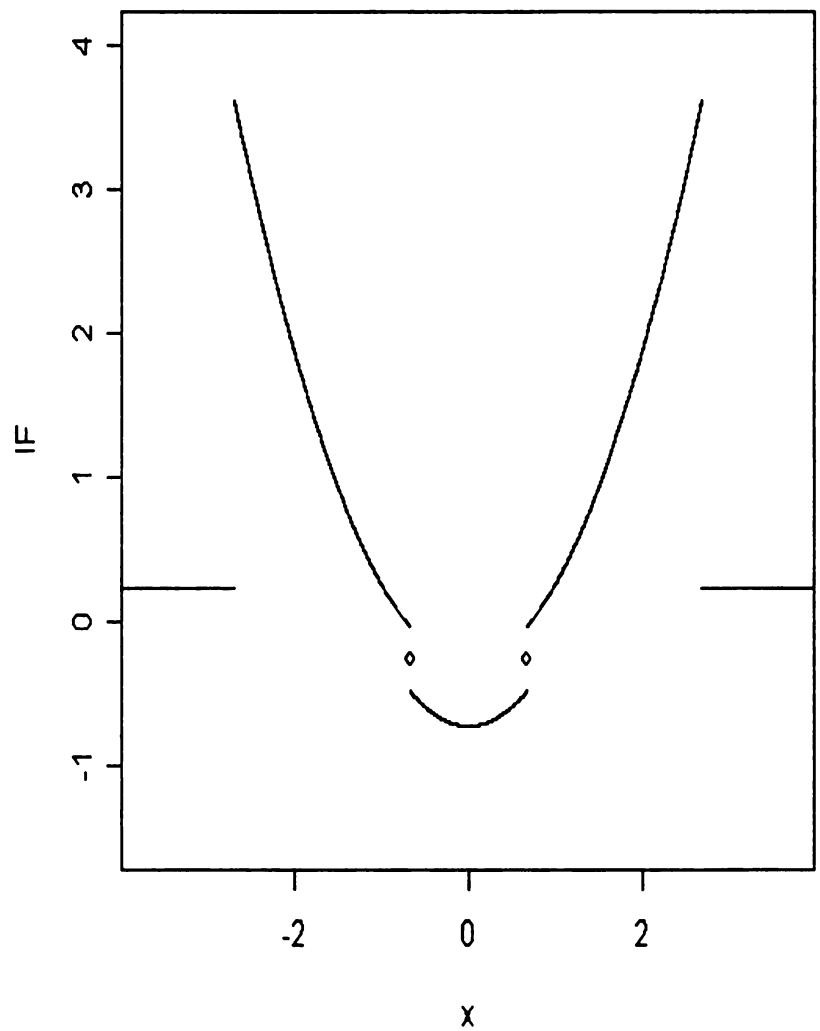

Figure 3.1. Influence functions of S for $N(0,1)$ with a constant weight and $\beta = 3$.

The proof of this corollary is omitted. A graph of this influence function $IF(x; S(F))$ is given in figure 3.1. Obviously, it is bounded.

To work out the influence function for $S_w^2(F)$ (or $S_w(F)$), we define

$$\ell_{w1i}(x) = \frac{1}{\delta_{wi}} \int [y\mathbf{I}(L \le y \le U) + L\mathbf{I}(y < L) + U\mathbf{I}(y > U) - T_{wi}]\, \mathrm{w}_i^{(1)}(D(y,F))h(x,y)dF(y) \quad (3.3.13)$$

$$\ell_{w2i}(x) = \frac{1}{\delta_{wi}} \int [IF(x,L)\mathbf{I}(y < L) + IF(x,U)\mathbf{I}(y > U)]\, \mathrm{w}_i(D(y,F))dF(y) \quad (3.3.14)$$

$$\ell_{w3i}(x) = \frac{1}{\delta_{wi}} [x\mathbf{I}(L \le x \le U) + L\mathbf{I}(x < L) + U\mathbf{I}(x > U) - T_{wi}]\, \mathrm{w}_i(D(x,F)). \quad (3.3.15)$$

and

$$\tau_{w1}(x) = \frac{1}{\delta_{w2}} \int [y^2\mathbf{I}(L \le y \le U) + L^2\mathbf{I}(y < L) + U^2\mathbf{I}(y > U) - s_w]\, \mathrm{w}_2^{(1)}(D(y,F))h(x,y)dF(y) \quad (3.3.16)$$

$$\tau_{w2}(x) = \frac{2}{\delta_{w2}} \int [IF(x,L)L\mathbf{I}(y < L) + IF(x,U)U\mathbf{I}(y > U)]\, \mathrm{w}_2(D(y,F))dF(y) \quad (3.3.17)$$

$$\tau_{w3}(x) = \frac{1}{\delta_{w2}} [x^2\mathbf{I}(L \le x \le U) + L^2\mathbf{I}(x < L) + U^2\mathbf{I}(x > U) - s_w]\, \mathrm{w}_2(D(x,F)). \quad (3.3.18)$$

One then has the influence function of the scaled deviation winsorized mean $T_w(F)$ as follows. The following result is in our location chapter.

Corollary 3.3.4. *Assume that $F' = f$ exists at μ, $\mu \pm \sigma$, $L(F)$ and $U(F)$ with $f(\mu)$ and $f(\mu+\sigma) + f(\mu-\sigma)$ positive and is continuous in a small neighborhood of $L(F)$ and $U(F)$, and that $w_i(\cdot)$, $(i = 1, 2.)$ are continuously differentiable with $rw^{(1)}(r)$ being bounded for $r \in R$. Then for a given $0 < \beta < \infty$,*

$$IF(x; T_{wi}(F)) = \ell_{w1i}(x) + \ell_{w2i}(x) + \ell_{w3i}(x). \quad (3.3.19)$$

Furthermore, if F is symmetric about the origin and w is a non-zero constant, one has

$$IF(x; T_w(F)) = \frac{\mathrm{sign}(x)}{f(0)} F(-\beta\sigma) + x\mathbf{I}(-\beta\sigma \le x \le \beta\sigma) - \beta\sigma\mathbf{I}(x < -\beta\sigma) + \beta\sigma\mathbf{I}(x > \beta\sigma). \quad (3.3.20)$$

One can have the influence function of the scaled deviation Winsorized variance $S_w^2(F)$ as follows.

Theorem 3.3.5. *Assume that $F' = f$ exists at μ, $\mu \pm \sigma$, $L(F)$ and $U(F)$ with $f(\mu)$ and $f(\mu+\sigma)+f(\mu-\sigma)$ positive and is continuous in a small neighborhood of $L(F)$ and $U(F)$, and that $w_i(\cdot)$, $(i = 1, 2.)$ are continuously differentiable with $r^2 w^{(1)}(r)$ being bounded for $r \in R$. Then for a given $0 < \beta < \infty$*

$$IF(x; S_w^2(F)) = c_w\big[IF(x; s_w(F)) - IF(x, \lambda_w(F))\big] \tag{3.3.21}$$

with

$$IF(x; s_w(F)) = \tau_{w1}(x) + \tau_{w2}(x) + \tau_{w3}(x) \tag{3.3.22}$$

The proof is given in section 5.2, chapter 5.

Under the conditions of Theorem 3.3.5, $IF(x; S_w^2(F))$ is readily seen to be bounded and S_w thus is locally robust. For symmetric F and constant w, the influence function simplifies greatly.

Corollary 3.3.6. *Let f be symmetric about the origin and w a constant. Under the conditions of Theorem 3.3.5, we have*

$$\begin{aligned} IF(x; S_w^2(F)) =& c_w\Big[\frac{\beta^2\sigma sign(|x|-\sigma)}{f(\sigma)}F(-\beta\sigma) + x^2\mathbf{I}(-\beta\sigma \le x \le \beta\sigma) \\ &+ (\beta\sigma)^2(\mathbf{I}(x \le -\beta\sigma) + \mathbf{I}(x \ge \beta\sigma)) - s_w\Big] \end{aligned} \tag{3.3.23}$$

The boundedness of the influence function $IF(x; S_w(F))$ is very clear and also shown in figure 3.2.

Besides measuring local robustness measures, the influence functions also provides the form of asymptotic variance of corresponding estimator. It is useful for establishing the asymptotic theory for $S(F)$ $(S_w(F))$ in the next section.

3.4 Asymptotic representation and limiting Distribution

It is challenging to establish the limiting distribution of the scaled deviation trimmed standard deviation and winsorized standard deviation. Here we fulfill the task by combining

an empirical process theory argument with the influence function results obtained in the last section. Once asymptotic representations of the estimators are established, the limiting distributions easily follow.

Theorem 3.4.1. *Let $F' = f$ exist at μ and be continuous in small neighborhoods of $\mu \pm \sigma$, L and U with $f(\mu)$ and $f(\mu-\sigma)+f(\mu+\sigma)$ positive. Let $w^{(1)}$ be continuous on* R. *Then for $1 \le \beta < \infty$*

$$S^2(F_n) - S^2(F) = \frac{1}{n}\sum_{i=1}^{n} IF(X_i; S^2(F)) + o_p(\frac{1}{\sqrt{n}})$$

where $IF(x; S^2(F))$ is given by Theorem 3.3.2. Consequently

$$\sqrt{n}(S^2(F_n) - S^2(F)) \to N(0, \tilde{\sigma}^2) \tag{3.4.1}$$

provided $\tilde{\sigma}^2 = E(IF(X_1, S^2(F)))^2 < +\infty$.

The proof is given in section 5.2, chapter 5.

Suppose one want to estimate η in distribution $F_0(x/\eta)$ by using the scaled-deviation trimmed standard deviation to estimate it. Since $S^2(F) = \eta^2$, invoking delta method, one has

$$\sqrt{n}(S(F_n) - \eta) \to N(0, \frac{1}{4\eta^2}\tilde{\sigma}^2) \tag{3.4.2}$$

So, if we set $\eta = 1$, then the asymptotic variance for $S(F_n)$ is one quarter of $\tilde{\sigma}^2$.

The distribution F is usually assumed to be symmetric about a point θ. By affine equivariance, we can let $\theta = 0$. In this case with a (non-zero) constant weight $\tilde{\sigma}^2$ takes a much simpler form and will be evaluated at a number of distributions in the next section.

Corollary 3.4.2. *Let f be symmetric about the origin and w a constant. Under the conditions of Theorem 3.4.1, we have*

$$\tilde{\sigma}^2 = c_t^2\Big[\int_{-\beta\sigma}^{\beta\sigma}(x^2-s)^2dF(x) + \frac{2((\beta\sigma)^2-s)f(\beta\sigma))\beta}{f(\sigma)}\Big(\int_{-\beta\sigma}^{-\sigma}(x^2-s)dF(x)$$
$$-\int_{-\sigma}^{0}(x^2-s)dF(x)\Big) + \Big(\frac{((\beta\sigma)^2-s)f(\beta\sigma))\beta}{2f(\sigma)}\Big)^2\Big]/(2F(\beta\sigma)-1)^2$$

where $c_t = 1/s(F_0)$.

The proof of this corollary is omitted.

For S_w^2, we can establish results similar to Theorem 3.4.1 and Corollary 3.4.2.

Theorem 3.4.3. *Let $F = f'$ exist at μ and be continuous in small neighborhoods of $\mu \pm \sigma$, L and U with $f(\mu)$ and $f(\mu-\sigma)+f(\mu+\sigma)$ positive. Let w be continuously differentiable with $r^2 w^{(1)}(r)$ being bounded for $r \in R$. Then for $0 < \beta < \infty$*

$$S_w^2(F_n) - S_w^2(F) = \frac{1}{n}\sum_{i=1}^{n} IF(X_i; S_w^2(F)) + o_p(\frac{1}{\sqrt{n}})$$

where $IF(x; S_w^2(F))$ is given by Theorem 3.3.5. Consequently

$$\sqrt{n}(S_w^2(F_n) - S_w^2(F)) \to N(0, \tilde{\sigma}_w^2) \tag{3.4.3}$$

provided $\tilde{\sigma}_w^2 = E(IF(X_1, S_w^2(F)))^2 < +\infty$.

The proof is given in section 5.2, chapter 5.

Similar to trimmed case, suppose one want to estimate η in distribution $F_0(x/\eta)$ by using the scaled-deviation winsorized standard deviation to estimate it. One has

$$\sqrt{n}(S_w(F_n) - \eta) \to N(0, \frac{1}{4\eta^2}\tilde{\sigma}_w^2) \tag{3.4.4}$$

So, if we set $\eta = 1$, then the asymptotic variance for $S_w(F_n)$ is one quarter of $\tilde{\sigma}_w^2$.

Let F be symmetric about a point θ. By affine equivariance, one can let $\theta = 0$. In this case with a (non-zero) constant weight $\tilde{\sigma}^2$ takes a relatively simple and explicit form and will be evaluated at a number of distributions in the next section.

Corollary 3.4.4. *Let f be symmetric about the origin and w a constant. Under the conditions of Theorem 3.4.3, we have*

$$\begin{aligned}\tilde{\sigma_w}^2 =& c_w^2\Big\{\Big(\frac{\beta^2\sigma F(-\beta\sigma)}{f(\sigma)}\Big)^2 + \int x^4 \mathbf{I}(-\beta\sigma \le x \le \beta\sigma) \\ &+ (\beta\sigma)^4(\mathbf{I}(x \le -\beta\sigma) + \mathbf{I}(x \ge \beta\sigma))dF(x) + s_w^2 \\ &+ \frac{4\beta^4\sigma^3(F(-\beta\sigma))^2}{f(\sigma)} + \frac{4\beta^2\sigma F(-\beta\sigma)}{f(\sigma)}\Big[\int_{-\beta\sigma}^{-\sigma} x^2 dF(x) - \int_{-\sigma}^{0} x^2 dF(x)\Big] \\ &- \frac{2s_w\beta^2\sigma F(-\beta\sigma)(4F(-\beta\sigma)-1)}{f(\sigma)} - 4s_w\Big[\int_0^{\beta\sigma} x^2 dF(x) + (\beta\sigma)^2 F(-\beta\sigma)\Big]\Big\}\end{aligned}$$

where $c_w = 1/s_w(F_0)$.

The proof of this corollary is omitted.

3.5 Comparison

In this section, we compare scaled-deviation trimmed and winsorized standard deviation with median absolute deviation (MAD), ordinary standard deviation (SD), and the scale estimators S_n^{RC} and Q_n proposed by Rousseeuw and Croux. We will discuss robustness through breakdown point, influence function and gross error sensitivity. We also will consider efficiency by asymptotic relative efficiency in section 3.5.3 and finite sample efficiency in section 3.5.4. For simplicity, let $w_i = 1$, $i = 1, 2$. Suppose $X \sim F(x) = F_0\big((x-\theta)/\eta\big)$ and F_0 is the underlying model distribution, let $c_t = 1/(s(F_0) - \lambda(F_0))$ and $c_w = 1/(s_w(F_0) - \lambda_w(F_0))$.

3.5.1 Breakdown Point

The finite sample breakdown point measures the global robustness of an estimator. Roughly speaking, the breakdown point of a scale estimator is the minimum fraction of "bad" data points that can render the estimator useless (0 or ∞). Precisely, the finite sample breakdown point of a scale estimator S_n at sample $X^n = \{X_1, \ldots, X_n\}$ in R is defined as

$$\varepsilon_n^*(S_n, X^n) = \min\{\varepsilon_n^+(S_n, X^n), \varepsilon_n^-(S_n, X^n)\}$$

where

$$\varepsilon_n^+(S_n, X^n) = \min\Big\{\frac{m}{n} : \sup_{X_m^n} S_n(X_m^n) = \infty\Big\}$$

and

$$\varepsilon_n^-(S_n, X^n) = \min\Big\{\frac{m}{n} : \sup_{X_m^n} S_n(X_m^n) = 0\Big\}$$

where X_m^n denotes a contaminated data set resulting from replacing m original points of X^n with arbitrary m points. The quantities ε_n^+ and ε_n^- are called the explosion and the implosion breakdown points. It is well known that the standard deviation has the lowest breakdown point $1/n$.

Since $\mu(F_n)$ and $\sigma(F_n)$ have the highest possible breakdown point, it is easy to see that $S(F_n)(S(F_{wn}))$ has the same breakdown point as $\sigma(F_n)$. So $S(F_n)(S(F_{wn}))$, together with S_n^{RC}, Q_n and MAD_n, achieves the best possible breakdown point of any translation equivalent scale estimators (see, e.g. Lopuhaä and Rousseuw (1991)).

Theorem 3.5.1. *At any sample X^n in which no two points coincide and for any $1 \leq \beta < \infty$, we have*

$$\varepsilon_n^+(S(F_n), X^n) = [(n+1)/2]/n, \quad \textit{and} \quad \varepsilon_n^-(S(F_n), X^n) = [n/2]/n$$

$$\varepsilon_n^+(S_w(F_n), X^n) = [(n+1)/2]/n, \quad \textit{and} \quad \varepsilon_n^-(S_w(F_n), X^n) = [n/2]/n$$

The breakdown point of the scale estimators $S(F_n)$ and $S_w(F_n)$ thus is given by

$$\varepsilon_n^*(S(F), X^n) = [n/2]/n, \quad \varepsilon_n^*(S_w(F), X^n) = [n/2]/n,$$

which is the highest possible value for any affine equivalent scale estimator. (The proof of this theorem is omitted.)

3.5.2 Influence Function and Gross Error Sensitivity

The breakdown point measures only the global robustness while the influence function can capture the local robustness of an estimator. The two together can provide a more complete picture of robustness. We now look at the influence functions of the trimmed and winsorized variances.

The boundedness of its influence function is the fundamental concern for a functional being locally robust. The mean functional and standard deviation functional have an unbounded influence function. The MAD and $S^{RC}(F)$, $Q(F)$ are known to have bounded influence functions; see, e.g., Serfling (1980) for MAD, Rousseeuw and Croux (1993) for $S^{RC}(F)$ and $Q(F)$.

In the light of Theorems 3.3.2 and 3.3.5, S and S_w have bounded influence functions for suitable w and β. Figure 3.3–3.6, which plots their influence functions at normal, Cauchy, exponential and contaminated normal models with $\beta = 4.5$, confirms this. For convenience, we also set $w_i = c \neq 0$ in S and S_w. Since a pure normal model is rare in practice, we thus consider light-tail, heavy-tail, and skewed models. They become different (but all are still bounded). In terms of the bounded influence function criterion, we conclude that all the trimmed and winsorized standard deviation are *equally* robust (locally).

Besides boundedness, one can also look at the magnitude of the supremum of $|IF(x; S(F))|$, the so-called the *gross error sensitivity* (GES) of S at F (Hampel et al.

(1986))

$$GES(S(F)) = \sup_{x \in R} |IF(x; S(F))|, \tag{3.5.1}$$

which measures the worst case effect on S of an infinitesimal point mass contamination. Generally speaking, a smaller GES is more desirable. For scaled-deviation trimmed standard deviation S and S_w, it is readily seen that their GES depends on the values of β and the weight function w. As a possible way to make a comparison, we again set ($w_i = c \neq 0$) in the following discussion.

Table 3.1. Gross Error Sensitivity

	SD	Q	S^{RC}	S ($\beta = 4.5$)	S_w ($\beta = 4.5$)	MAD
$N(0,1)$	$+\infty$	2.069	1.625	4.3503	4.1528	1.167
Cauchy		2.2214	1.8961	5.4185	2.5174	1.5708
Exponential	$+\infty$	2.3173	1.8447	5.2883	3.6954	1.8587
$0.9N(0,1) + 0.1N(4,9)$	$+\infty$	2.1007	1.6274	4.7827	4.6295	1.2457

The influence functions of MAD, Q, S^{RC}, S ($\beta < \infty$), are all bounded and has finite Gross Error sensitivity.

From those figure 3.3–3.6, and table 3.1 presented, it is easily seen that GES of scaled-deviation trimmed and winsorized scale which usually attained at the cutting end points L and U for most of β values is slightly larger than the other four estimators.

