Contributions to the Theory of Robust Inference

by

Matías Salibián-Barrera
Licenciado en Matemáticas, Universidad de Buenos Aires, Argentina, 1994

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

Doctor of Philosophy

in

THE FACULTY OF GRADUATE STUDIES

(Department of Statistics)

we accept this thesis as conforming
to the required standard

The University of British Columbia

July 2000

Department of STATISTICS

The University of British Columbia
Vancouver, Canada

Date JULY 20th 2000.

Abstract

We study the problem of performing statistical inference based on robust estimates when the distribution of the data is only assumed to belong to a contamination neighbourhood of a known central distribution. We start by determining the asymptotic properties of some robust estimates when the data are not generated by the central distribution of the contamination neighbourhood. Under certain regularity conditions the considered estimates are consistent and asymptotically normal. For the location model and with additional regularity conditions we show that the convergence is uniform on the contamination neighbourhood. We determine that a class of robust estimates satisfies these requirements for certain proportions of contamination, and that there is a trade-off between the robustness of the estimates and the extent to which the uniformity of their asymptotic properties holds. When the distribution of the data is not the central distribution of the neighbourhood the asymptotic variance of these estimates is involved and difficult to estimate. This problem affects the performance of inference methods based on the empirical estimates of the asymptotic variance. We present a new re-sampling method based on Efron's bootstrap (Efron, 1979) to estimate the sampling distribution of MM-location and regression estimates.

This method overcomes the main drawbacks of the use of bootstrap with robust estimates on large and potentially contaminated data sets. We show that our proposal is computationally simple and that it provides stable estimates when the data contain outliers. This new method extends naturally to the linear regression model.

Contents

List of Tables

List of Figures

Acknowledgements

Many people helped me during the years I spent at the Department.

No line of this thesis could have been written without my wife Verónica D'Angelo by my side and our son Lucas constantly teaching me what is really important.

Financial support was provided by the University of British Columbia through a University Graduate Fellowship; by the Department of Statistics via part-time appointments as Teaching Assistant; by Dr. Ruben Zamar with appointments as Research Assistant during many summers and by the Statistical Consultant and Research Laboratory with a part-time position.

I owe much more than what I can express here to our families back home (Alfredo, Marisa and Florencia Salibián, Oscar and Noemí D'Angelo) who always believed in us and unconditionally supported us all these years; to our friend Sonia Mazzi, who made our settling in this remote land a smooth and pleasant adventure; to the Palejkos (George, Gill, Nancy, Ingrid, and Roger) who became our family in Canada; and to our good friends Daniel and Iris Brendle-Moczuk, Geneviève Gagné, Stephane Lemieux (and Eugenie!), Rémi Gilbert, Caroline Grondin (Victor and Zacharias), the Hollmans (Jorge, Sandra and Rocío), Jorge Adrover, Héctor Palacio, Luis Suero, and Carlos Di Bella.

Many people at the Department also helped in different ways: Jarek Harezlak, being a good friend and colleague; Rick White, who was always ready for a chat; Christine Graham (what would we do without her?); Dr. Nancy Heckman and Dr. John Petkau, who on many occasions spent some time with the students at the pub; and last but not least: everybody at "The L. Gang".

Finally, I would like to thank Dr. Ruben Zamar for his encouragement, constant availability, patience, guidance and support; Dr. Nancy Heckman and Dr. John Petkau for their helpful comments, suggestions and advice; and Dr. Jim Zidek for his advice.

MATÍAS SALIBIÁN-BARRERA

The University of British Columbia
July 2000

For Lucas and Vero

Chapter 1

Introduction

In this chapter we introduce and illustrate the problems addressed in the rest of this thesis. In the first section we use a real-life example to show how an analysis based on a robust regression estimate compares with previous analyses of these data. In those analyses the data were carefully screened, suspicious observations were deleted, and least squares methods were used on the remaining data.