At normal model, S_w has the smaller GES than Q for $0 < \beta \leq 3.13$ and has the smaller GES than S^{RC} for $0 < \beta \leq 2.62$. But S alway has larger GES than Q, S^{RC} and MAD.

At Cauchy model, S_w has the smaller GES than Q for $0 < \beta \leq 3.66$ and has the smaller GES than S^{RC} for $0 < \beta \leq 2.67$. But S alway has larger GES than Q, S^{RC} and MAD.

At exponential model, S_w has the smaller GES than Q for $0 < \beta \leq 3.12$ and has the smaller GES than S^{RC} for $1.30 \leq \beta \leq 1.57$. But S alway has larger GES than Q, S^{RC} and MAD.

Scaled-deviaton winsorized standard deviation is overwhelmingly more robust than trimmed standard deviation in terms of GES as shown in figure 3.10–4.2.

Since different β can give us difference choices, so somehow, trimmed and winsorized standard deviation are more flexible than the other three.

The picture of GES versus the steps β of different distributions are presented by figure ?? . For normal distribution, GES increases dramatically when the β increases. But for Cauchy distribution, GES increases slowly; At the same time, since the influence function of the classical standard deviation for normal distribution is given by $IF(x, SD(\Phi)) = (x^2 - 1)/2$. When β takes large values, $IF(x, S(\Phi))$, $IF(x, S_w(\Phi))$ and $IF(x, SD(\Phi))$ will be very close. It implies the high efficiency of S (S_w) at normal model when β is large.

3.5.3 Large sample relative efficiency

Now we evaluate the performance of the trimmed (and winsorized) standard deviations in terms of their efficiency behavior (relative to the sample SD). First we examine the asymptotic relative efficiency (ARE). Table 3.2 lists the ARE results of S and S_w at a number of light- and heavy-tailed *symmetric* distributions with different β values. Here we again set w = $c > 0$. The table reveals that (i) at normal model with *large* β both S and S_w can be highly efficient relative to the SD; (ii) their efficiency increases as the tail of the distribution becomes heavier and exceeds 100% at heavy tailed distributions for large β; and (iii) S_w is more efficient than S for small β or normal distribution.

For Cauchy distribution, when $0.954 \leq \beta \leq 1.015$, S is more efficient than S_w.

The picture of ARE versus the steps β of different distributions are presented by figure 3.7–3.9. For normal distribution, ARE increases gradually when the β increases. But for Cauchy distribution, ARE decreases with the increasing of β. When β goes to infinity, ARE tends to zero. For both GES and ARE, S is unstable at the neighborhood of $\beta = 1$ mainly because the denominator of $S(F_n)$ might be 0. However, S_w is pretty stable when β changes. Table 3.3 gives the asymptotic relative efficiencies (ARE) of various high breakdown scale estimators with respect to the sample standard deviation for various distributions. Examining Table 3.3 reveals that in terms of efficiency: (i) Overall S_w performs *better* than S. (ii) S and S_w perform better than other robust estimators S^{RC}, Q and MAD at light tailed distribution while have satisfactory efficiency for heavy tailed distributions.

The behaviors of ARE of scaled-deviation trimmed standard deviation for different values of β to standard deviation at normal, to theoretical lower bound (inverse of Fisher informa-

Table 3.2. AREs of S and S_w relative to the standard deviation

β		1	2	3	4	5	6	7
$N(0,1)$	S	0.3068	0.2736	0.4509	0.7438	0.9403	0.9921	0.9994
	S_w	0.3717	0.5383	0.8261	0.9713	0.9973	0.9998	1.0000
$LG(0,1)$	S	0.4708	0.4132	0.5818	0.7822	0.9363	1.0090	1.0263
	S_w	0.5535	0.7368	0.9892	1.1069	1.1045	1.0691	1.0390
$DE(0,1)$	S	0.5647	0.4482	0.5412	0.6605	0.7879	0.9032	0.9883
	S_w	0.6191	0.7413	0.9318	1.0915	1.1758	1.1894	1.1638
$E(0,1)$	S	0.7315	0.7610	0.8303	0.9156	1.0067	1.0941	1.1679
	S_w	0.9163	1.1138	1.2903	1.4456	1.5420	1.5653	1.5291
t_5	S	0.7315	0.7610	0.8303	0.9156	1.0067	1.0941	1.1679
	S_w	0.9163	1.1138	1.2903	1.4456	1.5420	1.5653	1.5291
Cauchy*(0, 1)	S	0.8768	0.7053	0.7422	0.7542	0.7485	0.7334	0.7134
	S_w	0.8579	0.8900	0.9080	0.8969	0.8716	0.8403	0.8071

* compared with the inverse of fisher information 2.

Table 3.3. ARE's with respect to SD

	Q	S^{RC}	S ($\beta = 4.5$)	S ($\beta = 7$)	S_w ($\beta = 4.5$)	S_w ($\beta = 7$)	MAD
$N(0,1)$	0.8227	0.5823	0.8658	0.9994	0.9906	1.0000	0.3675
$LG(0,1)$	1.0210	0.8918	0.8690	1.0263	1.1144	1.0390	0.5431
$DE(0,1)$	1.5952	0.9206	0.7244	0.9883	1.1439	1.1638	0.6006
$E(0,1)$	1.4897	1.0591	0.9609	1.1679	1.5028	1.5291	0.9367
t_5	1.4066	2.0026	1.9388	2.0732	2.4113	2.0145	1.3332
Cauchy*	0.9784	0.9497	0.7530	0.7134	0.8854	0.8071	0.8106

* compared with the inverse of fisher information 2.

Table 3.4. β Values for S having better ARE than Other Scales

	$S \succeq Q$	$S \succeq S^{RC}$	$S \succeq MAD$
$N(0,1)$	$[4.31, \infty)$	$[3.47, \infty)$	$(0, \infty)$
$LG(0,1)$	$[6.38, 8.19)$	$[4.66, \infty)$	$[0, \infty)$
$DE(0,1)$	NA	$[6.18, \infty)$	$[0, \infty)$
$E(0,1)$	$[5.59, 18.17)$	NA	$[4.24, \infty)$
t_5	$[2.99, 18.62)$	$[4.81, 7.89)$	$[2.82, 22.11)$
Cauchy*	NA	NA	$[0.82, 1.06]$

* compared with the inverse of fisher information 2.

Table 3.5. β Values for S_w having better ARE than Other Scales

	$S_w \succeq Q$	$S_w \succeq S^{RC}$	$S_w \succeq MAD$	$S_w \succeq S$
$N(0,1)$	$[2.99, \infty)$	$[2.16, \infty)$	$(0, \infty)$	$(0, \infty)$
$LG(0,1)$	$[3.17, 7.93)$	$[2.58, \infty)$	$(0, \infty)$	$(0, \infty)$
$DE(0,1)$	NA	$[2.94, \infty)$	$(0, \infty)$	$(0, \infty) \setminus [11, 18]$
$E(0,1)$	$[1.68, 15.84)$	$[4.38, 7.64)$	$(0, \infty)$	$(0, 10.97)$
t_5	$[1.26, 15.01)$	$[2.41, 7.00)$	$(0, 17.78)$	$(0, 6.46)$
Cauchy*	NA	NA	$(0, 6.89)$	$(0, \infty) \setminus [0.96, 1.01]$

* compared with the inverse of fisher information 2.

tion which is equal to 2 in this case) at Cauchy, to standard deviation at exponential are shown in figure 3.7–3.9. For normal model, it is natural that $ARE_{S_w,SD}$ and $ARE_{S,SD}$ ($\beta > 1$) increase with β and less than 1. For Cauchy model, note that ARE_S and ARE_{S_w} roughly decrease to 0 with β, while its standard deviation does not even exist. For exponential distribution, $ARE_{S,SD}$ is increasing while $GES(S)$ decreases when β increases while $ARE_{S_w,SD}$ is J-shaped.

In practice, the sample size is often small and the distributions are not pure models. Scaled-deviation trimmed and winsorized scales perform quite well in the asymptotic sense when F's are from a pure model. This raises the concern as to whether these results are valid at finite sample practice and for F's with slight departure from a perfect model. We answer this question in the next section via finite sample simulation studies.

3.5.4 Finite sample relative efficiency

To check whether the estimator S and S_w are approximate unbiased for finite samples, we performed a modest simulation study. In Table 3.6, we calculated the average scale estimate on 10,000 batches of normal, Cauchy, and exponential observations. We see that S_n (S_{wn}) behaves better than other scales at normal model and we carried out a simulation to verify the efficiency gain at finite samples. For each n in the table 3.6 , we computed the variance $var_m(S_n)$ of the scale estimator S_n over $m = 10,000$ samples. Table 3.6 lists the standardized variances

$$nVar_m(S_n)/(ave_m(S_n))^2 \tag{3.5.2}$$

where $ave_m(S_n)$ is the average estimated value which is listed in the left half of the tables. The results show that the asymptotic variance provides a good approximation for (not too small) finite samples, and that S_n and S_{wn} are more efficient than S_n^{RC}, Q_n and MAD_n at normal model even for small n.

It is clearly reviewed that S_n and S_{nw} are considerably more efficient than S_n^{RC}, Q_n and MAD_n even for small n.

Table 3.6. Standard variance of MAD_n, S_n^{RC}, Q_n, S_n, S_{wn} and SD_n at normal model

	Average value					
n	Q_n	S_n^{RC}	S_n	S_{nw}	MAD_n	SD_n
10	0.899	1.286	0.913	0.905	0.905	0.969
20	0.957	1.166	0.961	0.959	0.959	0.987
40	0.978	1.087	0.979	0.979	0.980	0.992
60	0.986	1.059	0.986	0.984	0.987	0.994
80	0.990	1.046	0.990	0.990	0.992	0.997
100	0.993	1.037	0.993	0.991	0.993	0.998
200	0.997	1.020	0.997	0.996	0.997	0.999

	Standard variance					
n	Q_n	S_n^{RC}	S_n	S_{nw}	MAD_n	SD_n
10	0.849	0.920	0.617	0.533	1.241	0.518
20	0.749	0.877	0.535	0.507	1.293	0.504
40	0.700	0.850	0.521	0.514	1.320	0.514
60	0.658	0.849	0.497	0.494	1.337	0.494
80	0.649	0.845	0.509	0.506	1.344	0.506
100	0.647	0.855	0.501	0.500	1.352	0.500
200	0.627	0.859	0.494	0.493	1.380	0.493

We also conducted a similar study on Cauchy, exponential distribution. The results at these distributions confirm the asymptotic results presented in last section so they are not listed here. It is shown that S_n and S_{nw} can achieve satisfactory efficiency compared with other robust estimators.

We now conduct Monte Carlo studies to investigate the efficiency behavior of S and S_w at finite samples for normal and contaminated normal models. Here $\eta = 1$ is regarded as the target parameter to be estimated. For an estimator S the empirical mean squared error (EMSE) is: EMSE $= \frac{1}{m}\sum_{j=1}^{m} |S_j - \eta|^2$, where m is the number of samples generated and T_j is the estimate based on the jth sample. The relative efficiency (RE) of S is then obtained by dividing the EMSE of the sample standard deviation by that of S. We generated $m = 50,000$ samples from $(1-\varepsilon)N(0,1)+\varepsilon N(1,0.1)$ and $(1-\varepsilon)N(0,1)+\varepsilon\delta_{\{0\}}$ with $\varepsilon = 0, .1$ and .2 for different sample sizes n. Some results are listed in Table 3.7 and Table 3.8 with $\beta = 7$.

Table 3.7. REs of various robust scales ($\beta = 7$ for scaled-deviation trimmed/winsorized scale) at $(1-\varepsilon)N(0,1) + \varepsilon N(1,0.1)$

	Q	S^{RC}	S	S_w	MAD	Q	S^{RC}	S	S_w	MAD
n			$\varepsilon = 10\%$					$\varepsilon = 20\%$		
20	0.305	0.634	0.972	0.993	0.414	0.322	0.615	0.958	0.985	0.417
40	0.444	0.606	0.986	0.992	0.367	0.527	0.586	0.973	0.980	0.369
60	0.552	0.570	0.985	0.988	0.323	0.650	0.564	0.977	0.982	0.318
80	0.617	0.559	0.992	0.993	0.290	0.732	0.533	0.981	0.982	0.281
100	0.666	0.524	0.991	0.991	0.256	0.791	0.527	0.978	0.978	0.261

Table 3.8. REs of various robust scales ($\beta = 7$ for scaled-deviation trimmed/winsorized scale) at $(1-\varepsilon)N(0,1) + \varepsilon\delta_{\{0\}}$

	Q	S^{RC}	S	S_w	MAD	Q	S^{RC}	S	S_w	MAD
n			$\varepsilon = 10\%$					$\varepsilon = 20\%$		
20	0.502	0.494	0.825	0.908	0.361	0.604	0.410	0.758	0.875	0.328
40	0.721	0.413	0.872	0.924	0.319	0.711	0.309	0.823	0.903	0.278
60	0.779	0.342	0.897	0.930	0.295	0.633	0.248	0.856	0.917	0.252
80	0.746	0.296	0.908	0.935	0.283	0.562	0.216	0.873	0.927	0.241
100	0.699	0.263	0.916	0.940	0.270	0.502	0.195	0.888	0.934	0.229

For small β, simulation results show that scaled-deviation winsorized scale is more efficient

than scaled-deviation trimmed scale at the situation "bad" points from the areas around the center. When β gets large, the difference becomes small and all most achieve the same efficiency when $\beta > 7$.

Our simulation results for $\varepsilon = 0.0$ (not listed in the table) confirms the validity of the asymptotic ones in Table 3.2 for $N(0,1)$. On the other hand, when there is just a 10% or 20% contamination in the distribution, all other estimators Q, S^{RC}, MAD become less efficient than S and S_w which are the most efficient, reflecting the robustness of these estimators. It is remarkable that S_w^β is overwhelmingly more efficient than all other estimators if we suitably choose β.

The other three types of robust estimators that are built on (recursive) medians only use 50% of information around the center, so it is very robust and efficient when contaminating points are from either end. But its strength is also its weakness, when contaminating points are close to the center, it will use the "bad" points and lose its efficiency. The two types of estimators S, S_w introduced in this chapter achieve satisfactory efficiency when contaminating points are from either end, although it's a little bit less efficient than other three(Q, S^{RC}, MAD). But when outliers/contaminating points are from the areas near the center, S and S_w are far more efficient than them.

3.6 Concluding remarks

Unlike the standard deviation, all the trimmed and winsorized standard deviation and robust scales discussed in the chapter have bounded influence functions for suitable distributions and weight functions and hence are locally robust. In terms of the global robustness, the scaled deviation trimmed and winsorized standard deviations S and S_w are exceptional in the sense that they can enjoy the best possible breakdown point robustness for any $\beta \geq 1$ with other robust scales.

Relative to the standard deviation, the scaled deviation trimmed S and winsorized standard deviations S_w are highly efficient for large β's at light-tailed symmetric distributions and much more efficient at heavy-tailed ones for small β. Three popular scale estimators Q, S^{RC}, MAD which are built on (recursive) medians are highly efficient when "bad" points

are from either end. However they lose the capability to tell the truth when contaminating points are presented around the center. At this time, S and S_w can make a difference.

S and S_w are more flexible than other robust scale estimators since β can take different values, which also raises a question on the choice of β value. In light of our simulation studies, a β value between 4 to 7 would be recommended so that S and S_w can be very efficient at both light- and heavy-tailed distributions. Instead of a fixed value one might also adopt an adaptive data-driven approach to determine an appropriate β value. For a given data set, one determines a value for β based on the heaviness of the tail. Generally speaking, a large value of β is selected for a light-tailed data set while a smaller value for a heavy-tailed one.

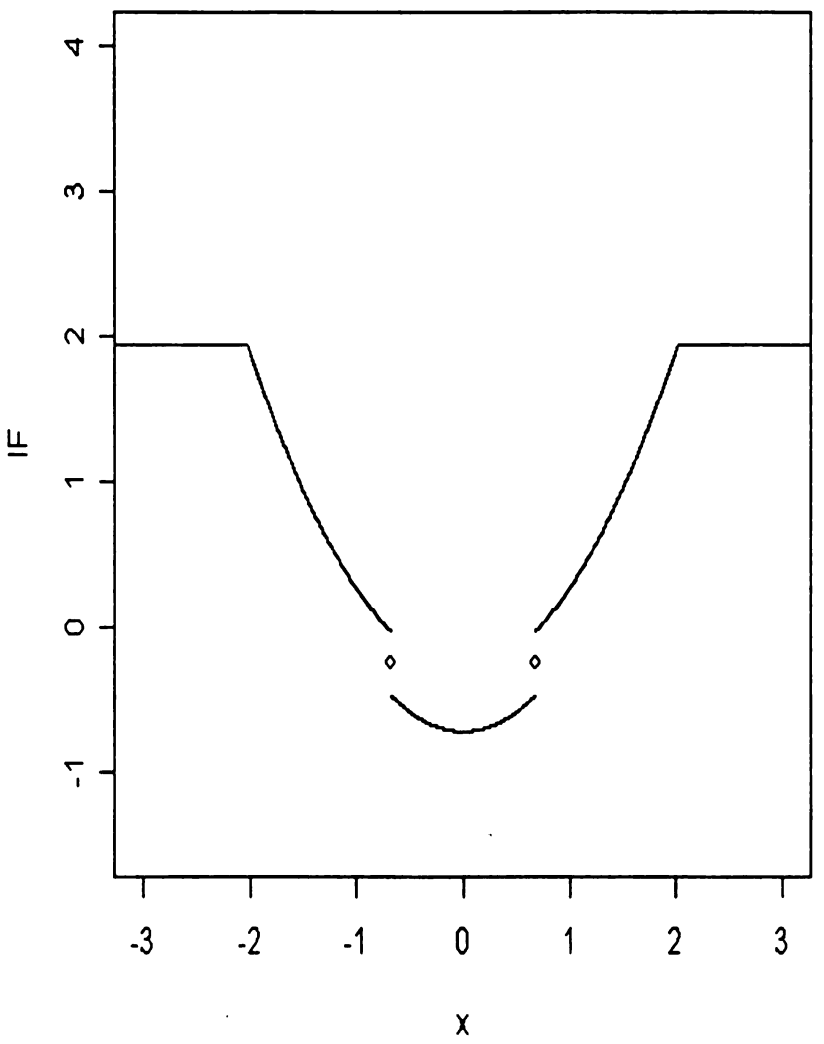

Figure 3.2. Influence functions of S_w for $N(0,1)$ with a constant weight and $\beta = 3$.

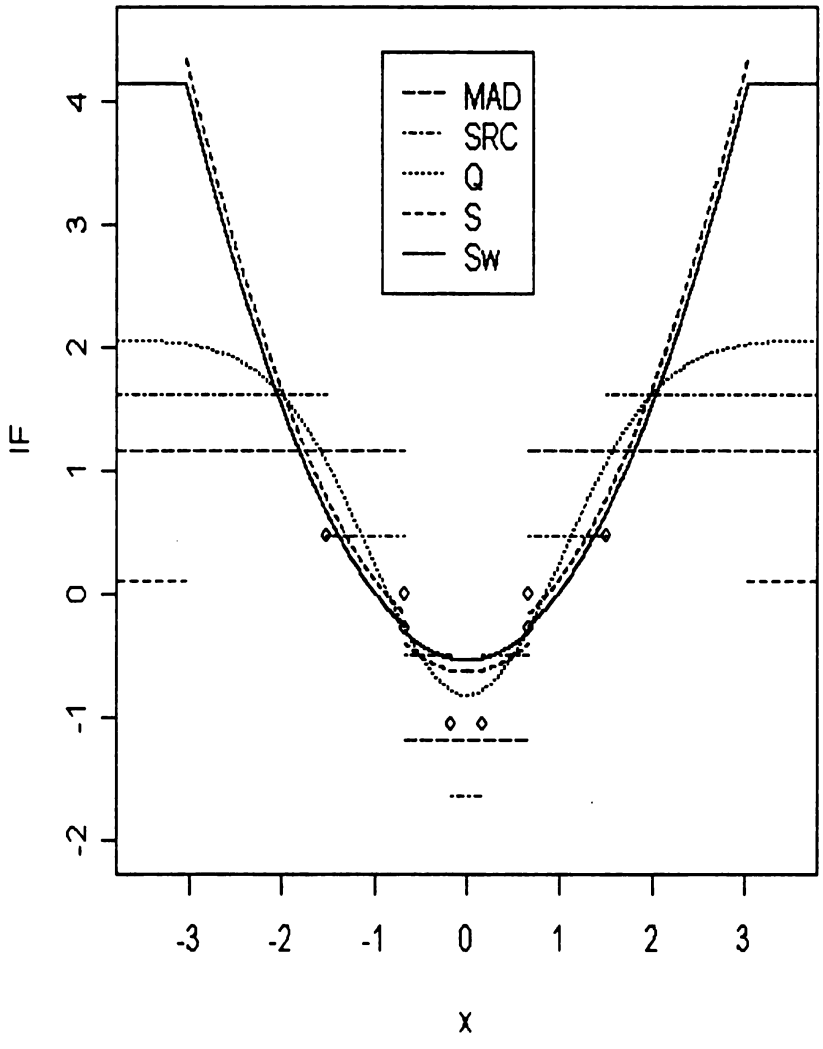

Figure 3.3. Influence functions of various scales for normal distribution. ($\beta = 4.5$ for S and S_w)

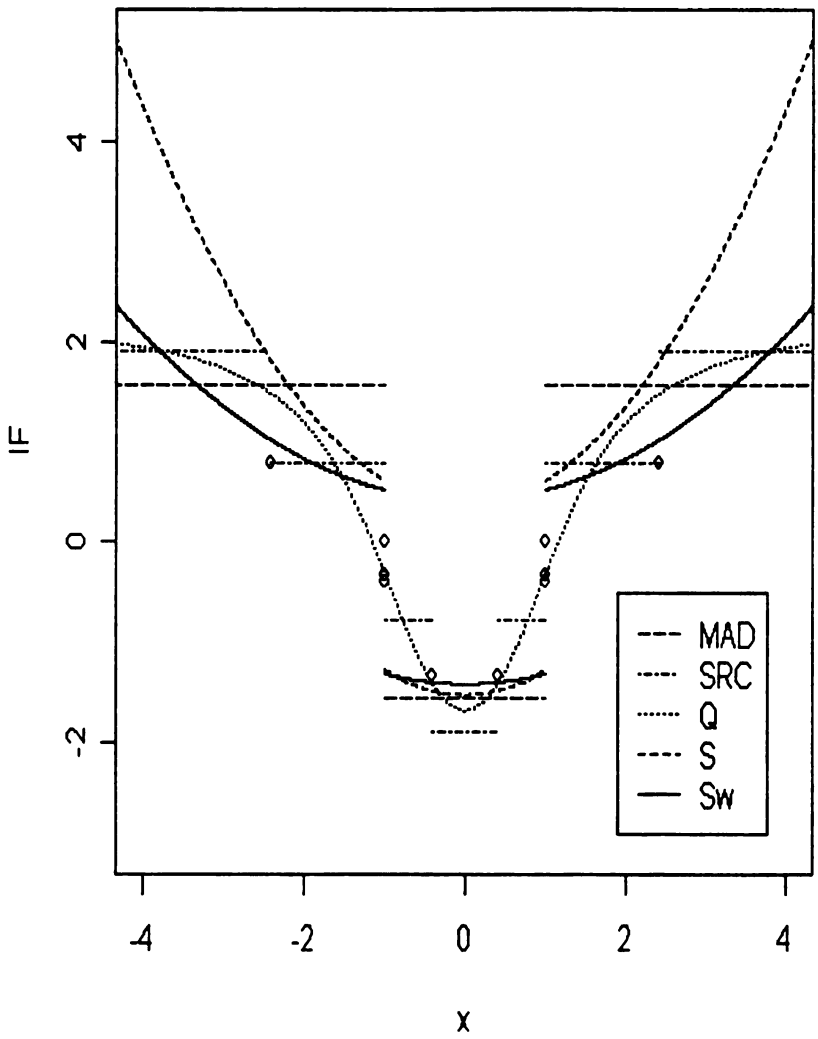

Figure 3.4. Influence functions of various scales for Cauchy distribution. ($\beta = 4.5$ for S and S_w)

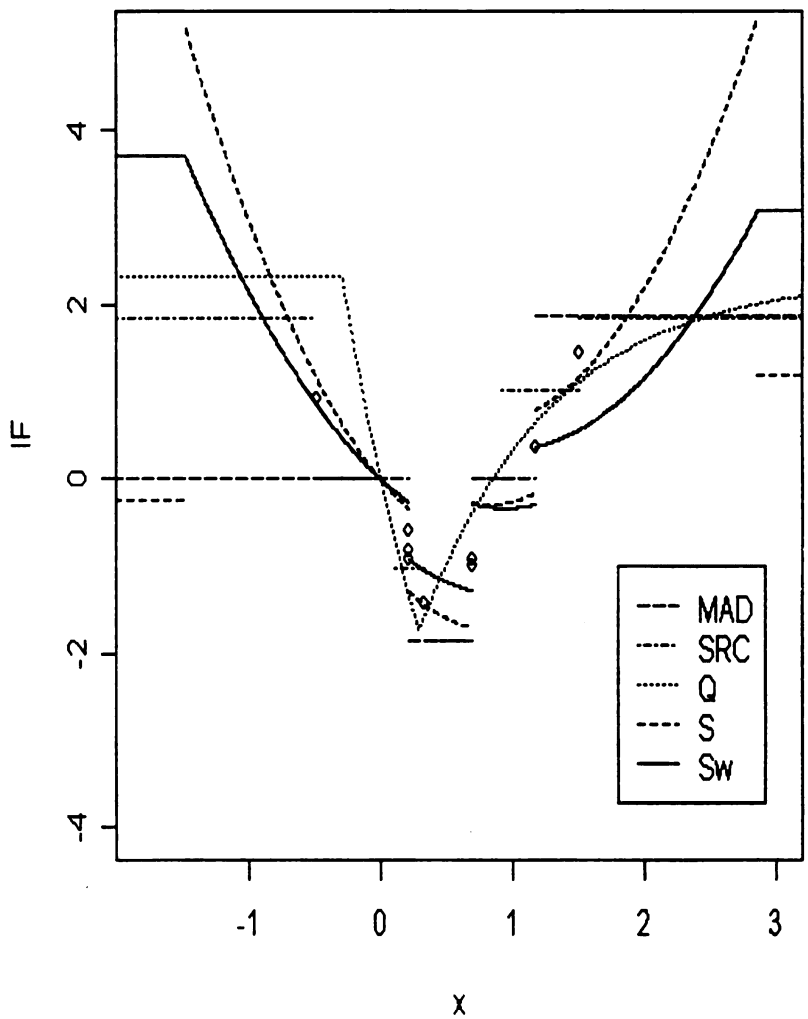

Figure 3.5. Influence functions of various scales for exponential distribution. ($\beta = 4.5$ for S and S_w)

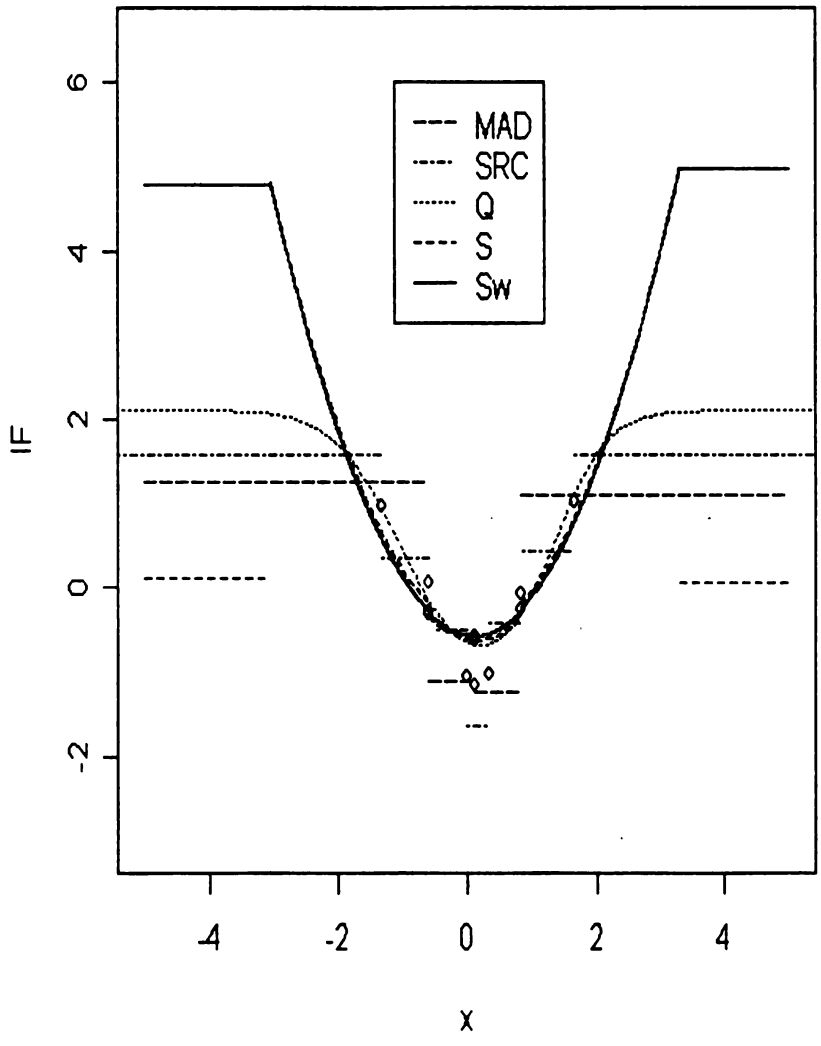

Figure 3.6. Influence functions of various scales for $0.9N(0,1) + 0.1N(1,0.1)$. ($\beta = 4.5$ for S and S_w)

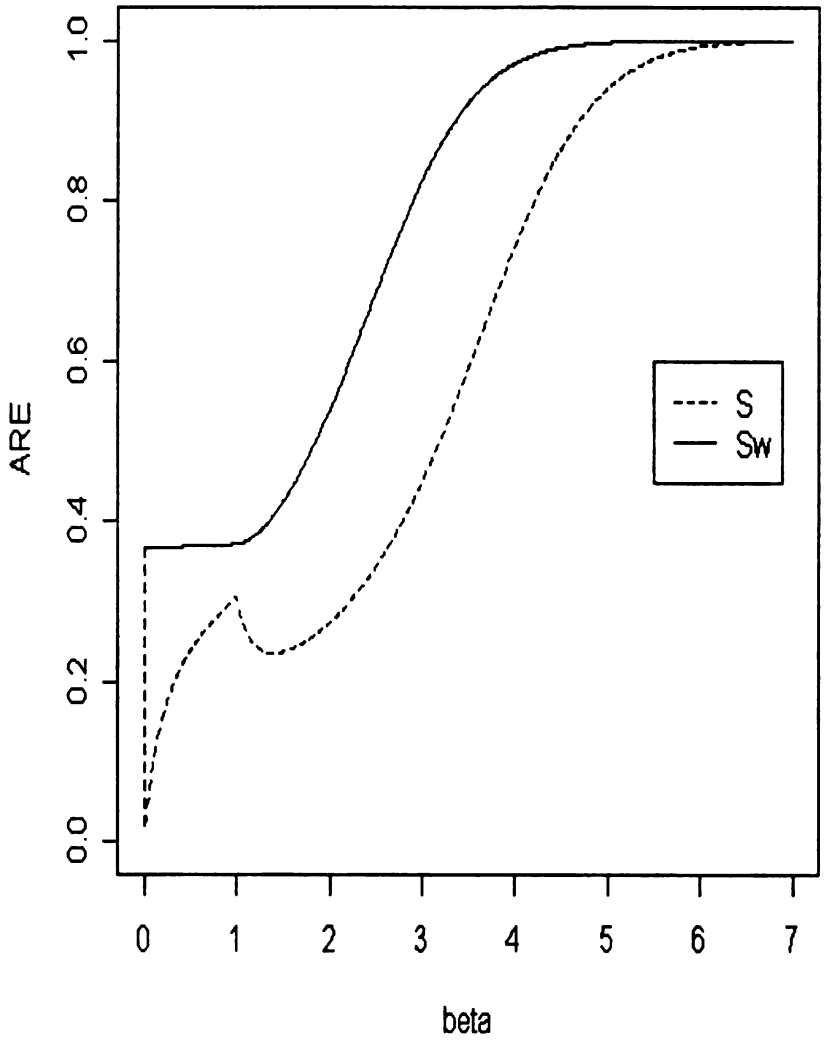

Figure 3.7. ARE of trimmed and winsorized standard deviations for normal distribution

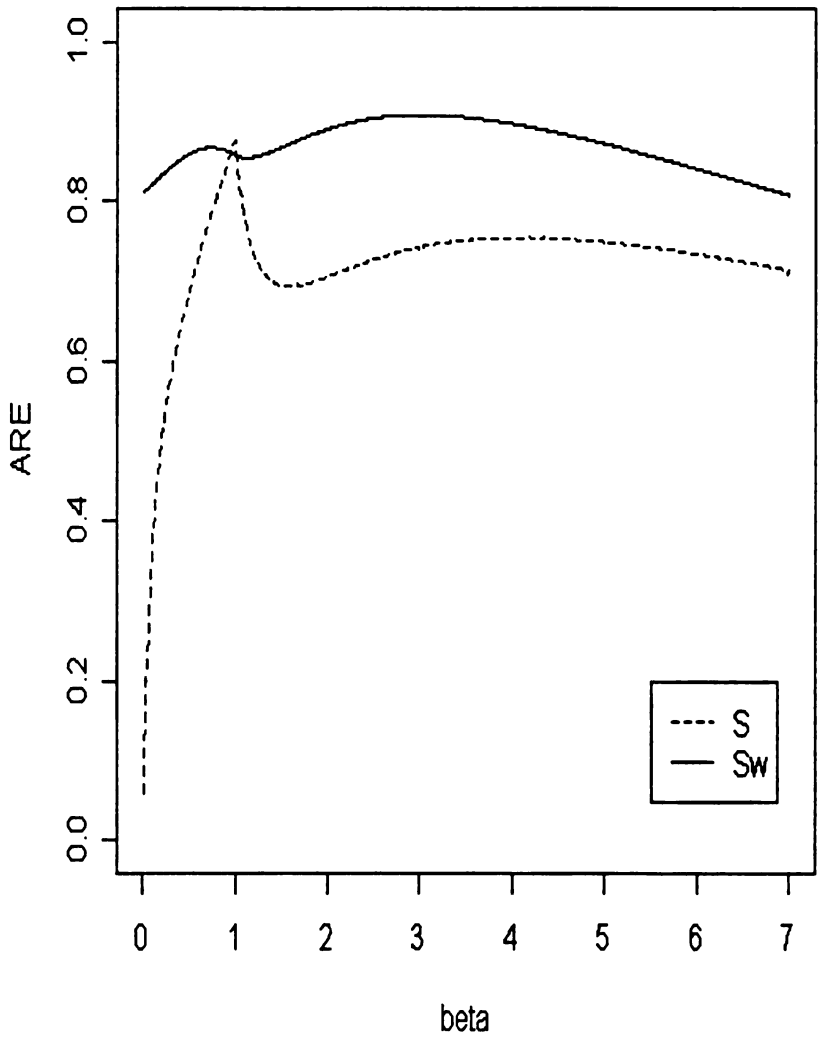

Figure 3.8. ARE of trimmed and winsorized standard deviations for Cauchy distribution

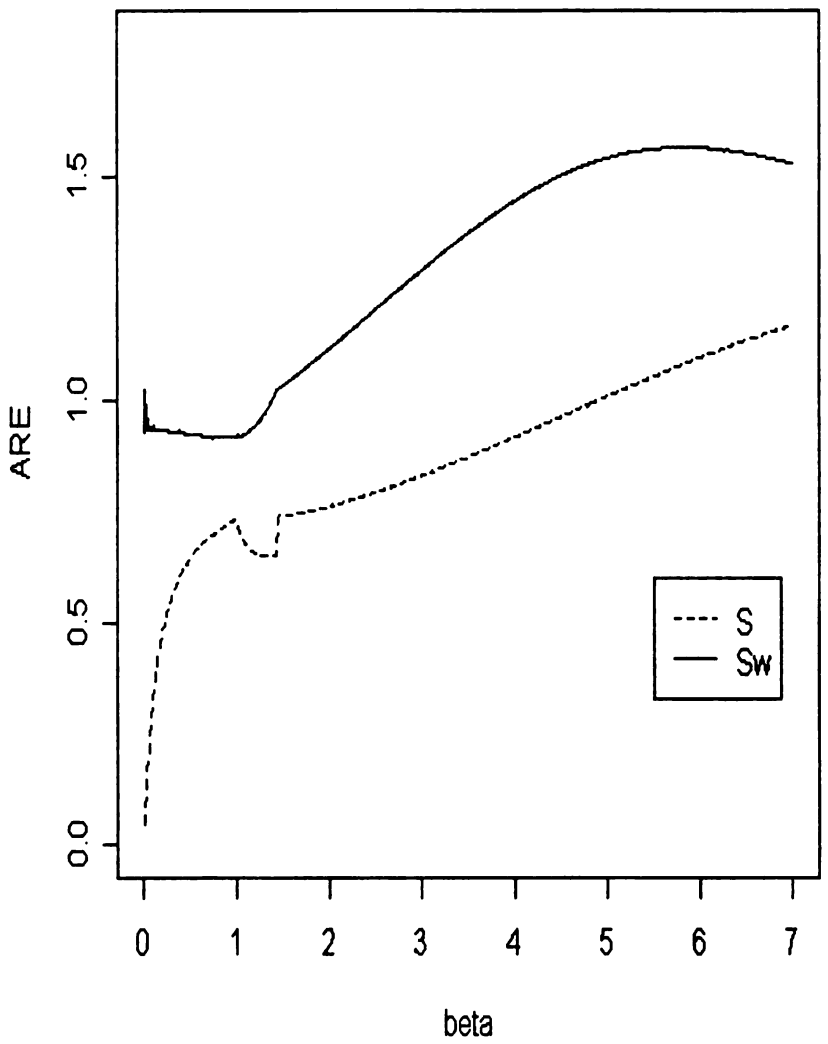

Figure 3.9. ARE of trimmed and winsorized standard deviations for exponential distribution