Unfortunately there are no proposals in the literature to consistently estimate the variability of the estimates obtained after deleting potential outliers. The second section of this chapter explores this problem.

An alternative method to perform statistical inference when the data are contaminated is to use robust estimates. Most attention in the robust literature has been paid to the case of errors with symmetric distributions. Section 1.3 briefly reviews some of the published studies for asymmetric distributions. In the same section we discuss our results regarding the asymptotic properties of some robust estimates for more general error distributions. In particular we study their consistency and asymptotic distribution. Empirical estimates of these asymptotic variances provide consistent estimates of the variability of these robust estimates. Unfortunately, simulation experiments suggest that they can be numerically unstable and hence yield poor estimates.

We also consider computer-intensive inference methods, in particular Efron's bootstrap (Efron, 1979). In Section 1.4 we discuss two drawbacks of the use of Efron's bootstrap with robust estimates. Both the presence of outliers in the data and the computational complexity of robust estimates are important challenges for this method.

In Section 1.5 we introduce a new computer-intensive method that overcomes these limitations and hence can be used with large data sets that might contain outliers. In this section we present the basic idea for the simple location model with known scale. Details of the application of this method to the location-scale and linear regression models are presented in Chapters 3 and 5 respectively.

Finally, Section 1.6 outlines the rest of this thesis.

1.1 Robust estimates and data screening

Consider the Stack Loss data, first published by Brownlee (1965, page 454). These data have been extensively studied in the literature (see Daniel and Wood, 1980, Chapters 5 and 7; Atkinson, 1985, pp. 129-136, 267-8; and Venables and Ripley, 1997, pp 262-264). They consist of 21 daily observations measured in a plant for the oxidation of ammonia to nitric acid. The response variable is ten times the percentage of ammonia lost. This is an indirect measure of the efficiency of the plant. There are 3 explanatory variables: air flow, temperature of the cooling water and acid concentration.

The linear model used in the literature is

$$\begin{aligned}\text{Ammonia Lost (\%)} = \beta_0 + \beta_1\,\text{Air flow} + \beta_2\,\text{Water temperature}\\ + \beta_3\,\text{Acid concentration} + \epsilon\,, \quad (1.1)\end{aligned}$$

where ϵ are independent identically distributed normal errors. The residuals of the least squares fit of model (1.1) presented some features worth further consideration. After a very careful analysis of the listing of the data, Daniel and Wood (1971, Chapter 5, page 81) noticed a different behavior in the response variable every time the water temperature was above 60. They concluded that the plant seemed to have needed a period of one day to stabilize after the water temperature reached 60. Hence they decided that the observations that were obtained with water temperature above 60 (cases 1, 3, 4 and 21) require special attention, and were removed from the analysis.

Figure 1.1 contains the plot of residuals versus fitted values for the least squares fit. The dotted lines correspond to twice the estimated standard deviation of the errors in (1.1). Note that observation number 21 appears to have a residual considerably larger than the others. Three other cases are somewhat outlying, but within 2 estimated standard deviations from zero. Classical outlier detection methods, such as the externally Studentized residuals test (Weisberg, 1985, page 115-6) only detect observation 21 as an outlier.

We also estimated the coefficients of model (1.1) using an MM-regression estimate with 50% breakdown point, 95% efficiency and Tukey's loss functions (see Sections 4.1 and 4.2 for the corresponding definitions). We worked with the complete data set. The plot of residuals versus fitted values is shown in Figure 1.2. The dotted lines correspond to twice the estimated standard deviation of the errors. With this robust estimate cases 1, 3, 4 and 21 are clearly identified as outliers.

This example illustrates the potential of robust estimates. Daniel and Wood (1980) had to rely on an careful analysis of the listing of the data until some pattern seemed apparent. The additional complications and limitations of this method when the data have either more explanatory variables or more cases are obvious. In this example the analysis based on a robust regression estimate yields the same conclusion as Daniel and Wood (1980, Chapter 5), namely that observations 1, 3, 4 and 21 seem to follow a different model from the rest of the data. Note that we did not require a detailed case-by-case analysis as Daniel and Wood did (1980, Chapter 5).