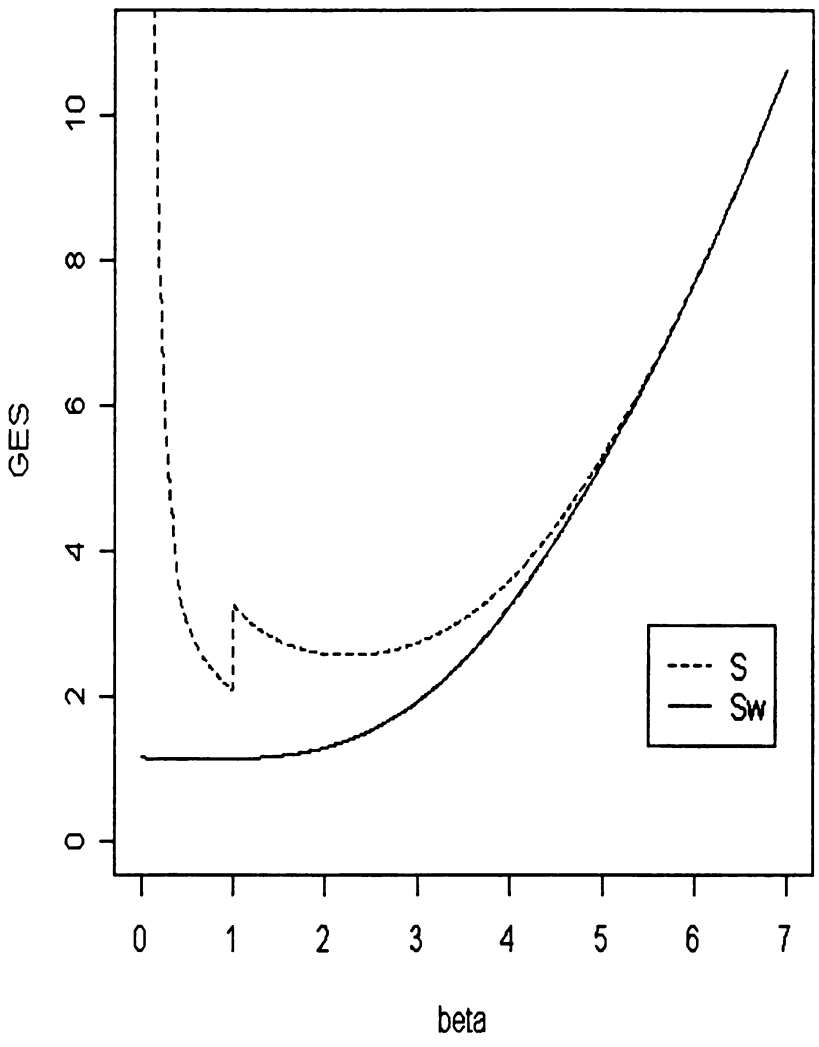

Figure 3.10. GES of trimmed and winsorized standard deviations for normal distribution

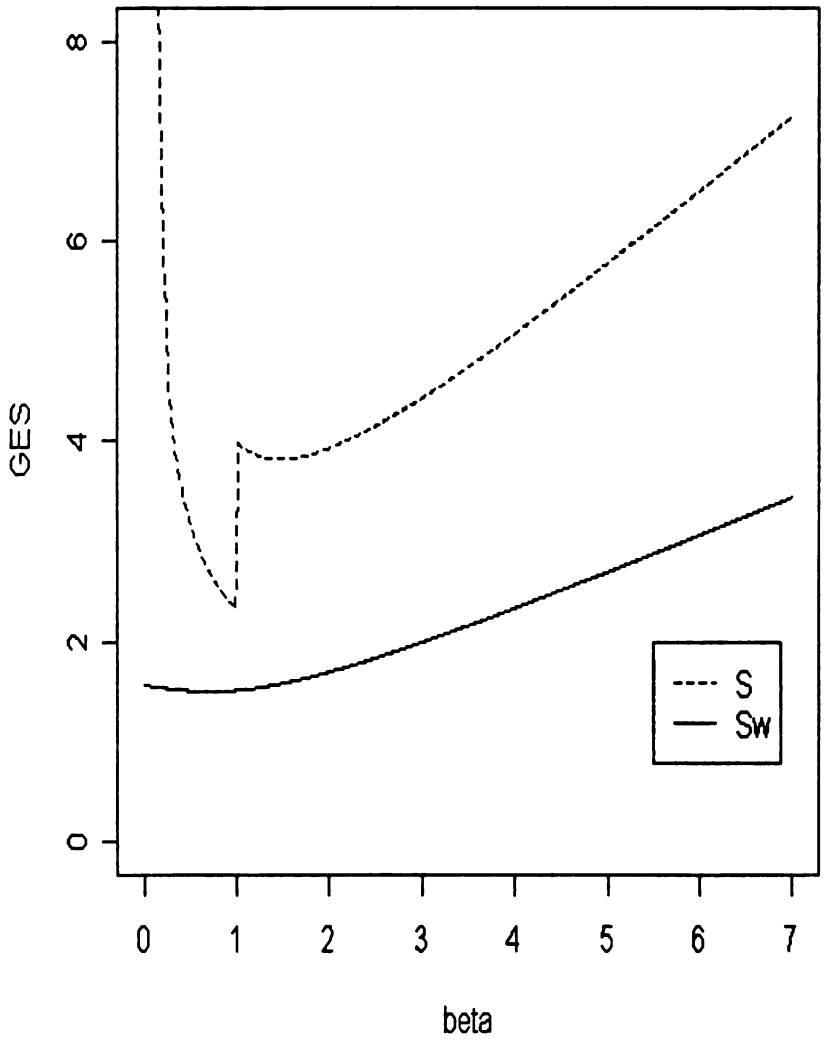

Figure 3.11. GES of trimmed and winsorized standard deviations for Cauchy distribution

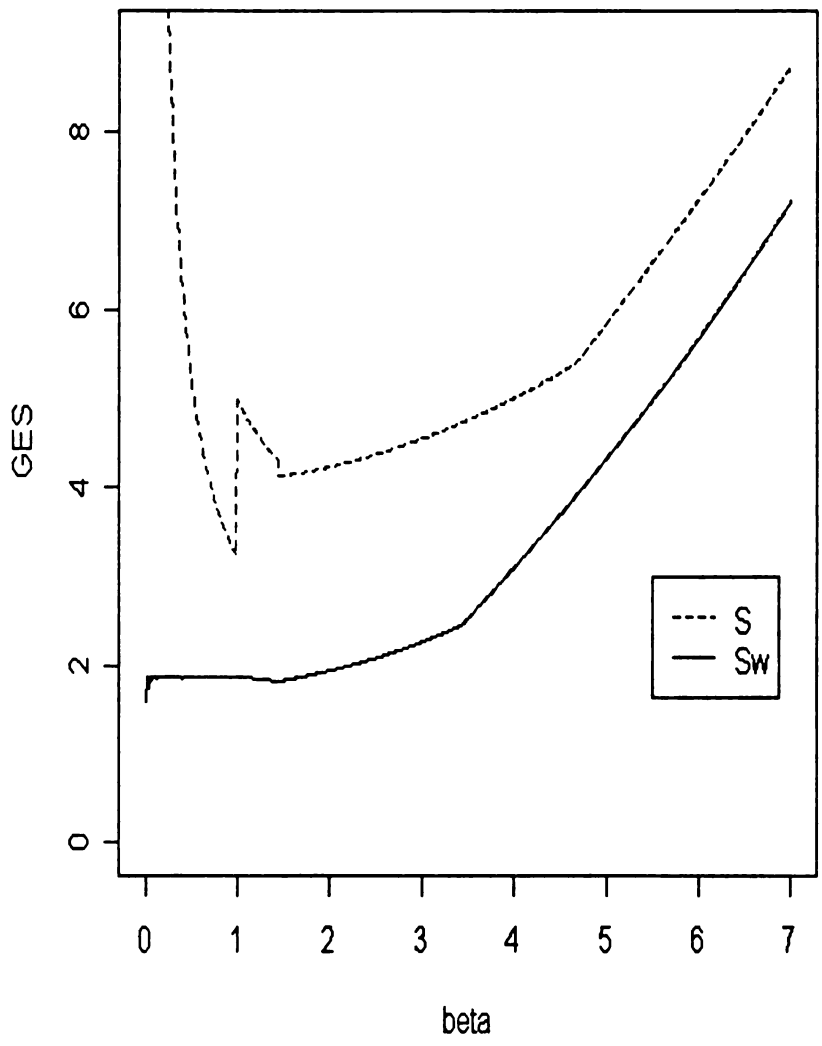

Figure 3.12. GES of trimmed and winsorized standard deviations for exponential distribution

CHAPTER 4

The Multiple Least Trimmed Squares Estimator

4.1 Introduction

Consider the multiple regression model

$$y_i = \mathcal{B}^t x_i + \varepsilon_i, \quad i = 1, \ldots, n$$

with $x_i = (x_{i1}, \ldots, x_{ip})^t \in \mathrm{R}^p$ and $y_i \in \mathrm{R}$. The matrix $\mathcal{B} \in \mathrm{R}^p$ contains the regression coefficients. The error terms $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. with zero center and a positive scale σ. Furthermore, we assume that the errors are independent of the carriers. Note that this model generalizes the location model ($x_i = 1$). Denote the entire sample $\{Z_n = (x_i; y_i); i = 1, \ldots, n\}$ and write $X = (x_1, \ldots, x_n)^t$ for the design matrix and $Y = (y_1, \ldots, y_n)^t$ for the vector of responses. The classical estimator for $\mathcal{B}$ is the least-squares (LS) estimator B_{LS} which is given by

$$\hat{\mathcal{B}}_{LS} = (X^t X)^{-1} X^t Y \tag{4.1.1}$$

while σ^2 is unbiasedly estimated by

$$\hat{\sigma}^2_{LS} = \frac{1}{n-p}(Y - X\hat{\mathcal{B}}_{LS})^2 \tag{4.1.2}$$

Since the least squares estimator is extremely sensitive to outliers, we aim to construct a robust alternative. An overview of strategies to robustify the multiple regression method

is given by Maronna and Yohai (1997) in the context of simultaneous equations models. Koenker and Portnoy (1990) apply a regression M-estimator to each coordinate of the responses and Bai et al. (1990) minimize the sum of the euclidean norm of the residuals. However, these two methods are not affine equivariant. Our approach will be different from the latter, since it will be affine equivariant. Agullo, J., and Croux, C., al.(2002) discussed some properties of multivariate trimmed least squares estimator. From computational aspect, Peter J. Rousseeuw, P. J. and Driessen, K. V. introduced an fast algorithm which is based on "C-step". "C-step" will gurantee the reduction of objective function during iterations.

However, as in location and scale settings, the general trimmed regression estimator suffers from very low efficiency while keeping the highest breakdown point. As shown in this chapter, the efficiency for ordinary least squares estimator is only 7.1% which is far less than satisfactory. Meanwhile, small proportion of trimming will cause low breakdown point issue and, hence, low robustness. In this Chapter, we are trying to work out this dilemma and find an estimator as in location and scale setting that has the highest breakdown point and can have high level efficiency. It turned out that this kind of estimator is hard to define and hard to discuss, don't not even mention to come up with an algorithm. The estimator defined in this chapter is expected to have highest breakdown point (strict proof is not available due to technical reasons) and truly efficient.

However we found that starting from any initial subset, keeping those points with residuals close to zero and after several steps of iterating, we found that the subsequent subset is relatively "stable". From figure 4.1 we can see that the mean squre of residuals is stabilized after a few number of iterations. Then we define an estimator with the minimum mean square error in the collection of "stable" sets, we obtain the estimator defined in this chapter. This definition is different from the general least squares estimator defined by Agullo, J., and Croux, C., al.(2002).

In Section 4.2 we give a formal definition of the multiple least trimmed squares (LTS) estimator. In Section 4.3, we derive the influence function and study the ARE. A time efficient algorithm to compute the LTS is presented in Section 4.4.

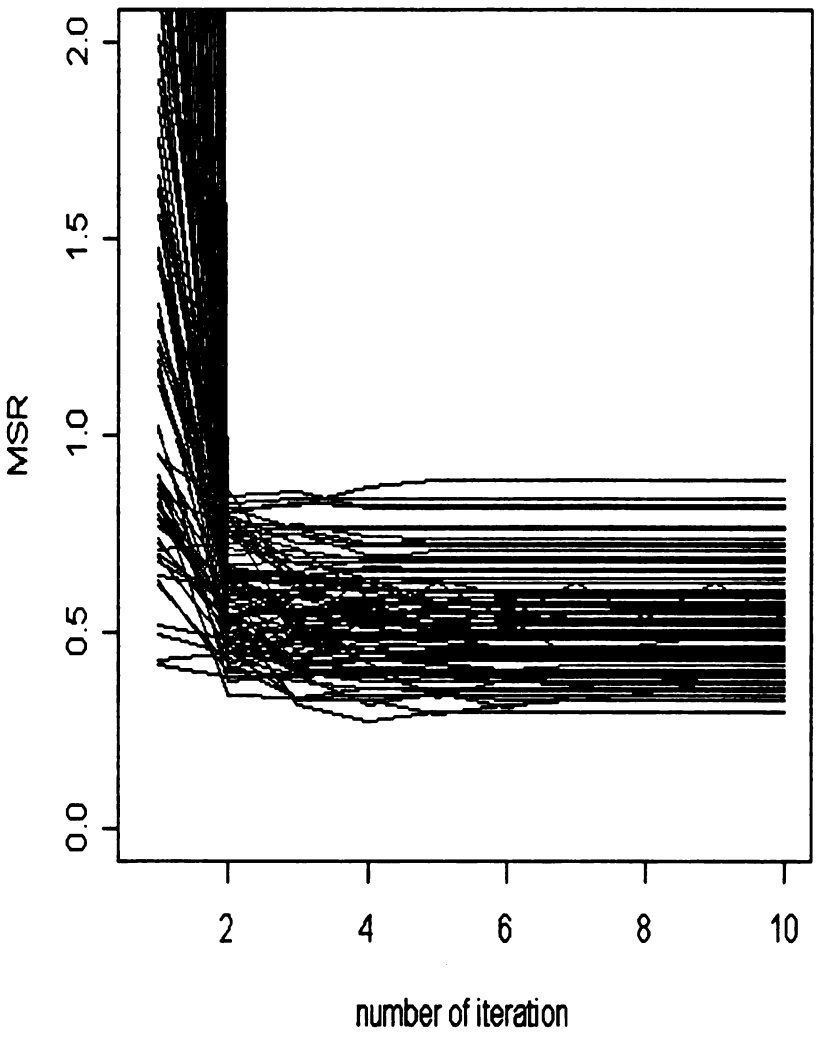

Figure 4.1. MSR vs number of iterations with 100 arbitrary initial subsets

4.2 Definition and properties

Our approach consists of finding the subset **H** of observations having the property that the Mean Square of its residuals from a LS-fit $\hat{\mathcal{B}}_{sub}$ solely based on this subset is minimal, the subset **H** of observations satisfies $\mathbf{H} = \{i : d_i(\hat{\mathcal{B}}_{sub}) \leq \ell d_{(h)}(\hat{\mathcal{B}}_{sub})\}$, let $h = [(n+p+1)/2]$ or $[(n+p+2)/2]$, $d_i(\hat{\mathcal{B}}_{sub}) = (Y - X\hat{\mathcal{B}}_{sub})^2$. When **H** is not unique, we take the one that has Minimum Mean Square of its residuals. Denote the collection of **H** by $\mathcal{H}$.

The resulting estimator will then be simply the LS-estimator computed from the optimal subset which defined by its least squares estimator. When $X = (1, \ldots, 1)^t \in \mathrm{R}^p$, it reduces to a multiple regression model with only an intercept that is a location model.

When $\ell = 1$, our approach is equivalent to the general least trimmed squares estimator (see, e.g. Agullo, etc) , which is a generalization of the LTS estimator (Rousseeuw 1984) for robust regression. Consider a dataset $Z_n = \{(x_i; y_i);\ i = 1, \ldots, n\} \in R^{p+1}$ and for any $\mathcal{B} \in \mathrm{R}^p$ denote $r_i(\mathcal{B}) = y_i - \mathcal{B}^t x_i$ the corresponding residuals.

Definition 4.2.1. *With the notation above the multiple least trimmed squares estimator (LTS) is defined as*

$$\hat{\mathcal{B}}_{LTS}(Z_n) = \hat{\mathcal{B}}_{sub}(\hat{H}) \quad \textit{where } \hat{H} \in \mathrm{argmin}_{\mathbf{H}\in\mathcal{H}} \hat{\sigma}^2_{LS}(H) \tag{4.2.1}$$

with $\hat{\sigma^2}_{LS}(\mathbf{H}) = \sum_{i\in H} r_i^2(\mathcal{B})/(\#(H) - p)$ *for any* $\mathbf{H} \in \mathcal{H}$. *The variance of the errors can then be estimated by*

$$\hat{\sigma}^2_{LTS}(Z_n) = c_\ell \hat{\sigma}^2_{sub}(\hat{\mathbf{H}}) \tag{4.2.2}$$

where c_ℓ *is a consistency factor.*

Note that if the minimization problem has more than one solution, in which case we look at $\mathrm{argmin}_H \hat{sigma}^2_{LS}(H)$ as a set, we arbitrarily select one of these solutions to determine the LTS estimator. In Section 5 a consistency factor c_ℓ will be proposed to attain Fisher consistency at the specified model. Note that for $\ell = +\infty$, we find back the

classical least squares estimator. Throughout the text we will suppose that the data set $Z_n = \{(x_i; y_i); i = 1, \ldots, n\} \in \mathrm{R}^{p+1}$ is in general position in the sense that no h points of Z_n are lying on the same hyperplane of R^{p+1}. Formally, this means that for all $\beta \in \mathrm{R}^p$, $\gamma \in \mathrm{R}$, it holds that

$$\#\{(x_j, y_j) | \beta^t x_j + \gamma y_j = 0\} < h \tag{4.2.3}$$

unless if β and γ are both zero vectors.

4.3 The influence function and asymptotic variances

The functional form of the LTS estimator can be defined as follows. Let K be an arbitrary $(p+1)$ dimensional continuously distribution which represents the joint distribution of the carriers and response variables.

Let us denote $d_A^2(x, y) = (y - \mathcal{B}(A)^t x)^2$, then it follows that $A = \{(x, y) \in \mathrm{R}^{p+1} | d_A^2(x, y) \leq \ell q(A)\}$ where $q(A) = (D_A^2)^{-1}(0.5)$ with $D_A^2(t) = P_K(d_A^2(x, y) \leq t)$.

Define

$$D_K(\ell) = \{A | A = \{(x, y) | d_A^2(x, y) \leq \ell q(A)\} \tag{4.3.1}$$

To define the LTS estimator at the distribution K we require that

$$P_K(\beta^t x = 0) < 1/2 \text{ for all } \beta \in \mathrm{R}^p \setminus \{0\} \tag{4.3.2}$$

For each $A \in D_K(\ell)$, the least squares solution over the set A is then given by

$$\mathcal{B}_A(K) = \Big(\int_A x x^t dK(x, y) \Big)^{-1} \int_A x y^t dK(x, y) \tag{4.3.3}$$

and A

$$\sigma_A^2(K) = \frac{\int_A (y - \mathcal{B}_A(K)^t x)^2 dK(x; y)}{P_K(A)} \tag{4.3.4}$$

Furthermore, a set $\hat{A} \in D_K(\ell)$ is called an LTS solution if $\sigma^2_{\hat{A}}(K) \leq \sigma^2_A(K)$ for any other $A \in D_K(\ell)$. The LTS functionals at the distribution K are then defined as

$$\mathcal{B}_{LTS}(K) = \mathcal{B}_{\hat{A}}(K) \text{ and } \sigma^2_{LTS}(K) = c_\ell \, \sigma^2_{\hat{A}}(K) \tag{4.3.5}$$

The constant c_ℓ can be chosen such that consistency will be obtained at the specified model. If the distribution K is not continuous, then the definition of $D_K(\ell)$ can be modified as in Croux and Haesbroeck (1999) to ensure that the set $D_K(\ell)$ is non-empty. Now consider the regression model

$$y = \mathcal{B}^t x + \varepsilon$$

where $x = (x_1, \ldots, x_p)^t$ is the p-dimensional vector of explanatory variables, and ε is the error term. We suppose that ε is independent of x and has a distribution F_σ with density

$$f_\sigma(u) = g(u^2/\sigma^2)/\sigma$$

where $\sigma > 0$. The function g is assumed to have a strictly negative derivative g' such that F_σ is a unimodal elliptically symmetric distribution around the origin. The distribution of $z = (x, y)$ is denoted by H. A regularity condition (to avoid degenerate situations) on the model distribution H is that

$$P_H(\beta^t x + \gamma^t y = 0) < 1/2 \tag{4.3.6}$$

for all $\beta \in \mathrm{R}^p$ and $\gamma \in \mathrm{R}$ not both equal to zero at the same time. This general position condition says that the maximal amount of probability mass of H lying on the same hyperplane must be lower than 1/2.

Theorem 4.3.1. *Denote*

$$c_\ell = \frac{F(\ell q) - 1}{\int_{u^2 \le \ell q} u^2 dF_0(u)}$$

Where F is the symmetric error distribution and $q = K^{-1}(0.5)$ with $K(t) = P_{F_0}(\varepsilon^t \varepsilon \le t)$. Then the functionals $\mathcal{B}_{LTS}$ and σ^2_{LTS} are Fisher-consistent estimators for the parameters $\mathcal{B}$ and σ^2 at the model distribution K:

$$\mathcal{B}_{LTS}(K) = \mathcal{B} \qquad and \quad \sigma^2_{LTS} = \sigma^2$$

Proof. First of all, due to equivariance, we may assume that $\mathcal{B} = 0$ and $\sigma^2 = 1$, so $y = \varepsilon \sim F$. It now suffices to show that $\mathcal{B}_{LTS}(K) = 0$. Then we will have that $\sigma^2(K)$ is that LTS functional at the distribution of $y - B_{LTS}(K)^t x = y = \varepsilon$, the consistent coefficient c can be easily derived. Since $\mathcal{B}_{LTS}$ is defined solely based on the set $C = \{(x, y) \in R^{p+1} | (y - B^t_{LTS} x)^2 \le \ell q\}$. Therefore

$$\int_C x(y - \mathcal{B}^t_{LTS} x) dK(x, y) = 0 \tag{4.3.7}$$

Now suppose that $\mathcal{B}_{LTS} \neq 0$. From (4.3.7) it follows that

$$\int_C \mathcal{B}^t_{LTS} x(y - \mathcal{B}^t_{LTS} x) dF(y) dG(y) = 0$$

Which can be rewritten as

$$\int_R \mathcal{B}^t_{LTS} x \mathbf{I}(x) dG(y) = 0 \tag{4.3.8}$$

with

$$\mathbf{I}(x) = \int_{d-\sqrt{\ell q}}^{d+\sqrt{\ell q}} (y - \mathcal{B}^t_{LTS} x) dF(y),$$

where $C_x = \{y \in R | (x, y) \in C\}$, Fix x and set $d = \mathcal{B}^t_{LTS} x$. Since y is symmetric, we have that

$$\begin{aligned} \mathbf{I}(x) &= \int_{d-\sqrt{\ell q}}^{d+\sqrt{\ell q}} (y - \mathcal{B}^t_{LTS} x) dF(y) \\ &= \int_0^{\sqrt{\ell q}} t(g((d+t)^2) - g((d-t)^2)) dt \end{aligned}$$

If $d > 0$ we have $(d+t)^2 > (d-t)^2$ (for $t > 0$) and since g is strictly decreasing this implies $\mathbf{I}(x) < 0$. Similarly, we can show that $d < 0$ implies $\mathbf{I}(x) > 0$. Also, $\mathcal{B}^t_{LTS} x = 0$ implies

$\mathbf{I}(x) = 0$. However, due to condition (4.3.6), the latter event occurs with probability less than 0.5. Therefore, we obtain $\int_C x(y - \mathcal{B}_{LTS}^t x)dK(x, y) < 0$ which contradicts (4.3.8), so we conclude that $\mathcal{B}_{LTS} = 0$. □

The influence function of a functional T at the distribution K measures the effect on T of adding a small mass at $z = (x, y)$. If we denote the point mass at z by Δ_z and consider the contaminated distribution $K_{\varepsilon,z} = (1 - \varepsilon)K + \varepsilon\Delta_z$ then the influence function is given by

$$IF(z; T, K) = \lim_{\varepsilon \to 0} \frac{T(K_{\varepsilon,z}) - T(K)}{\varepsilon} = \frac{\partial}{\partial \varepsilon} T(K_{\varepsilon,z})|_{\varepsilon=0}.$$

(See Hampel et al. 1986.) It can easily be seen that the LTS is equivariant for affine transformations of the regressors and responses and for regression transformations which add a linear function of the explanatory variables to the responses. Therefore, it suffices to derive the influence function at a model distribution K_0 for which $\mathcal{B} = 0$ and the error distribution $F = F_0$ with density $f_0(y) = g(y^2)$. The following theorem gives the influence function of the LTS regression functional at K_0.

Theorem 4.3.2. *With the notations from above, we have that*

$$IF(z; \mathcal{B}_{LTS}, K_0) = E_G[XX^t]^{-1} \frac{xy}{(2F(\sqrt{\ell q}) - 1) - 2\sqrt{\ell q} f(\sqrt{\ell q})} \mathbf{I}(y^2 \leq \ell q) \tag{4.3.9}$$

Proof. Consider the contaminated distribution $K_\varepsilon = (1 - \varepsilon)K_0 + \varepsilon\Delta_{z_0}$ with $z_0 = (x_0, y_0)$ and denote $B_\varepsilon := \mathcal{B}_{LTS}(K_\varepsilon)$ and $\sigma_\varepsilon^2 := \sigma_{LTS}^2(K_\varepsilon)$. Then (4.3.3) results in

$$\hat{B}_\varepsilon = \Big(\int_{\hat{A}_\varepsilon} xx^t dK_\varepsilon(x; y) \Big)^{-1} \int_{\hat{A}_\varepsilon} xy^t dK_\varepsilon(x; y)$$

where $\hat{A}_\varepsilon \in \mathcal{D}_{K_\varepsilon}(\alpha)$ is a LTS solution. Differentiating w.r.t. ε and evaluating at 0 yields

$$\begin{aligned} IF(z_0; \mathcal{B}_{LTS}, K_0) = & \Big(\int_{\hat{A}} xx^t dK_0(x; y) \Big)^{-1} \frac{\partial}{\partial \varepsilon} \int_{\hat{A}_\varepsilon} xy^t dK_\varepsilon(x; y) \Big|_{\varepsilon=0} \\ & + \frac{\partial}{\partial \varepsilon} \Big[\Big(\int_{\hat{A}_\varepsilon} xx^t dK_\varepsilon(x; y) \Big)^{-1} \Big] \Big|_{\varepsilon=0} \int_{\hat{A}} xy^t dK_0(x, y) \end{aligned}$$

Fisher-consistency yields that $\hat{A} = \{(x, y) : \mathrm{R}^{p+q}; y^2 \leq \ell q\}$ where $q = (D_F^2)^{-1}(1/2)$ with $D_F^2(t) = P_F(y^2 \leq t)$. Hence $\hat{A} = R^p \times \{y \in \mathrm{R}; y^2 \leq \ell q\} =: \mathrm{R}^p \times A$. This implies

$$\int_{\hat{A}} xy^t dK_0(x,y) = \int_{\mathrm{R}^p} x dG(x) \int_A y dF(y) = 0$$

by symmetry of F and

$$\int_{\hat{A}} xx^t dK_0(z) = \int_{\mathrm{R}^p} xx^t dG(x) \int_A dF(y) = (2F(\sqrt{\ell q}) - 1) E_G[xx^t]$$

Therefore, we obtain

$$IF(z_0; \mathcal{B}_{LTS}, K_0) = \frac{E_G[XX^t]}{(2F(\sqrt{\ell q}) - 1)} \frac{\partial}{\partial \varepsilon} \int_{\hat{A}_\varepsilon} xy^t dK_\varepsilon(x;y)\Big|_{\varepsilon=0} \tag{4.3.10}$$

$$= \frac{E_G[XX^t]}{(2F(\sqrt{\ell q}) - 1)} \frac{\partial}{\partial \varepsilon} \Big((1-\varepsilon) \int_{\hat{A}_\varepsilon} xy^t dK_0(x;y) + \varepsilon x_0 y_0 \mathbf{I}(z_0 \in \hat{A}_\varepsilon) \Big|_{\varepsilon=0} \tag{4.3.11}$$

$$= \frac{E_G[XX^t]}{(2F(\sqrt{\ell q})} \Big(\frac{\partial}{\partial \varepsilon} \int_{\hat{A}_\varepsilon} xy^t dK_0(x;y) + x_0 y_0 \mathbf{I}(y^2 \le \ell q) \Big|_{\varepsilon=0} \tag{4.3.12}$$

Let us denote $d_\varepsilon^2(x;y) = (y - \mathcal{B}_\varepsilon^t x)^2$, then it follows that $\hat{A}_\varepsilon = \{(x,y) \in \mathrm{R}^{p+1}; d_\varepsilon^2(x;y) \le \ell q(\varepsilon)\}$ where $q(\varepsilon) = (D_{K_\varepsilon}^2)^{-1}(0.5)$ with $D_{K_\varepsilon}^2(t) = P_{K_\varepsilon}(d_\varepsilon^2(x;y) \le t)$. For x fixed we define the set $\mathcal{E}_{\varepsilon,x} := \{y \in R | d_\varepsilon^2(x;y) \le \ell q(\varepsilon)\}$. Then it follows that

$$\int_{\hat{A}_\varepsilon} xy^t dK_0(x;y) = \int_{R^p} \int_{\mathcal{E}_{\varepsilon,x}} xy dF(y) dG(x)$$

$$= \int_{R^p} \int_{\mathcal{E}_{\varepsilon,x}} y g(y^2) dy x dG(y) \tag{4.3.13}$$

Using the transformation $v = y - \mathcal{B}_\varepsilon^t x$, we obtain that

$$I(\varepsilon) := \int_{\mathcal{E}_{\varepsilon,x}} y g(y^2) dy$$

$$= \int_{v^2 \le \ell q(\varepsilon)} (v + \mathcal{B}_\varepsilon^t x) g((v + \mathcal{B}_\varepsilon^t x)^2) dv$$

$$= \int_{\sqrt{-\ell q(\varepsilon)}}^{\sqrt{\ell q(\varepsilon)}} (v + \mathcal{B}_\varepsilon^t x) g((v + \mathcal{B}_\varepsilon^t x)^2) dv$$

Note that

$$\begin{aligned}
\frac{I(\varepsilon)-I(0)}{\varepsilon} =& \frac{1}{\varepsilon}\Big[\Big(\int_{-\sqrt{\ell q(\varepsilon)}}^{\sqrt{\ell q(\varepsilon)}}(v+\mathcal{B}_\varepsilon^t x)g((v+\mathcal{B}_\varepsilon^t x)^2)dv - \int_{-\sqrt{\ell q}}^{\sqrt{\ell q}}(v+\mathcal{B}_\varepsilon^t x)g((v+\mathcal{B}_\varepsilon^t x)^2)dv\Big) \\
&+ \Big(\int_{-\sqrt{\ell q}}^{\sqrt{\ell q}}(v+\mathcal{B}_\varepsilon^t x)g((v+\mathcal{B}_\varepsilon^t x)^2)dv - \int_{-\sqrt{\ell q}}^{\sqrt{\ell q}}(v+\mathcal{B}^t x)g((v+\mathcal{B}^t x)^2)dv\Big)\Big] \\
=& \frac{1}{\varepsilon}\Big(\int_{-\sqrt{\ell q}}^{\sqrt{\ell q(\varepsilon)}}(v+\mathcal{B}_\varepsilon^t x)g((v+\mathcal{B}_\varepsilon^t x)^2)dv - \int_{-\sqrt{\ell q}}^{-\sqrt{\ell q(\varepsilon)}}(v+\mathcal{B}_\varepsilon^t x)g((v+\mathcal{B}_\varepsilon^t x)^2)dv\Big) \\
&+ \Big(\int_{-\sqrt{\ell q}}^{\sqrt{\ell q}}\frac{1}{\varepsilon}\Big((v+\mathcal{B}_\varepsilon^t x)g((v+\mathcal{B}_\varepsilon^t x)^2)dv - (v+\mathcal{B}^t x)g((v+\mathcal{B}^t x)^2)\Big)dv \\
=& (\theta_1+\mathcal{B}_\varepsilon^t x)g((\theta_1+\mathcal{B}_\varepsilon^t x)^2)\frac{\sqrt{\ell q(\varepsilon)}-\sqrt{\ell q}}{\varepsilon} - (-\theta_2 \\
&+ \mathcal{B}_\varepsilon^t x)g((\theta_2+\mathcal{B}_\varepsilon^t x)^2)\frac{(-\sqrt{\ell q(\varepsilon)}-(-\sqrt{\ell q})}{\varepsilon} \\
&+ \Big(\int_{-\sqrt{\ell q}}^{\sqrt{\ell q}}\frac{1}{\varepsilon}\Big((v+\mathcal{B}_\varepsilon^t x)g((v+\mathcal{B}_\varepsilon^t x)^2)dv - (v+\mathcal{B}^t x)g((v+\mathcal{B}^t x)^2)\Big)dv \\
=& \int_{-\sqrt{\ell q}}^{\sqrt{\ell q}}\Big(IF(z_0;\mathcal{B}_{LTS},K_0)^t x g(v^2) + 2v^2 g'(v^2)IF(z_0;\mathcal{B}_{LTS},K_0)^t x\Big)dv + o_{z_0}(1)
\end{aligned}$$

So we have that

$$\frac{\partial}{\partial\varepsilon}\mathbf{I}(\varepsilon)|_{\varepsilon=0} = \big((2F(\sqrt{\ell q})-1)+2c_2\big)IF(z_0;\mathcal{B}_{LTS},K_0)^t x$$

where $c_2 = \int_{-\sqrt{\ell q}}^{\sqrt{\ell q}} g'(v^2)v^2dv = \int_0^{\sqrt{\ell q}} vdf(v) = \sqrt{\ell q}f(\sqrt{\ell q}) - \frac{1}{2}(2F(\sqrt{\ell q})-1)$

□

Note that the influence function is bounded in y but unbounded in x. Closer inspection of (6.1) shows, however, that only good leverage points, which have outlying x but satisfy the regression model, can have a high effect on the LTS estimator. Bad leverage points will give a zero influence.

Remark: The influence function of the LTS location estimator T at a symmetric distribution F_0 can be obtained easily, it is given by

$$IF(y;T;F_0) = \frac{y}{(2F(\sqrt{\ell q})-1)-2\sqrt{\ell q}f(\sqrt{\ell q})}\mathbf{I}(y^2 \le \ell q)$$

Therefore, it follows that the influence function of $\mathcal{B}_{LTS}$ can be rewritten as

$$IF(z;\mathcal{B}_{LTS},K_0) = E_G[XX^t]^{-1}xIF(y;T,F_0): \qquad (4.3.14)$$

The asymptotic variance-covariance matrix of $\mathcal{B}_{LTS}$ can now be computed by means of $ASV(\mathcal{B}_{LTS}, K_0) = E_K[IF(z; \mathcal{B}_{LTS}, K_0) \otimes IF(z; \mathcal{B}_{LTS}, K_0)^t]$ (see e.g. Hampel et al. 1986). Here $A \otimes B$ denotes the Kronecker product of a $(p \times 1)$ matrix A with a $(1 \times p)$ matrix B, which results in a $(p \times p)$ matrix with (i,j)-th block equals $a_i b_j$, where a_j are the elements of the matrix A and b_j are the elements of the matrix B. Let us denote $\Sigma_x^2 := E_G[XX^t]$, then expression (4.3.14) implies that

$$ASV(\mathcal{B}_{LTS}, K_0) = ASV(T, F_0))\Sigma_x^{-1}) \tag{4.3.15}$$

From (4.3.15) it follows that for every $1 \leq i \leq p$ the asymptotic variance of $(\mathcal{B}_{LTS})_i$ equals

$$ASV((\mathcal{B}_{LTS})_i, K_0) = E_K[IF^2(z; (\mathcal{B}_{LTS})_i, K_0)] = (\Sigma_x^{-1})_{ii} ASV(T, F_0))$$

For $i \neq i'$ we obtain the asymptotic covariances

$$\begin{aligned} ASV((\mathcal{B}_{LTS})_i, (\mathcal{B}_{LTS})_{i'}, K_0) &= E_K[IF(z; (\mathcal{B}_{LTS})_i, K_0) IF(z; (\mathcal{B}_{LTS})_{i'}, K_0)] \\ &= (\Sigma_x^{-1})_{ii'} ASV(T, F_0) \end{aligned}$$

and all other asymptotic covariances (for $j' \neq j$) equal 0. Due to affine equivariance, we may consider w.l.o.g. the case where $\sigma = 1$. Then all asymptotic covariances are zero, while $ASV((\mathcal{B}_{LTS})_i, K_0) = ASV(T, F_0)$ for all $1 \leq i \leq p$. The limit case $\ell = \infty$ yields the asymptotic variance of the least squares estimator $ASV((\mathcal{B}_{LS})_i, K_0) = ASV(M; F_0)$ where M is the functional form of the sample mean. Therefore, we can compute the asymptotic relative efficiency of the LTS estimator at the model distribution K_0 with respect to the least squares estimator as

$$ASV((\mathcal{B}_{LTS})_i, K_0) = \frac{ASV((\mathcal{B}_{LS})_i; K_0)}{ASV((\mathcal{B}_{LTS})_i; K_0)} = \frac{ASV(M; F_0)}{ASV(T; F_0)} = ARE(T, F_0)$$

for all $1 \leq i \leq p$. Hence the asymptotic relative efficiency of the LTS estimator in $p+1$ dimensions does not depend on the distribution of the carriers, but only on the distribution of the errors and equals the asymptotic relative efficiency of the LTS location estimator at the error distribution F_0. For the normal distribution these relative efficiencies are given in

Table 1. Note that the efficiency of LTS does not depend on p, the number of explanatory variables, but only on the number of dependent variables.

Table 4.1. Asymptotic relative efficiency of the LTS estimator w.r.t. the Least Squares estimator at the normal distribution for several values of ℓ.

ℓ	1	3	5	7	10	20	30
ARE	0.071	0.286	0.483	0.636	0.792	0.973	0.997

4.4 Finite-sample simulations

4.4.1 Algorithm

In algorithmic terms, the procedure can be described as follows:

Step I

1. Create an initial subsets H_0. Draw a random p-subset J, and compute $\hat{\theta}_0 :=$ the coefficients of the hyperplane through J. If J does not define a unique hyperplane (i.e., when the rank of X_J is less than p), redraw J random observations until it does. Then compute the residuals $r_0(i) := y_i - \hat{\theta}_0^t x_i^i$ for $i = 1, \ldots, n$. Sort the absolute values of these residuals, which yields a permutation π for which $r_0^2(\pi(1)) \leq r_0^2(\pi(2)) \leq \cdots \leq r_0^2(\pi(n))$, $H_0 = \{i | r_0^2(i) \leq \ell r_0^2(\pi(h))\}$, $h = [n/2] + [(p+1)/2]$.

3. Compute $\hat{\beta}_0 :=$ LS regression estimator based on H_0.

4. Iterate K (say 20) steps or until $H_k = H_{k-1}$ or $H_k = H_{k-2}$, record β, H in the last step and k(the actual number of iteration).

Step II

Repeat M (say 300) times step I by choose different initial subsets K_0. For simplicity, write β_i, H_i and k_i, $i = 1, \ldots, M$

Step III

If $min(k_i) < K$, then Choose $\text{argmin}_{1 \leq i \leq M, k_i < K} \sum_{j \in H_i} r_j^2(\beta_i)/(\#(H_i) - p)$.

Else If $min(k_i) = K$, then Choose $\text{argmin}_{1 \leq i \leq M} \sum_{j \in H_i} r_j^2(\beta_i)/(\#(H_i) - p)$.