From the discussion above one might conclude that the main role of robust estimates is to help to identify outliers or suspicious observations. These cases could

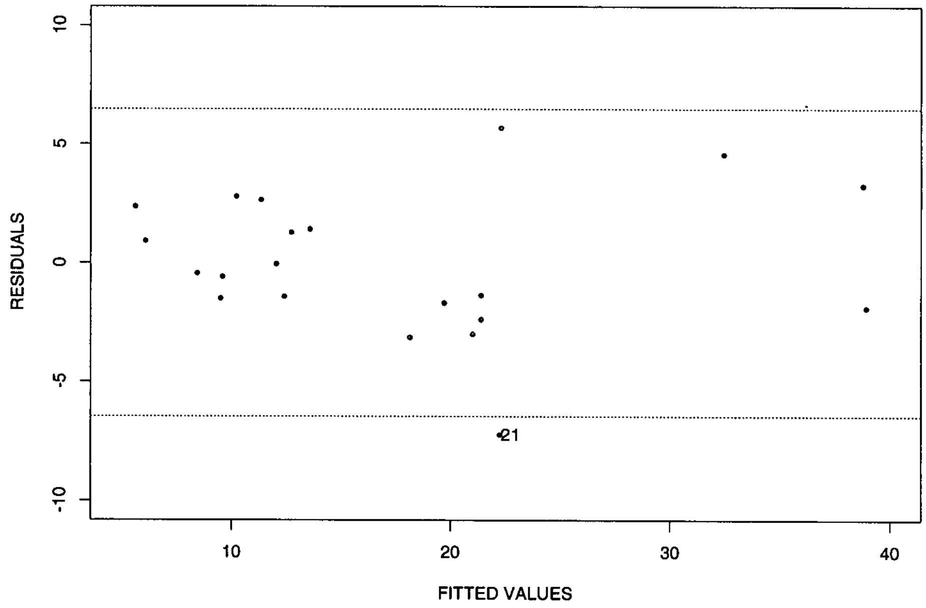


Figure 1.1: Residuals of the least squares fit for the Stack Loss data. The dotted lines correspond to twice the estimated standard deviation of the errors

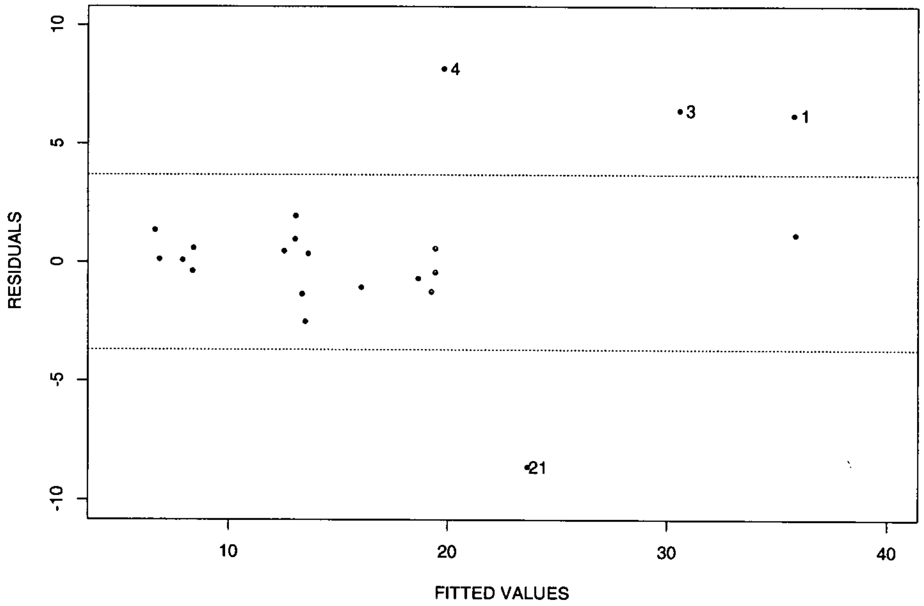


Figure 1.2: Residuals of a robust fit for the Stack Loss data. The dotted lines correspond to twice the estimated standard deviation of the errors

then be discarded and classical methods applied to the "clean" data set. In the next section we discuss some drawbacks of this approach.