The key to the algorithm is to find H–a subset having the property that the Mean Square of its residuals from a LS-fit $\hat{\mathcal{B}}_{sub}$ solely based on this subset is minimal, the subset H of observations satisfies $\mathbf{H} = \{i : d_i(\hat{\mathcal{B}}_{sub}) \leq \ell d_{(h)}(\hat{\mathcal{B}}_{sub})\}$, where $h = [(n+p+1)/2]$ or $[(n+p+2)/2]$, $d_i(\hat{\mathcal{B}}_{sub}) = (Y - X\hat{\mathcal{B}}_{sub})^2$. We call H "stable set" which won't be changed if we iterate by use H as an initial set. Sometimes, H doesn't exist, however, we can find H_1 and H_2 and they appear alternatively–If we choose H_1 as an initial set and iterate one step we reach H_2 and if we iterate once more by choosing H_2 as an initial set, we come back to H_1. H_1 and H_2 form 'closed sets'. This idea is revealed in Step I–4.

4.4.2 Finite-sample performance

In this section we investigate the finite-sample performance of the LTS estimator. Therefore, we will compare the asymptotic efficiency obtained in the previous section with finite sample efficiencies obtained by simulation. To this end, we performed the following simulations. For various sample sizes n, and for $p = 3$, we generated $m = 10000$ regression datasets of size n. The response variables were generated from the standard normal distribution $N(0,1)$, and w.l.o.g. we took $\mathcal{B} = 0$ in the multiple regression model. We set the p-th regressor equal to one, so we consider a regression model with intercept. The remaining $p-1$ explanatory variables were generated from the following distributions:

The multiple standard normal distribution $N(0; I_{p-1})$.

In this simulation setup, the last element of $\mathcal{B}$ is the intercept vector and the matrix formed by the $p-1$ first rows of $\mathcal{B}$, which we will denote by $\mathcal{B}_0$, is the slope matrix. $h = [(n+p+2)/2]$. For the parameter ℓ, we let it vary.

For each simulated dataset $Z^{(l)}$, $l = 1, \ldots, m$ we computed the regression vector $\hat{B}^{(l)}_{LTS}$. The Monte Carlo variance of a regression coefficient $(\hat{B}_{LTS})_j$ is measured as $Var((\hat{B}_{LTS})_j) = n\,\mathrm{var}_l((\hat{B}_{LTS})^l_j)$ for $j = 1, \ldots, p-1$. The variance of the estimated slope matrix $\hat{B}^0_{LTS}$ is then summarized by $ave_j(Var((\hat{B}_{LTS})_j))$ for $1 \leq j \leq p-1$ while its inverse measures the finite-sample efficiency of the slope. Similarly we computed the finite-sample efficiency of the intercept vector.

Table 4.2 shows the finite-sample efficiencies of the LTS estimator obtained by simulation for sample size n equal to 100, 300, and 500. We see that the finite-sample efficiencies of the

LTS converge to the corresponding asymptotic efficiencies which are listed under $n = \infty$

Table 4.2. AREs of LTS relative to the LS for $p = 3$

ℓ		100	300	500	∞
1	Slope	0.047	0.020	0.013	0.071
	Intercept	0.068	0.030	0.019	0.071
3	Slope	0.193	0.110	0.077	0.286
	Intercept	0.254	0.156	0.112	0.286
5	Slope	0.409	0.341	0.247	0.483
	Intercept	0.480	0.395	0.377	0.483
7	Slope	0.589	0.575	0.549	0.636
	Intercept	0.662	0.592	0.647	0.636
10	Slope	0.748	0.758	0.767	0.792
	Intercept	0.706	0.714	0.805	0.792
20	Slope	0.944	0.935	0.952	0.973
	Intercept	0.946	0.874	0.998	0.973
30	Slope	0.941	0.974	1.000	0.997
	Intercept	0.918	0.952	1.000	0.997

Note that efficiencies for large ℓ are always higher than the corresponding efficiency for small ℓ. We also did a study by choosing different p value. The results of them confirm the theoretical ARE.

To illustrate the difference between ordinary least trimmed squares estimator (OLTS) and the trimmed squares estimator introduced in this chapter. We generated a regression data set with $n = 1000$ observations of which 30% bad leverage points. The first 700 observations were generated by the formula

$$y_i = 1x_i + 1 + \varepsilon_i \qquad i = 1, 2, \ldots, 700,$$

where $x_i \sim N(0,\ 100)$ and $\varepsilon_i \sim N(0,\ 1)$. The other 300 observations $(x_i,\ y_i)$ were drawn from the bivariate normal distribution with $\mu = (50, 0)$ and $\Sigma = 25\mathbf{I}_2$. The entire data sets were shown in figure. We compute 75% OLTS and $\ell = 30$ LTS.

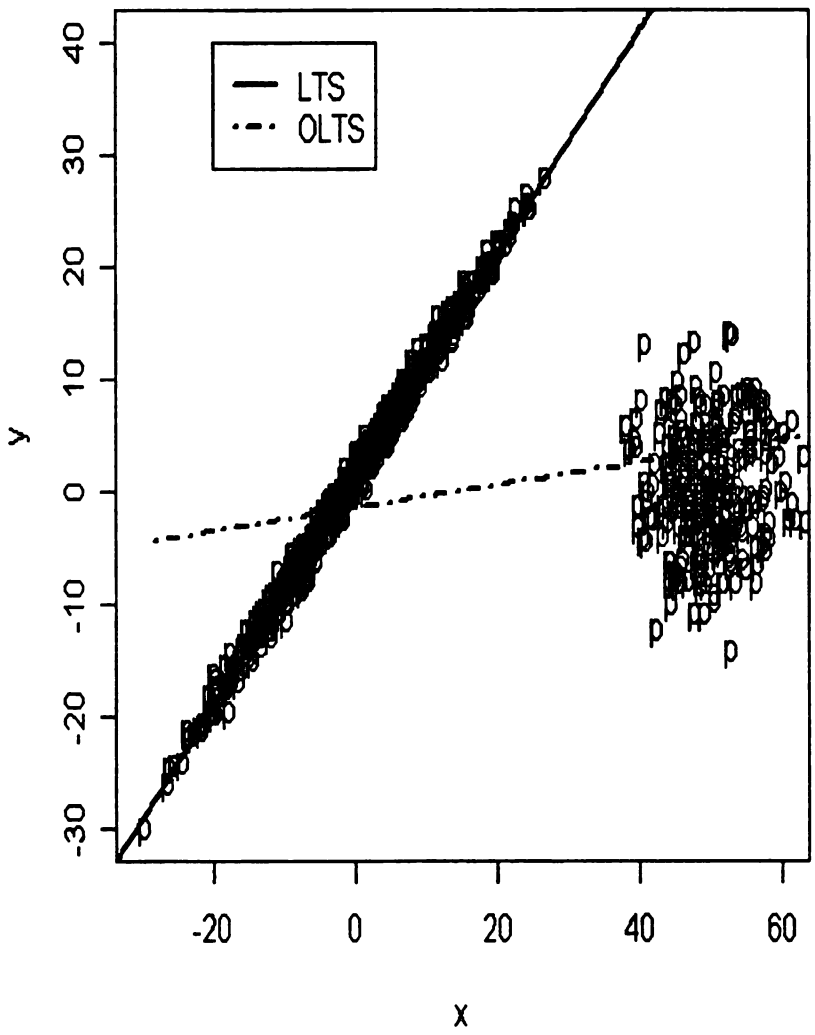

Figure 4.2. LTS estimator v.s OLTS estimator

CHAPTER 5

Selected proofs of main results and lemmas

5.1 Selected proofs for results of chapter 2

PROOF OF THEOREM 2.3.1. The proof follows the lines of that given in Zuo (2003). For simplicity we sometimes write F_ε for $F(\varepsilon, \delta_x)$ for a given $x \in R$. Denote by $o_x(1)$ a quantity that may depend on x but approaches 0 as $\varepsilon \to 0$ for the given x. We need the following results whose more general versions are given and treated in Zuo (2003).

Lemma 5.1.1. *For fixed $x \in$ R and sufficiently small ε, We have for fixed $0 < \beta < \infty$*
(a) $D(y, F)$ and $D(y, F(\varepsilon, \delta_x))$ are Lipschitz continuous in $y \in$ R;
(b) $\sup_{y \in S} |D(y, F(\varepsilon, \delta_x)) - D(y, F)| = o_x(1)$ for any bounded set $S \subset R$;
(c) $|L(F(\varepsilon, \delta_x) - L(F)| = o_x(1)$, $|U(F(\varepsilon, \delta_x)) - U(F)| = o_x(1)$.

First we write

$$T(F_\varepsilon) - T(F) = \frac{\int_{L(F_\varepsilon)}^{U(F_\varepsilon)} (y - T(F)) w(D(y, F_\varepsilon)) dF_\varepsilon(y)}{\int_{L(F_\varepsilon)}^{U(F_\varepsilon)} w(D(y, F_\varepsilon)) dF_\varepsilon(y)} \tag{5.1.1}$$

We focus on the numerator. The denominator can be treated in the same (but less involved)

manner. The numerator can clearly be decomposed into three terms

$$I_{1\varepsilon} = \Big(\int_{L(F_\varepsilon)}^{U(F_\varepsilon)} - \int_{L(F)}^{U(F)}\Big)(y - T(F))w(D(y, F_\varepsilon))dF_\varepsilon(y),$$

$$I_{2\varepsilon} = \int_{L(F)}^{U(F)} (y - T(F))(w(D(y, F_\varepsilon)) - w(D(y, F)))dF_\varepsilon(y),$$

$$I_{3\varepsilon} = \int_{L(F)}^{U(F)} (y - T(F))w(D(y, F))dF_\varepsilon(y).$$

It follows immediately that

$$\frac{1}{\varepsilon} I_{3\varepsilon} = \mathbf{I}_{\{x \in [L(F), U(F)]\}}(x - T(F))w(D(x, F)), \quad \text{uniformly in } y \text{ for the given } x. \quad (5.1.2)$$

In light of the continuity of $\mathrm{w}^{(1)}(\cdot)$ and Lemma 5.1.1, we have that

$$\mathrm{w}(D(y, F_\varepsilon)) - \mathrm{w}(D(y, F)) = (\mathrm{w}^{(1)}(D(y, F)) + o_x(1))(D(y, F_\varepsilon)) - D(y, F)), \quad (5.1.3)$$

uniformly in y for y in a bounded set S for the given x. This, combining with the boundedness of $L(F)$ and $U(F)$ and Lemma 5.1.1, immediately gives

$$\frac{1}{\varepsilon} I_{2\varepsilon} = \int_{L(F)}^{U(F)} (y - T(F))w^{(1)}(D(y, F))IF(x; D(y, F))dF(y) + o_x(1) \quad (5.1.4)$$

In virtue of equation (5.1.3), boundedness of $L(F)$ and $U(F)$, Lemma 5.1.1, and the argument used above, we have for sufficiently small $\varepsilon > 0$

$$\frac{1}{\varepsilon} I_{1\varepsilon} = \frac{1}{\varepsilon} \int \Delta(y, \varepsilon)(y - T)w(D(y, F))dF(y) - \int \Delta(y, \varepsilon)(y - T)w(D(y, F))dF(y)$$
$$+ \int \Delta(y, \varepsilon)(y - T(F))w^{(1)}(D(y, F))IF(x; D(y, F))dF(y) + o_x(1)$$

where $\Delta(y, \varepsilon) = \mathbf{I}_{\{y \in [L(F_\varepsilon), U(F_\varepsilon)]\}} - \mathbf{I}_{\{y \in [L(F), U(F)]\}}$. Call the three terms with integration $I_{1\varepsilon 1}$, $I_{1\varepsilon 2}$, $I_{1\varepsilon 3}$, respectively. It's obvious that $I_{1\varepsilon 2}$ and $I_{2\varepsilon 3}$ are $o_x(1)$ because of the boundedness of $L(F)$ and $U(F)$, Lemma 5.1.1, and Lebesgue's dominated convergence theorem.

Conditions on f and w, the mean value theorem, and Lemma 5.1.1 imply that

$$
\begin{aligned}
I_{1\varepsilon_1} =& \frac{1}{\varepsilon}\Big(\int_{U(F)}^{U(F_\varepsilon)}(y-T)w(D(y,F))dF(y) - \int_{L(F)}^{L(F_\varepsilon)}(y-T)w(D(y,F))dF(y)\Big) \\
=& (\theta_{2\varepsilon}-T)w(D(\theta_{2\varepsilon},F))f(\theta_{2\varepsilon})(IF(x;U(F))+o_x(1)) \\
& -(\theta_{1\varepsilon}-T)w(D(\theta_{1\varepsilon},F))f(\theta_{1\varepsilon})(IF(x;L(F))+o_x(1)) \\
=& (U-T)w(D(U,F))f(U)IF(x;U(F)) \\
& -(L-T)w(D(L,F))f(L)IF(x;L(F))+o_x(1)
\end{aligned}
$$

where $\theta_{2\varepsilon}$ is a point between $U(F)$ and $U(F_\varepsilon)$, and $\theta_{1\varepsilon}$ between $L(F)$ and $L(F_\varepsilon)$. The desired result now follows. □

PROOF OF THEOREM 2.3.3. The proof is similar to that of Theorem 2.3.1 and we adopt the same notation. Let $\mathrm{w}(D(y,F_\varepsilon))-\mathrm{w}(D(y,F)) = \mathrm{w}^{(1)}(\theta(y,F_\varepsilon))(D(y,F_\varepsilon)-D(y,F))$. We need the following lemma whose proof is omitted here.

Lemma 5.1.2. *Under the conditions of Theorem 2.3.3, we have*

(*a*) $\sup_{y\in\mathrm{R}}|\,\mathrm{w}^{(1)}(\theta(y,F_\varepsilon))-\mathrm{w}^{(1)}(D(y,F))| = o_x(1)$;

(*b*) $\sup_{y\in\mathrm{R}}|y\,\mathrm{w}^{(1)}(\theta(y,F_\varepsilon))| < \infty$ *for sufficiently small* $\varepsilon > 0$;

(*c*) $(D(y,F_\varepsilon)-D(y,F))/\varepsilon = IF(x;D(y,F))+yo_x(1)+o_x(1)$.

First we write

$$
\begin{aligned}
T_w(F_\varepsilon)-T_w(F) \;=\; & \frac{1}{\int \mathrm{w}(D(y,F_\varepsilon))dF_\varepsilon(y)}\Big[\int_{L(F_\varepsilon)}^{U(F_\varepsilon)}\mathrm{w}(D(y,F_\varepsilon))(y-T_w)dF_\varepsilon(y) \\
& +\int_{-\infty}^{L(F_\varepsilon)}\mathrm{w}(D(y,F_\varepsilon))(L(F_\varepsilon)-T_w))dF_\varepsilon(y) \\
& +\int_{U(F_\varepsilon)}^{\infty}\mathrm{w}(D(y,F_\varepsilon))(U(F_\varepsilon)-T_w)dF_\varepsilon(y)\Big]
\end{aligned}
$$

Lebesgue's dominated convergence theorem implies immediately that

$$
\int \mathrm{w}(D(y,F_\varepsilon))dF_\varepsilon(y) \;=\; \int \mathrm{w}(D(y,F))dF(y)+o_x(1) \tag{5.1.5}
$$

We now focus on the numerator. Call the three terms $I_i(F_\varepsilon,y)$, $i=1,2,3$, respectively. By

the proof of Theorem 2.3.1, we see immediately that

$$\begin{aligned}\frac{1}{\varepsilon}I_1(F_\varepsilon,y) =&(U-T_w)\,\mathrm{w}(\beta)f(U)IF(x;U)-(L-T_w)\,\mathrm{w}(\beta)f(L)IF(x;L)\\ &+\int_L^U (y-T_w)\,\mathrm{w}^{(1)}(D(y,F))h(x,y)dF(y)+I(L\le x\le U)(x-T_w)\,\mathrm{w}(D(x,F))\\ &+\frac{1-\varepsilon}{\varepsilon}\int_L^U (y-T_w)\,\mathrm{w}(D(y,F))dF(y)+o_x(1). \qquad (5.1.6)\end{aligned}$$

Now it suffices to treat $I_2(F_\varepsilon,y)$. Following the proof of Theorem 2.3.1 and employing Lemmas 5.1.1 and 5.1.2, we have

$$\begin{aligned}\frac{1}{\varepsilon}I_2(F_\varepsilon,y) &= I(x<L)\,\mathrm{w}(D(x,F))(L-T_w)+IF(x;L)\int_{-\infty}^L \mathrm{w}(D(y,F))dF(y)\\ &\quad+(L-T_w)\int_{-\infty}^L \mathrm{w}^{(1)}(D(y,F))h(x,y)dF(y)+(L-T_w)\,\mathrm{w}(\beta)f(L)IF(x;L)\\ &\quad+\frac{1-\varepsilon}{\varepsilon}\int_{-\infty}^L (L-T_w)\,\mathrm{w}(D(y,F))dF(y)+o_x(1). \qquad (5.1.7)\end{aligned}$$

Likewise we have

$$\begin{aligned}\frac{1}{\varepsilon}I_3(F_\varepsilon,y) &= I(x>U)\,\mathrm{w}(D(x,F))(U-T_w)+IF(x;U)\int_U^\infty \mathrm{w}(D(y,F))dF(y)\\ &\quad+(U-T_w)\int_U^\infty \mathrm{w}^{(1)}(D(y,F))h(x,y)dF(y)-(U-T_w)\,\mathrm{w}(\beta)f(U)IF(x;U)\\ &\quad+\frac{1-\varepsilon}{\varepsilon}\int_U^\infty (U-T_w)\,\mathrm{w}(D(y,F))dF(y)+o_x(1). \qquad (5.1.8)\end{aligned}$$

Combining the last four displays, we have the desired result. □

PROOF OF THEOREM 2.4.1. For the sake of convenience, we define

$$\nu_n=\sqrt{n}\,(F_n-F),\quad H_n(\cdot)=\sqrt{n}\,(D(\cdot,F_n)-D(\cdot,F)) \qquad (5.1.9)$$

The following result, a special version of a general result in Zuo (2003), is needed in the proof.

Lemma 5.1.3. *Assume that $F'=f$ exists at μ and continuous in small neighborhoods of $\mu\pm\sigma$ with $f(\mu)$ and $f(\mu+\sigma)+f(\mu-\sigma)$ positive. Then for $0<\beta<\infty$ and any numbers $L_1<U_1$*

(a) $\sup_{x\in[L_1,U_1]}(1+|x|)|H_n(x)|=O_p(1)$; *and*

(b) $H_n(x)=\int IF(y,D(x,F))\nu_n(dy)+o_p(1)$, *uniformly over* $x\in[L_1,U_1]$.

PROOF: For $x \in [L_1, U_1]$, it is readily seen that

$$D(x, F_n) - D(x, F) = -(D(x, F)(\sigma_n - \sigma) + (\mu_n - \mu))/\sigma_n.$$

(a) follows immediately since the given conditions allow asymptotic representations for both μ_n and σ_n (see, e.g., page 92 of Serfling (1980)), which lead to (b). □

The proof of the theorem follows the lines given in Zuo (2003). First, we can write

$$\sqrt{n}\,(T_n - T) = \sqrt{n}\int_{L_n}^{U_n}(y - T)w(D(y, F_n))dF_n(y)\Big/\int_{L_n}^{U_n} w(D(y, F_n))dF_n(y) \quad (5.1.10)$$

and the numerator then can be decomposed into three terms

$$I_{1n} = \sqrt{n}\int_{L_n}^{U_n}(y - T)w(D(y, F_n))F_n(dy) - \sqrt{n}\int_{L}^{U}(y - T)w(D(y, F_n))F_n(dy);$$

$$I_{2n} = \sqrt{n}\int_{L}^{U}(y - T)w(D(y, F_n))F_n(dy) - \sqrt{n}\int_{L}^{U}(y - T)w(D(y, F))F_n(dy);$$

$$I_{3n} = \sqrt{n}\int_{L}^{U}(y - T)w(D(y, F))F_n(dy).$$

It follows immediately that

$$\frac{I_{3n}}{\delta} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\ell_3(X_i). \quad (5.1.11)$$

For I_{2n}, we note that

$$\begin{aligned}
I_{2n} =& \sqrt{n}\int_{L}^{U}(y - T)w(D(y, F_n))F_n(dy) - \sqrt{n}\int_{L}^{U}(y - T)w(D(y, F))F_n(dy) \\
=& \int_{L}^{U}(y - T)w'(\theta_n(y))H_n(y)F_n(dy) \\
=& \int_{L}^{U}(y - T)w'(D(y, F))H_n(y)F_n(dy) \\
&+ \int_{L}^{U}(y - T)(w'(\theta_n(y)) - w'(D(y, F)))H_n(y)F_n(dy) \\
\triangleq& J_{1n} + J_{2n},
\end{aligned}$$

where $\theta_n(y)$ is a point between $D(y, F_n)$ and $D(y, F)$. For J_{2n}, by Lemma 5.1.3, we have

$$\begin{aligned}
J_{2n} =& \int_{L}^{U}(y - T)(w'(\theta_n(y)) - w'(D(y, F)))H_n(y)F_n(dy) \\
\leq& \int_{L}^{U}(|y| + |T|)|H_n(y)|(w'(\theta_n(y)) - w'(D(y, F)))F_n(dy) \;=\; o_p(1)
\end{aligned}$$

On the other hand, by Lemma 5.1.3, continuity of w' and boundedness of L and U, Fubini's Theorem and the central limit theorem, we obtain

$$\begin{aligned}
&\int_L^U (y-T)w'(D(y,F)))H_n(y)(F_n-F)(dy) \\
&= \int_L^U (y-T)w'(D(y,F)))\Big(\int IF(x;D(y,F))\nu_n(dx)\Big)(F_n-F)(dy)+o_p(1) \\
&= \frac{1}{\sqrt{n}}\int\int_L^U (y-T)w'(D(y,F))IF(x;D(y,F))\nu_n(dy)\nu_n(dx)+o_p(1) \\
&= \frac{1}{\sigma\sqrt{n}}\Big(-\int_L^U (y-T)w'(D(y,F))D(y,F)\nu_n(dy)\int IF(x;\sigma(F))\nu_n(dx)+ \\
&-\int_L^U (y-T)w'(D(y,F))\nu_n(dy)\int IF(x;\mu(F))\nu_n(dx)\Big)+o_p(1) \;=\; o_p(1),
\end{aligned}$$

which, in conjunction with Lemma 5.1.3 and the Fubini's Theorem, yields

$$\begin{aligned}
J_{1n} &= \int_L^U (y-T)w'(D(y,F)))H_n(y)F(dy)+o_p(1) \\
&= \int\Big(\int_L^U (y-T)w'(D(y,F)))IF(x;D(y,F))F(dy)\Big)\nu_n(dx)+o_p(1) \\
&= \delta\frac{1}{\sqrt{n}}\sum_{i=1}^n \ell_2(X_i)+o_p(1).
\end{aligned}$$

Hence

$$\frac{I_{2n}}{\delta} = \frac{1}{\sqrt{n}}\sum_{i=1}^n \ell_2(X_i)+o_p(1). \tag{5.1.12}$$

For I_{1n}, we note that

$$\begin{aligned}
I_{1n} &= \sqrt{n}\int_{L_n}^{U_n}(y-T)w(D(y,F_n))F_n(dy)-\sqrt{n}\int_L^U (y-T)w(D(y,F_n))F_n(dy) \\
&= \sqrt{n}\int_{L_n}^{L}(y-T)w(D(y,F_n))F_n(dy)+\sqrt{n}\int_U^{U_n}(y-T)w(D(y,F_n))F_n(dy) \\
&\triangleq V_{1n}+V_{2n}
\end{aligned}$$

Now we deal with V_{1n} only since V_{2n} can be treated similarly. By mean value theorem,

$$\begin{aligned}
V_{1n} &= \sqrt{n}\int_{L_n}^{L}(y-T)w(D(y,F_n))F(dy)+\int_{L_n}^{L}(y-T)w(D(y,F_n))\nu_n(dy) \\
&= -(\eta_n-T)w(D(\eta_n,F_n))f(\eta_n)\sqrt{n}\,(L_n-L)+\int_{L_n}^{L}(y-T)w(D(y,F_n))\nu_n(dy),
\end{aligned}$$

where η_n is a point between L_n and L. Note that by the conditions given, we have

$$\begin{aligned}&(\eta_n - T)w(D(\eta_n, F_n))f(\eta_n)\sqrt{n}(L_n - L)\\ &= (L - T)w(D(L, F))f(L)\frac{1}{\sqrt{n}}\sum_{i=1}^{n} IF(X_i; L(F)) + o_p(1).\end{aligned}$$

Since $P(X = L) = 0$, it is readily seen that for large n and $L^* = -1 - |L|$,

$$\begin{aligned}&\int_{L_n}^{L}(y - T)w(D(y, F_n))\nu_n(dy)\\ &= -\int[\mathbf{I}_{(L^*, L_n)}(y) - \mathbf{I}_{(L^*, L)}(y)](y - T)w(D(y, F_n))\nu_n(dy) + o_p(1) = o_p(1),\end{aligned}$$

by an empirical process theory argument; see Pollard (1984) or van der Vaart and Wellner (1996). Thus

$$V_{1n} = -(L - T)w(D(L, F))f(L)\frac{1}{\sqrt{n}}\sum_{i=1}^{n} IF(X_i; L(F)) + o_p(1),$$

which, combining with a similar result from V_{2n}, gives

$$\frac{I_{1n}}{\delta} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\ell_1(X_i) + o_p(1). \tag{5.1.13}$$

In the same but much less involved manner we can show that

$$\int_{L_n}^{U_n} w(D(x, F_n))F_n(dx) = \int_{L}^{U} w(D(x, F))F(dx) + O_p(1/\sqrt{n}). \tag{5.1.14}$$

Now (5.1.11), (5.1.12), (5.1.13) and (5.1.14) give the desired result. □

PROOF OF THEOREM 2.4.3. The proof is very similar to that of Theorem 2.4.1. We adopt the notation in the proof of Theorem 2.4.1. Let $\mathrm{w}(D(y, F_n)) - \mathrm{w}(D(y, F)) = \mathrm{w}^{(1)}(\theta(y, F_n))(D(y, F_n) - D(y, F))$. We need the following lemma whose proof is skipped here.

Lemma 5.1.4. *Under the conditions of Theorem 2.4.3, we have*

(*a*) $\sup_{y\in\mathrm{R}}(1 + |y|)|\,\mathrm{w}^{(1)}(\theta(y, F_n)) - \mathrm{w}^{(1)}(D(y, F))| = o_p(1)$; *and*

(*b*) $H_n(y) = yO_p(1) + O_p(1)$.

We first write

$$T_w(F_n) - T_w(F) = \frac{1}{\int \mathrm{w}(D(y,F_n))dF_n(y)}\Big[\int_{L(F_n)}^{U(F_n)} \mathrm{w}(D(y,F_n))(y-T_w)dF_n(y)$$
$$+\int_{-\infty}^{L(F_n)} \mathrm{w}(D(y,F_n))(L(F_n)-T_w))dF_n(y)$$
$$+\int_{U(F_n)}^{\infty} \mathrm{w}(D(y,F_n))(U(F_n)-T_w)dF_n(y)\Big]$$

The given conditions guarantee that $\mathrm{w}(D(y_n,F_n)) \to \mathrm{w}(D(y,F))$ a.s. for every $y \in \mathrm{R}$ and every sequence $y_n \to y$. Skorokhod representation theorem and Lebesgue's dominated convergence theorem imply immediately that

$$\int \mathrm{w}(D(y,F_n))dF_n(y) = \int \mathrm{w}(D(y,F))dF(y) + o(1), \quad a.s. \tag{5.1.15}$$

We now focus on the numerator. Call the three terms $I_i(F_n,y)$, $i=1,2,3$, respectively. By the proof of Theorem 2.4.1, we see immediately that

$$I_1(F_n,y) = \frac{1}{n}\sum_{i=1}^{n}\Big((U-T_w)\,\mathrm{w}(\beta)f(U)IF(X_i;U) + I(L \le X_i \le U)(X_i - T_w)\,\mathrm{w}(D(X_i,F))$$
$$+\int_L^U (y-T_w)\,\mathrm{w}^{(1)}(D(y,F))h(X_i,y)dF(y) - (L-T_w)\,\mathrm{w}(\beta)f(L)IF(X_i;L)\Big)$$
$$+o_p(n^{-1/2}). \tag{5.1.16}$$

Now it suffices to treat $I_2(F_n,y)$. Following the proof of Theorem 2.4.1 and employing Lemmas 5.1.3 and 5.1.4, we have

$$I_2(F_n,y) = \frac{1}{n}\sum_{i=1}^{n}\Big(I(X_i < L)\,\mathrm{w}(D(X_i,F))(L-T_w) + IF(X_i;L)\int_{-\infty}^{L} \mathrm{w}(D(y,F))dF(y)$$
$$+(L-T_w)\int_{-\infty}^{L} \mathrm{w}^{(1)}(D(y,F))h(X_i,y)dF(y) + (L-T_w)\,\mathrm{w}(\beta)f(L)IF(X_i;L)\Big)$$
$$+o_p(n^{-1/2}). \tag{5.1.17}$$

Likewise we have

$$I_3(F_n,y) = \frac{1}{n}\sum_{i=1}^{n}\Big(I(X_i > U)\,\mathrm{w}(D(X_i,F))(U-T_w) + IF(X_i;U)\int_{U}^{\infty} \mathrm{w}(D(y,F))dF(y)$$
$$+(U-T_w)\int_{U}^{\infty} \mathrm{w}^{(1)}(D(y,F))h(X_i,y)dF(y) - (U-T_w)\,\mathrm{w}(\beta)f(U)IF(X_i;U)\Big)$$
$$+o_p(n^{-1/2}). \tag{5.1.18}$$

Combining the last four displays, we have the desired result. □

5.2 Selected proofs for results of chapter 3

Proof of Theorem 3.3.2. For simplicity we sometimes write F_ε for $F(\varepsilon, \delta_x)$ for a given $x \in R$. Denote by $o_x(1)$ a quantity that may depend on x but approaches 0 as $\varepsilon \to 0$ for the given x. We need the following results whose proof is omitted.

Lemma 5.2.1. *For fixed $x \in$ R and sufficiently small ε, we have*

(a) $D(y, F)$ and $D(y, F(\varepsilon, \delta_x))$ are Lipschitz continuous in $y \in$ R;

(b) $\sup_{x \in S} |D(y, F(\varepsilon, \delta_x)) - D(y, F)| = o_x(1)$ for any bounded set $S \in R$;

(c) $|L(F_\varepsilon) - L(F)| = o_x(1)$, $|U(F_\varepsilon) - U(F)| = o_x(1)$.

Proof. We only need to show that

$$IF(x; s(F)) = \tau_1(x) + \tau_2(x) + \tau_3(x)$$

Step 1: First we write

$$s(F_\varepsilon) - s(F) = \frac{\int_{L(F_\varepsilon)}^{U(F_\varepsilon)} (y^2 - s)\, \mathrm{w}_2(D(y, F_\varepsilon)) dF_\varepsilon(y)}{\int_{L(F_\varepsilon)}^{U(F_\varepsilon)} \mathrm{w}_2(D(y, F_\varepsilon)) dF_\varepsilon(y)} \tag{5.2.1}$$

The continuity of $\mathrm{w}^{(1)}(\cdot)$ and Lemma 5.2.1 yields

$$\mathrm{w}_2(D(y, F_\varepsilon) = \mathrm{w}_2(D(y, F) + o_x(1), \quad \text{uniformly in } y \text{ for the given } x.$$

Therefore, we have

$$\begin{aligned}
&\int_{L(F_\varepsilon)}^{U(F_\varepsilon)} \mathrm{w}_2(D(y, F_\varepsilon)) dF_\varepsilon(y) \\
&= (1 - \varepsilon) \int_{L(F_\varepsilon)}^{U(F_\varepsilon)} \mathrm{w}_2(D(y, F)) dF(y) + \varepsilon \mathbf{I}_{\{x \in [L(F_\varepsilon), U(F_\varepsilon)]\}} \mathrm{w}_2(D(y, F_\varepsilon)) + o_x(1), \\
&= \int_{L(F)}^{U(F)} \mathrm{w}_2(D(y, F)) dF(y) + o_x(1)
\end{aligned}$$

Step 2: For the numerator of equation (5.2.1), it can clearly be discomposed in three terms

$$\begin{aligned}
I_{1\varepsilon} &= \Big(\int_{L(F_\varepsilon)}^{U(F_\varepsilon)} (y^2 - s)\, \mathrm{w}_2(D(y, F_\varepsilon)) dF_\varepsilon(y) - \int_{L(F)}^{U(F)} (y^2 - s)\, \mathrm{w}_2(D(y, F_\varepsilon)) dF_\varepsilon(y) \Big) \\
I_{2\varepsilon} &= \Big(\int_{L(F)}^{U(F)} (y^2 - s)\, \mathrm{w}_2(D(y, F_\varepsilon)) dF_\varepsilon(y) - \int_{L(F)}^{U(F)} (y^2 - s)\, \mathrm{w}_2(D(y, F)) dF_\varepsilon(y) \Big) \\
I_{3\varepsilon} &= \int_{L(F)}^{U(F)} (y^2 - s)\, \mathrm{w}_2(D(y, F)) dF_\varepsilon(y)
\end{aligned}$$

It follows immediately that

$$\frac{1}{\varepsilon}I_{3\varepsilon} = \mathbf{I}_{\{x\in[L(F),U(F)]\}}(x^2 - s)\,\mathrm{w}_2(D(x,F)), \quad \text{uniformly in } y \text{ for the given } x. \tag{5.2.2}$$

In light of the continuity of $\mathrm{w}_2^{(1)}(\cdot)$ and Lemma 5.2.1, we have that

$$\mathrm{w}_2(D(y,F_\varepsilon)) - \mathrm{w}_2(D(y,F)) = (\mathrm{w}_2^{(1)}(D(y,F)) + o_x(1))(D(y,F_\varepsilon)) - D(y,F)), \tag{5.2.3}$$

uniformly in y for y in a bounded set S for the given x. This, combining with the boundedness of $L(F)$ and $U(F)$ and Lemma 5.2.1, immediately gives

$$\frac{1}{\varepsilon}I_{2\varepsilon} = \int_{L(F)}^{U(F)}(y^2 - s)\,\mathrm{w}_2^{(1)}(D(y,F))IF(x;D(y,F))dF(y) + o_x(1) \tag{5.2.4}$$

In virtue of equation (5.2.3), boundedness of $L(F)$ and $U(F)$, Lemma 5.2.1, and the argument used above, we have for sufficiently small $\varepsilon > 0$

$$\begin{aligned}\frac{1}{\varepsilon}I_{1\varepsilon} = &\frac{1}{\varepsilon}\int \triangle(y,\varepsilon)(y^2 - s)\,\mathrm{w}_2(D(y,F))dF(y) - \int \triangle(y,\varepsilon)(y^2 - s)\,\mathrm{w}_2(D(y,F))dF(y)\\ &+ \int \triangle(y,\varepsilon)(y^2 - s)\,\mathrm{w}_2^{(1)}(D(y,F))h(x,y)dF(y) + o_x(1)\end{aligned}$$

where $\triangle(y,\varepsilon) = \mathbf{I}(x \in [L(F_\varepsilon),U(F_\varepsilon)] - \mathbf{I}(x \in [L(F),U(F)])$. Call the three terms with integration $I_{1\varepsilon_i}$, $i = 1,2,3$ respectively. It's obvious that $I_{1\varepsilon_2}$ and $I_{2\varepsilon_3}$ are $o_x(1)$ because of the boundedness of $L(F)$) and $U(F)$, Lemma 5.2.1, and Lebesgue's dominated convergence theorem. Conditions on f and w_2, the mean value theorem and Lemma 5.2.1 imply that

$$\begin{aligned}I_{1\varepsilon_1} =&\frac{1}{\varepsilon}\int(\mathbf{I}_{\{x\in[L(F_\varepsilon),U(F_\varepsilon)]\}} - \mathbf{I}_{\{x\in[L(F),U(F)]\}})(y^2 - s)\,\mathrm{w}_2(D(y,F))dF(y)\\ =&\frac{1}{\varepsilon}\Big(\int_{U(F)}^{U(F_\varepsilon)}(y^2 - s)\,\mathrm{w}_2(D(y,F))dF(y) - \int_{L(F)}^{L(F_\varepsilon)}(y^2 - s)\,\mathrm{w}_2(D(y,F))dF(y)\Big)\\ =&(\theta_{2\varepsilon}^2 - s)\,\mathrm{w}_2(D(\theta_{2\varepsilon},F))f(\theta_{2\varepsilon})(IF(x,U(F)) + o_x(1))\\ &- (\theta_{1\varepsilon}^2 - s)\,\mathrm{w}_2(D(\theta_{1\varepsilon},F))f(\theta_{1\varepsilon})(IF(x,L(F)) + o_x(1))\\ =&(U(F)^2 - s)\,\mathrm{w}_2(D(U,F))f(U)IF(x,U(F))\\ &- (L(F)^2 - s)\,\mathrm{w}_2(D(L,F))f(L)IF(x,L(F)) + o_x(1)\end{aligned}$$

where $\theta_{2\varepsilon}$ is a point between $U(F)$ and $U(F_\varepsilon)$, and $\theta_{1\varepsilon}$ between $L(F)$ and $L(F_\varepsilon)$. Therefore the desired result now follows. □

Proof of Theorem 3.3.5

Proof. The proof is very similar to that of Theorem 3.3.2. We adopt the notation in the proof of Theorem 3.3.2. Let $w_2(D(y,F_n)) - w_2(D(y,F)) = w_2^{(1)}(\theta(y,F_n))(D(y,F_n) - D(y,F))$. We need the following facts whose proofs are omitted here.

Lemma 5.2.2. *Under the conditions of Theorem 3.3.5, we have*

(a) $\sup_{y\in R}(1+y^2)(w_2^{(1)}(\theta(y,F_\varepsilon)) - w_2^{(1)}(D(y,F))) = o_x(1)$; *and*

(b) $\sup_{y\in R}(1+y^2)|y^2 w_2^{(1)}(\theta(y,F_\varepsilon))| < \infty$ *for sufficiently small* $\varepsilon > 0$;

(c) $(D(y,F_\varepsilon) - D(y,F))/\varepsilon = IF(x,D(y,F)) + yo_x(1) + o_x(1)$.