1.2 The variability caused by cleaning the data

There are two classes of methods to detect outliers: "subjective" and "objective" procedures. In this section we will focus on outlier detection methods applied to linear regression analyses.

Subjective methods rely on the judgment of data analysts. They normally use a classical fit followed by an analysis of the residuals. Using plots and other devices the researcher identifies outliers or suspicious observations. These observations are then removed and classical methods applied to the remaining data.

A formal study of the variability introduced into the final least squares estimate by these data-cleaning methods seems impossible with the mathematical tools available today (but see Relles and Rogers (1977) for a Monte Carlo experiment on subjective outlier rejection rules).

On the other hand, objective methods are based on a well defined rule, such as: "discard all observations with a residual larger than 2 standard deviations", or "reject all observations with associated Cox Distance larger than 1". Because it is expected that if there are outliers in the data the classical fit will be misleading, another set of objective rules are based on the residuals from a robust fit as follows:

1. Fit a robust estimate.
2. Calculate a robust estimate of the standard deviation of the residuals, $\hat{\sigma}$.
3. Fix a number $c > 0$ and drop any observation with a residual larger than $c\,\hat{\sigma}$ (typically $2 \leq c \leq 3$).
4. Apply classical methods to the remaining data.

We will refer to this last family of methods as "hard rejection rules" (HRR). See Hampel *et al.* (1986, page 31) for a Monte Carlo study of objective rejection rules for the location model.

If we apply steps 1-4 to the Stack Loss data with the same MM-regression estimate we used before, and set $c = 2$ in step 3, we find that observations 1, 3, 4 and 21 should be removed. The least squares fit of the remaining 17 data points yields regression estimates that are indistinguishable from the MM-regression fit. However, the estimates of the standard deviations of the regression estimates given by the least squares analysis are consistently smaller than those reported by the robust procedure (see Table 1.1).

It is important to note that the standard errors of the estimates reported by the least squares analysis of the "cleaned" data do not take into account the variability introduced by the "cleaning" step. In other words, the column of estimated standard deviations in the computer output may not reflect the actual variability of the reported point estimates.

	Estimated Standard Deviations	
Coefficient	LS on the "cleaned" data	Robust fit
Intercept	4.732	5.003
Air flow	0.067	0.071
Water temp.	0.166	0.176
Acid conc.	0.062	0.065
Residuals	1.253	1.837

Table 1.1: Comparison of the estimated standard deviations of the linear regression estimates for the Stack Loss data

To illustrate this problem we performed a small Monte Carlo experiment (also see Dupuis and Hamilton (2000) for a theoretical assessment of this inference procedure). The objective of the experiment is to show that the estimates of the standard deviations of the regression estimates calculated by the HRR method consistently underestimate the actual standard deviations of those regression estimates.

In order to do so, we first estimated the actual variability of the point estimates obtained by using a HRR. We considered a linear model of the form

$$y_i = \beta_0 + \beta_1 \, x_{1i} + \cdots + \beta_p \, x_{pi} + \epsilon_i \, , \qquad i = 1, \ldots, n \, , \tag{1.2}$$

where ϵ_i are independent standard normal random variables. Note that in the above model there are no outliers in the data. We used all the combinations with $n = 20$, $n = 50$, $p = 1$ and $p = 3$. The robust estimate used in the HRR procedure was a 95% efficient MM-regression estimate with 50% breakdown point and scale calculated with Tukey's loss function (see Sections 4.1 and 4.2 for the corresponding definitions). For each combination of sample size and number of predictors we generated 100,000 samples following model (1.2). With each sample we followed steps 1 to 4 above. In step 3 we used $c = 2.5$ and the robust estimate of the standard deviation of the errors associated with the MM-regression estimate. Our estimate of the actual variability of these estimates is the Monte Carlo standard deviation of these 100,000 coefficient estimates. In Table 1.2, the column labeled "Monte Carlo estimate of the standard deviation" contains this estimated standard deviation for each coefficient in the different models.

The next step in the experiment is to show that the estimates of those standard deviations as reported by the HRR analysis are consistently smaller that the estimates obtained in the first part of our study. With the same design matrices we generated 100,000 new samples following model (1.2). We applied steps 1 to 4 as before to each of these new samples, and recorded the estimates of the standard errors of each coefficient as reported by the least squares analysis in step 4. Column "HRR estimates of the standard deviation" in Table 1.2 contains the mean and standard error of these 100,000 estimated standard deviations.

From Table 1.2 it is clear that the estimates of the standard deviations reported by the least squares fit after cleaning the data consistently underestimate the actual variability of this estimation procedure. Hence we might obtain optimistic confidence intervals and smaller p-values than their actual value. The researcher should be concerned that this difference can affect the validity of his or her conclusions.

An alternative method of performing inference that can deal with outliers in the data is to use robust estimates. These methods naturally incorporate the variability of

	n	β	HRR estimates of the standard deviation	Monte Carlo estimate of the standard deviation
$p = 2$	20	β_0	0.205 (0.046)	0.256
		β_1	0.227 (0.051)	0.283
	50	β_0	0.133 (0.017)	0.152
		β_1	0.138 (0.017)	0.156
$p = 4$	20	β_0	0.182 (0.057)	0.322
		β_1	0.164 (0.051)	0.410
		β_2	0.173 (0.054)	0.478
		β_3	0.177 (0.056)	0.295
	50	β_0	0.135 (0.018)	0.159
		β_1	0.144 (0.019)	0.170
		β_2	0.145 (0.019)	0.171
		β_3	0.132 (0.018)	0.157

Table 1.2: Comparison of actual and estimated standard deviations using the HRR method to "clean" the data. The first column contains the Monte Carlo mean of the HRR estimates of the standard deviations and the corresponding Monte Carlo standard deviation within parentheses. The second column contains the estimate of the standard deviations obtained from a separate simulation experiment. These last values are the "actual" standard deviations.

the down-weighting step into the estimated standard deviations. In the next section we discuss some limitations of the existing asymptotic theory for robust estimates.

1.3 Inference based on robust estimates

The finite sample distribution of robust estimates is unknown and hence inference must be based on their asymptotic distribution (see Hampel *et al.*, 1986, Chapter 3; Ronchetti, 1982; Markatou and Hettmansperger, 1990; among others).

The asymptotic distribution of robust regression estimates is well known when the distribution of the errors is symmetric (Huber, 1967; Maronna and Yohai, 1981; Davies, 1993). In this case the estimates of the regression coefficients and of the scale of the errors are asymptotically independent.

Because outliers need not be balanced on both sides of the regression line, many data sets with outliers do not satisfy this symmetry assumption. If one relaxes this condition, the calculation of the asymptotic variance of robust location and regression estimates becomes very involved. The main difficulty seems to be that the scale estimate is no longer asymptotically independent of the estimate of the location or regression parameters. This problem has received little attention in the literature. Carroll (1979), Huber (1981) and Rocke and Downs (1981) are among the few who studied it. Carroll (1979) compared several variance estimates of both location and simple linear regression robust estimates. He showed that the asymptotic variance derived under the symmetry assumption underestimates the true variance. In the location case, this effect can be ameliorated by jackknifing. However, this technique does not seem to work for the intercept of the simple linear regression model. Huber (1981, page 140) gave a formula to compute the influence functions of location and scale estimates when they are calculated simultaneously. Rocke and Downs (1981) also studied variance estimation for robust location estimates when the distribution of the data is asymmetric. Their simulation study concluded that estimating the variance of robust location estimates in this situation is very difficult. In particular, for symmetric distributions the empirical estimate of asymptotic variance estimate worked better than the bootstrap, but for asymmetric distributions the performances reversed. Their numerical results do not show a variance estimation method that yields good estimates for both symmetric and asymmetric distributions.