First we write

$$\begin{aligned} s_w(F_\varepsilon) - s_w(F) = \frac{1}{\int \mathrm{w}_2(D(y,F_\varepsilon))dF_{\varepsilon(y)}}\Big[&\int_{L(F_\varepsilon)}^{U(F_\varepsilon)} \mathrm{w}_2(D(y,F_\varepsilon))(y^2 - s_w)dF_\varepsilon(y) \\ &+ \int_{-\infty}^{L(F_\varepsilon)} \mathrm{w}_2(D(y,F_\varepsilon))(L^2(F_\varepsilon) - s_w))dF_\varepsilon(y) \\ &+ \int_{U(F_\varepsilon)}^{\infty} \mathrm{w}_2(D(y,F_\varepsilon))(U^2(F_\varepsilon) - s_w)dF_\varepsilon(y)\Big] \end{aligned} \tag{5.2.5}$$

Lebesgue's dominated convergence theorem implies immediately that

$$\int \mathrm{w}_2(D(y,F_\varepsilon))dF_\varepsilon(y) = \int \mathrm{w}_2(D(y,F))dF(y) + o_x(1) \tag{5.2.6}$$

Step 2: We now focus on the numerator of equation (5.2.5), Call three terms $I_i(F_\varepsilon,y)$, $i = 1, 2, 3$, respectively. By the proof of Theorem 3.4.1, we see immediately that

$$\begin{aligned} \frac{1}{\varepsilon}I_1(F_\varepsilon,y) =&(U^2 - s_w)\,\mathrm{w}_2(\beta)f(U)IF(x;U) - (L^2 - s_w)\,\mathrm{w}_2(\beta)f(L)IF(x;L) \\ &+ \int_L^U (y^2 - s_w)\,\mathrm{w}_2^{(1)}(D(y,F))h(x,y)dF(y) + I(L \le x \le U)(x^2 - s_w)\,\mathrm{w}_2(D(x,F)) \\ &+ \frac{1-\varepsilon}{\varepsilon}\int_L^U (y^2 - s_w)\,\mathrm{w}_2(D(y,F))dF(y) + o_x(1). \end{aligned} \tag{5.2.7}$$

Now it suffices to treat $I_2(F_\varepsilon,y)$. Following the proof of Theorem 3.4.1 and employing Lemma 5.2.1 and 5.2.3, we have

$$\begin{aligned} \frac{1}{\varepsilon}I_2(F_\varepsilon,y) = &(L^2 - s_w)\mathbf{I}(x < L)\,\mathrm{w}(P(x,F)) + (L^2 - s_w)\,\mathrm{w}_2(D(L,F))IF(x;L) \\ &+ (L^2 - s_w)\int_{-\infty}^{L(F)} \mathrm{w}_2^{(1)}(D(y,F))h(x,y)dF(y) \\ &+ 2L(F)IF(x,L(F))\int_{-\infty}^{L(F)} \mathrm{w}_2(D(y,F))dF(y) \end{aligned} \tag{5.2.8}$$

$$+\frac{1-\varepsilon}{\varepsilon}\int_L^U (L^2-s)w(D(y,F))dF(y)+o_x(1) \tag{5.2.9}$$

Likewise we have

$$\begin{aligned}\frac{1}{\varepsilon}I_3(F_\varepsilon,y) &= (U^2-s_w)\mathbf{I}(x>U)\,\mathrm{w}(P(x,F)) - (U^2-s_w)\,\mathrm{w}_2(D(U,F))IF(x;U)\\ &\quad+(U^2-s_w)\int_{U(F)}^{\infty}\mathrm{w}_2^{(1)}(D(y,F))h(x,y)dF(y)\\ &\quad+2U(F)IF(x,U(F))\int_{U(F)}^{\infty}\mathrm{w}_2(D(y,F))dF(y) \qquad (5.2.10)\\ &\quad+\frac{1-\varepsilon}{\varepsilon}\int_L^U (U^2-s)w(D(y,F))dF(y)+o_x(1) \qquad (5.2.11)\end{aligned}$$

Combining the last four displays, we have the desired result. □

Proof of Theorem 3.4.1

Proof. For sake of convenience, we define

$$\nu_n=\sqrt{n}(F_n-F),\quad H_n(\cdot)=\sqrt{n}(D_n(\cdot,F_n)-D(\cdot,F)) \tag{5.2.12}$$

The following result, is needed in the proof.

Lemma 5.2.3. *Assume that $F'=f$ exists at μ and continuous in small neighborhoods of $\mu\pm\sigma$ with $f(\mu)$ and $f(\mu+\sigma)+f(\mu-\sigma)$ positive. Then for $0<\beta<\infty$, we have*
(a) $\sup_{x\in[L,U]}(1+x^2)|H_n(x)|=O_p(1)$; *and*
(b) $H_n(x)=\int h(y,x)\nu_n(dy)+o_p(1)$, *uniformly on* $x\in[L(F),U(F)]$.

Proof. For $x\in[L,U]$, it is readily seen that

$$D(x,F_n)-D(x,F)=-(D(x,F)(\sigma_n-\sigma)+(\mu_n-\mu))/\sigma_n.$$

(a) follows immediately since the given conditions allow asymptotic representations for both μ_n and σ_n (see, e.g., page 92 of Serfling (1980)), which lead to (b). □

Since $S^2(F)$ can be breaken into two parts ($s(F)$, $\lambda(F)$) and the proof of the two parts are similar, we only prove

$$\begin{aligned} s(F_n) - s(F) &= \frac{1}{n}\sum_{i=1}^{n} IF(X_i; s(F)) + o_p(\frac{1}{\sqrt{n}}) \\ &= \frac{1}{n}\sum_{i=1}^{n} \tau_1(X_i) + \frac{1}{n}\sum_{i=1}^{n} \tau_2(X_i) + \frac{1}{n}\sum_{i=1}^{n} \tau_3(X_i) + o_p(\frac{1}{\sqrt{n}}). \end{aligned}$$

First, observe that

$$\sqrt{n}(s(F_n) - s(F)) = \sqrt{n}\int_{L_n}^{U_n} (y^2 - s) w_2(D(y, F_n)) F_n(dy) / \int_{L_n}^{U_n} w_2(D(y, F_n)) F_n(dy) \tag{5.2.13}$$

and the numerator then can be decomposed into three terms

$$\begin{aligned} I_{1n} &= \sqrt{n}\int_{L_n}^{U_n} (y^2 - s) w_2(D(y, F_n)) F_n(dy) - \sqrt{n}\int_{L}^{U} (y^2 - s) w_2(D(y, F_n)) F_n(dy) \\ I_{2n} &= \sqrt{n}\int_{L}^{U} (y^2 - s) w_2(D(y, F_n)) F_n(dy) - \sqrt{n}\int_{L}^{U} (y^2 - s) w_2(D(y, F)) F_n(dy) \\ I_{3n} &= \sqrt{n}\int_{L}^{U} (y^2 - s) w_2(D(y, F)) F_n(dy) \end{aligned}$$

It follows immediately that

$$I_{3n} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \tau_3(X_i). \tag{5.2.14}$$

For I_{2n}, we note that

$$\begin{aligned} I_{2n} =& \sqrt{n}\int_{L}^{U} (y^2 - s)\, \mathrm{w}_2(D(y, F_n)) F_n(dy) - \sqrt{n}\int_{L}^{U} (y^2 - s)\, \mathrm{w}_2(D(y, F)) F_n(dy) \\ =& \int_{L}^{U} (y^2 - s)\, \mathrm{w}_2^{(1)}(\theta_n(y)) H_n(y) F_n(dy) \\ =& \int_{L}^{U} (y^2 - s)\, \mathrm{w}_2^{(1)}(D(y, F)) H_n(y) F_n(dy) \\ &+ \int_{L}^{U} (y^2 - s)(\mathrm{w}_2^{(1)}(\theta_n(y)) - \mathrm{w}_2^{(1)}(D(y, F))) H_n(y) F_n(dy) \\ \stackrel{\Delta}{=}& J_{1n} + J_{2n} \end{aligned}$$

where $\theta_n(y)$ is a point between $D(y, F_n)$ and $D(y, F)$. For J_{2n}, by using Lemma 5.2.3, we

have

$$\begin{aligned} J_{2n} &= \int_L^U (y^2 - s)(\mathrm{w}_2^{(1)}(\theta_n(y)) - \mathrm{w}_2^{(1)}(D(y,F)))H_n(y)F_n(dy) \\ &\leq \int_L^U (y^2 + s)|H_n(y)|(\mathrm{w}_2^{(1)}(\theta_n(y)) - \mathrm{w}_2^{(1)}(D(y,F)))F_n(dy) \\ &= o_p(1) \end{aligned}$$

On the other hand, by Lemma 6.3, continuity of $w_2^{(1)}$ and boundedness of L and U, Fubini's Theorem and the central limit theorem, we obtain

$$\begin{aligned} &\int_L^U (y^2 - s)\,\mathrm{w}_2^{(1)}(D(y,F)))H_n(y)(F_n - F)(dy) \\ &= \int_L^U (y^2 - s)\,\mathrm{w}_2^{(1)}(D(y,F)))\Big(\int h(x,y)\nu_n(dx) + o_p(1)\Big)(F_n - F)(dy) \\ &= \frac{1}{\sqrt{n}}\int\int_L^U (y^2 - s)\,\mathrm{w}_2^{(1)}(D(y,F)))IF(x,D(y,F))\nu_n(dy)\nu_n(dx) + o_p(1) \\ &= \frac{1}{\sigma\sqrt{n}}\Big(-\int_L^U (y^2 - s)\,\mathrm{w}_2^{(1)}(D(y,F))D(y,F)\nu_n(dy)\int IF(x,\sigma(F))\nu_n(dx) \\ &- \int_L^U (y^2 - s)w_2^{(1)}(D(y,F))\nu_n(dy)\int IF(x,\mu(F))\nu_n(dx)\Big) + o_p(1) \\ &= o_p(1) \end{aligned}$$

which, in conjunction with Lemma 5.2.3 and the Fubini's Theorem, yields

$$\begin{aligned} J_{1n} &= \int_L^U (y^2 - s)w_2^{(1)}(D(y,F)))H_n(y)F(dy) + o_p(1) \\ &= \int\Big(\int_L^U (y^2 - s)w_2^{(1)}(D(y,F)))h(x,y)F(dy)\Big)\nu_n(dx) + o_p(1) \\ &= \delta_2\frac{1}{\sqrt{n}}\sum_{i=1}^n \tau_2(X_i) + o_p(1). \end{aligned}$$

Hence

$$\frac{\mathrm{I}_{2n}}{\delta_2} = \frac{1}{\sqrt{n}}\sum_{i=1}^n \ell_2(X_i) + o_p(1) \tag{5.2.15}$$

For I_{1n}, note that

$$\begin{aligned} I_{1n} &= \sqrt{n}\int_{L_n}^{U_n} (y^2 - s)\,\mathrm{w}_2(D(y,F_n))F_n(dy) - \sqrt{n}\int_L^U (y^2 - s)\,\mathrm{w}_2(D(y,F_n))F_n(dy) \\ &= \sqrt{n}\int_{L_n}^{L} (y^2 - s)\,\mathrm{w}_2(D(y,F_n))F_n(dy) + \sqrt{n}\int_U^{U_n} (y^2 - s)\,\mathrm{w}_2(D(y,F_n))F_n(dy) \\ &\triangleq V_{1n} + V_{2n} \end{aligned}$$

Next we only deal with V_{1n} since V_{2n} can be treated similarly. By mean value theorem,

$$\begin{aligned} V_{1n} &= \sqrt{n}\int_{L_n}^{L}(y^2-s)\,\mathrm{w}_2(D(y,F_n))F_n(dy) \\ &= \sqrt{n}\int_{L_n}^{L}(y^2-s)\,\mathrm{w}_2(D(y,F_n))F(dy) + \int_{L_n}^{L}(y^2-s)\,\mathrm{w}_2(D(y,F_n))\nu_n(dy) \\ &= -(\eta_n^2-s)\,\mathrm{w}_2(D(\eta_n,F_n))f(\eta_n)\sqrt{n}(L_n-L) + \int_{L_n}^{L}(y^2-s)\,\mathrm{w}_2(D(y,F_n))\nu_n(dy) \end{aligned}$$

where η_n is a point between L_n and L. Note that by the conditions given, we have

$$(\eta_n^2-s)\,\mathrm{w}_2(D(\eta_n,F_n))f(\eta_n)\sqrt{n}(L_n-L) = (L^2-s)\,\mathrm{w}_2(D(L,F))f(L)\frac{1}{\sqrt{n}}\sum_{i=1}^{n} IF(X_i,L(F))+o_p(1),$$

Since $P(X=L)=0$, it is readily seen that for large n and $L^* = -1-|L|$, ,

$$\begin{aligned} &\int_{L_n}^{L}(y^2-s)\,\mathrm{w}_2(D(y,F_n))\nu_n(dy) \\ &= \int[\mathbf{I}_{\{(L^*,L_n)\}}(y)-\mathbf{I}_{\{(L^*,L_n)\}}(y)](y^2-s)w_2(D(y,F_n))\nu_n(dy)+o_p(1) \\ &= o_p(1), \end{aligned}$$

by an empirical process theory argument; See Pollard (1984) or van der Vaart and Wellner (1986). Thus

$$V_{1n} = -(L^2-s)w_2(D(L,F))f(L)\frac{1}{\sqrt{n}}\sum_{i=1}^{n} IF(X_i,L(F))+o_p(1), \tag{5.2.16}$$

which combining with a similar results from V_{2n}, gives

$$\frac{\mathbf{I}_{1n}}{\delta} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\ell_1(X_i)+o_p(1) \tag{5.2.17}$$

In the same but much less involved manner, we can show that

$$\int_{L_n}^{U_n}\mathrm{w}_2(D(x,F))F_n(dx) = \int_{L}^{U}\mathrm{w}_2(D(x,F))F(dx)+O_p(1/\sqrt{n}) \tag{5.2.18}$$

Now (5.2.14), (5.2.15), (5.2.17) and (5.2.18) give the desired result. □

Proof of Theorem 3.4.3. The proof is very similar to that of Theorem 3.4.1. We adopt the notation in the proof of Theorem 3.4.1. Let $w(D(y,F_n)) - w(D(y,F)) = w^{(1)}(\theta(y,F_n))(D(y,F_n)-D(y,F))$. We need the following lemma whose proof is skipped here.

Lemma 5.2.4. *Under the conditions of Theorem 3.4.3, we have*
(a) $\sup_{y\in R}(1+y^2)|\,\mathrm{w}_2^{(1)}(\theta(y,F_n)) - \mathrm{w}_2^{(1)}(\mathrm{w}_2(D(y,F)))| = o_p(1)$; *and*
(b) $H_n(y) = yO_p(1) + O_p(1)$.

We first write

$$\begin{aligned} s_{wn} - s_w = \frac{1}{\int \mathrm{w}_2(D(y,F_n))dF_n(y)}\Big[&\int_{L_n}^{U_n} \mathrm{w}_2(D(y,F_n))(x^2 - s_w)dF_n(x) \\ &+ \int_{-\infty}^{L_n} \mathrm{w}_2(D(y,F_n))(L^2(F_n) - s_w))dF_n(x) \\ &+ \int_{U_n}^{\infty} \mathrm{w}_2(D(y,F_n))(R_{2n}^2 - s_w)dF_n(x)\Big] \end{aligned}$$

The given conditions guarantee that $w(D(y_n,F_n)) \to w(D(y,F))$ *a.s.* for every $y \in R$ and every sequence $y_n \to y$. Skorohod representation theorem and Lebesgue's dominated convergence theorem imply immediately that

$$\int w(D(y,F_n))dF_n(y) = \int \mathrm{w}_2(D(y,F))dF(x) \tag{5.2.19}$$

We now focus on the numerator. Call the three terms $\mathrm{I}_i(F_n,y)$, $i = 1, 2, 3$, respectively. By the proof of Theorem 3.4.1, we see immediately that

$$\mathrm{I}_1(F_n,y) = \frac{1}{n}\sum_{i=1} n(\int_{L(F)}^{U(F)} (y^2 - s_w)\,\mathrm{w}_2^{(1)}(D(y,F))h(X_i,y)dF(y)) + o_p(n^{-1/2}) \tag{5.2.20}$$

Now it suffices to treat $\mathrm{I}_2(F_n,y)$. Following the proof of Theorem 3.4.1, and employing Lemmas 5.2.3, 5.2.4; we have

$$\begin{aligned} \mathrm{I}_2(F_n,y) =& \frac{1}{n}\sum_{i=1} n(L^2 - s_w)\mathrm{I}(X_i < L)\,\mathrm{w}(P(X_i,F)) \\ &+ (L^2 - s_w)\int_{-\infty}^{L(F)} \mathrm{w}_2^{(1)}(D(y,F))h(X_i,y)dF(y) \\ &+ 2L(F)IF(X_i,L(F))\int_{-\infty}^{L(F)} \mathrm{w}_2(D(y,F))dF(y) + o_p(n^{-1/2}) \end{aligned} \tag{5.2.21}$$

Likewise we have

$$\mathrm{I}_3(F_n,y) = (U^2 - s_w)\mathrm{I}(X_i > U)\,\mathrm{w}_2(D(X_i,F)) \tag{5.2.22}$$

$$+ (U^2 - s_w) \int_{U(F)}^{\infty} \mathrm{w}_2^{(1)}(D(y,F))h(X_i,y)dF(y)$$

$$+ 2U(F)IF(X_i,U(F)) \int_{U(F)}^{\infty} \mathrm{w}_2(D(y,F))dF(y) + o_p(n^{-1/2}) \qquad (5.2.23)$$

Combining the last four displays, we have the desired result.

BIBLIOGRAPHY

[1] Agullo, J., Croux, C., and Van Aelst, S. (2002). The multivariate least trimmed squares estimator. Submitted.

[2] Bai, Z.D., Chen, N.R., Miao, B.Q., and Rao, C.R. (1990), Asymptotic Theory of Least Distance Estimate in Multivariate Linear Models, Statistics, 21, 503-519.

[3] Bickel, P. J. (1965). On some robust estimates of location. *Ann. Math. Statist.* **36** 847-858.

[4] Donoho, D. L., and Huber, P. J. (1983). The notion of breakdown point. In *A Festschrift for Erich L. Lehmann* (P. J. Bickel, K. A. Doksum and J. L. Hodges, Jr., eds.) 157-184. Wadsworth, Belmont, CA.

[5] Jaeckel, L. A. (1971). Some flexible estimates of location. *Ann. Math. Statis.* **42** 1540-1552.

[6] Jurečková, J., Koenker, R. and Welsh, A. H. (1994). Adaptive choice of trimming proportions. *Ann. Inst. Statist. Math.* **46** 737-755.

[7] Hampel, F. R., Ronchetti, E. Z., Rousseeuw, P. J. and Stahel, W. A. (1986). Robust Statistics: The approach based on influence function. Wiley, New York.

[8] Hogg, R. V. (1974). Adaptive robust Procedures: A Partial review and some suggestions for future applications and theory. *J. Amer. Statist. Assoc.* **69** 909-923.

[9] Kim, S. (1992). The metrically trimmed mean as a robust estimator of location. *Ann. Statist.* **20** 1534-1547.

[10] Koenker, R., Portnoy, S. (1990). M-estimation of multivariate regressions. Journal of the American Statistical Association, 85, 1060-1068.

[11] Marrona, R.A., and Yohai, V.J. (1997), Robust Estimation in Simultaneous Equations Models, Journal of Statistical Planning and Inference, 57, 233-244.

[12] Pollard, D. (1984). *Convergence of Stochastic Processes.* Springer-Verlag, New York.

[13] Rousseeuw, P.J. (1984). Least median of squares regression. Journal of the American Statistical Association, 79, 871-880.

[14] Rousseeuw, P.J., Van Aelst, S., Van Driessen, K., and Agullo, J. (2001). Robust multivariate regression. submitted.

[15] Rousseeuw, P.J. and Van Driessen, K. (2002), Computing LTS Regression for Large Data Sets, Estadistica, 54, 163-190.

[16] Serfling, R. (1980). *Approximation Theorems of Mathematical Statistics*, John Wiley & Sons, New York.

[17] Shorack, G. R. (1974). Random Means. *Ann. Statist.* **2** 661-675.

[18] Peter J. Rousseeuw and Christophe Croux. Alternative to the Median Absolute Deviation. J. AM. Stat. Assoc. 88, 1993.

[19] Stigler, S. M. (1973). The asymptotic distribution of the trimmed mean. *Ann. Statist.* **1** 472-477.

[20] Stigler, S. M. (1977). Do robust estimators work with real data? *Ann. Statist.* **5** 1055-1077.

[21] Tukey, J. W. (1948). Some elementary problems of importance to small sample practice. *Human Biology* **20** 205-214.

[22] van der Vaart, A. W., and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes With Applications to Statistics.* Springer.

[23] Welsh, AH; Morrison, HL Robust L estimation of scale with an application in astronomy. J. Amer. Statist. Assoc. 85 (1990), no. 411, 729–743. 62F35.

[24] Zuo, Y. (2003). Projection depth trimmed means for multivariate data: robustness and efficiency (the latest version: Multi-dimensional trimming based on projection depth was tentatively accepted by the Annals of Statistics in 2004).