In Sections 2.3 and 4.3 we study the consistency and asymptotic distribution of the S-scale, S- and MM-location and regression estimators (see Sections 2.1 and 4.1 for the corresponding definitions). We assume that the distribution of the errors belongs to a contamination neighbourhood of a symmetric central distribution and show that these estimates are consistent for any distribution in this neighbourhood. For the location-scale model, with further regularity conditions we show that these results hold uniformly on the neighbourhood. That is, the speed of the convergence does not depend on the particular distribution F in the contamination neighborhood $\mathcal{H}_\epsilon$ (see Section 2.2). Formally, the uniformity result we obtain is as follows. Let $\hat{\mu}_n$ be the robust location estimate calculated on a sample of size n generated by a distribution $F \in \mathcal{H}_\epsilon$. Let μ be the almost sure asymptotic value of $\hat{\mu}_n$ when $n \to \infty$.

Let $\delta > 0$ be arbitrary, then

$$\lim_{m\to\infty} \sup_{F\in\mathcal{H}_\epsilon} P_F\left\{\sup_{n\geq m} |\hat{\mu}_n - \mu| > \delta\right\} = 0\,.$$

We also find that under certain regularity conditions the MM-location estimates are asymptotically normal and we derive an explicit formula for their asymptotic variance. For the location model it has the form

$$V(\mu, \sigma, F) = \sigma^2 a^2 E_F\left\{[U - b \times W]^2\right\}, \qquad (1.3)$$

where U and W are certain random variables (see equation 2.70 on page 71), the constants a and b are given by

$$a = 1/E_F\left\{\psi'\left((X-\mu)/\sigma\right)\right\},$$

and

$$b = \frac{E_F\left\{\psi'\left((X-\mu)/\sigma\right)(X-\mu)/\sigma\right\}}{E_F\left\{\rho'\left((X-\tilde{\mu})/\sigma\right)(X-\tilde{\mu})/\sigma\right\}}\,,$$

where $\tilde{\mu}$ is the almost sure asymptotic value of the S-location estimate $\tilde{\mu}_n$ associated with the MM-location estimate $\hat{\mu}_n$. The functions ψ and ρ are bounded, and continuously differentiable (see Definition 2.9).

Another result we derive for the location-scale model is that the asymptotic normality of these estimates holds uniformly on the distribution generating the data. That is, we have

$$\lim_{n\to\infty} \sup_{F\in\mathcal{H}_\epsilon} \sup_{x\in\mathbb{R}} \left| P_F\left\{\sqrt{n}\,(\hat{\mu}_n - \mu) \, < x\sqrt{V}\right\} - \Phi(x)\right| = 0\,.$$

where Φ denotes the standard normal cumulative distribution function and $V = V(\mu, \sigma, F)$ is the asymptotic variance given by (1.3).

In general, consistent estimates for the asymptotic standard deviations of robust estimates can be obtained from the corresponding empirical asymptotic variances. For example, to estimate V above we can use $\hat{V} = V(\hat{\mu}_n, \hat{\sigma}_n, F_n)$ where F_n is the empirical distribution function of the sample $x_1, \ldots, x_n$. However, for the case of asymmetric error distributions, we found some numerical problems that seem to arise from the involved form of V in (1.3). In particular, the denominators in a and b can become small for asymmetric distributions F. In Section 3.6.1 we describe a Monte Carlo experiment that illustrates the extent of this instability.

1.4 Bootstrapping robust estimates

Another approach to estimate the variability of estimates is given by the bootstrap (Efron, 1979). This method has been extensively studied for diverse models. In particular, the theory for bootstrap distribution of robust estimates has been studied by Shorack (1982), Parr (1985) and Yang (1985) among others.

Two problems of practical relevance arise when bootstrapping robust regression estimates. First, the proportion of outliers in the bootstrap samples may be higher than that in the original data set causing the bootstrap quantiles to be very inaccurate. Intuitively the reasoning is as follows. Both outlying and non-outlying observations have the same chance of being in any bootstrap sample. With a certain positive probability, the proportion of outliers in a bootstrap sample will be larger than the fraction of contamination the robust estimate tolerates. In other words, a certain proportion of the re-calculated values of the robust estimate will be heavily influenced by the outliers in the data. Thus, the tails of the bootstrap distribution can be heavily influenced by the outliers.

This "lack of robustness" of the classical bootstrap was noted by Ghosh *et al.* (1984), and Shao (1990, 1992) in the context of estimating asymptotic variances, and by Singh (1998) for quantile estimates. Ghosh *et al.* (1984) showed that a condition is needed on the tails of the distribution of the data for the bootstrap variance estimate of the median to converge. Note that no matter how robust the estimate being bootstrapped, a tail condition is still needed. Shao (1990) proved that if one truncates the tails of the bootstrap distribution (with the truncation limit going to infinity as the sample size increases, so that asymptotically there are no discarded bootstrapped estimates) then the bootstrap variance converges to the asymptotic variance of the estimate of interest. Unfortunately it is not clear how to implement this method in a finite sample setting. Singh (1998) quantified this robustness problem for the estimates of the quantiles of the asymptotic distribution of location estimates. He defined the breakdown point for bootstrap quantiles and showed that it is disappointingly low even for highly robust location estimates. He proposed to Winsorize the observations using the robust location and scale estimates and then to re-sample from these Winsorized observations. He showed that the quantiles obtained from this method have a much higher breakdown point and that they converge to the quantiles of the asymptotic distribution of the estimate.

The second difficulty is caused by the heavy computational requirements of the bootstrap which are compounded with robust estimates. Robust regression estimates are generally determined by the solution of a non-linear optimization problem in several dimensions. In the particular case of MM-estimates (Yohai, 1987) for each sample we have to solve two such problems. Moreover, one of them is only implicitly defined as the solution of a non-linear equation. We see that bootstrapping MM-estimates involves repeatedly solving two non-linear optimization problems in several dimensions. We have also found additional computational issues that needed special attention. For example, a bootstrap sample may not be in general position (see Definition 5.1 in Section 5.4) and this has consequences in determining the scale of the residuals. This large number of non-linear optimization problems may render the method unfeasible for high dimensional problems. As an example of the computational times that can be expected, the evaluation of 5,000 bootstrap re-calculations of an MM-regression estimate on a simulated data set with 200 observations and 10 explanatory variables took 9120 CPU seconds (≈2.5 hours) on a Sun Sparc Ultra workstation. The same number of re-calculations performed with the robust bootstrap we introduce in the next section took 416 CPU seconds (approximately 7 minutes) under the same conditions.

1.5 A new computer intensive method

The basic ideas are best presented using the simple location model. Let $x_1, \ldots, x_n$ be a random sample satisfying

$$x_i = \mu + \epsilon_i, \qquad i = 1, \ldots, n, \tag{1.4}$$

where ϵ_i are independent and identically distributed random variables with known variance. Let $\psi : \mathbb{R} \to \mathbb{R}$ be odd, bounded, and non-decreasing. The associated M-location estimate for μ is defined as the solution $\hat{\mu}_n$ of

$$\sum_{i=1}^{n} \psi\left(x_i - \hat{\mu}_n\right) = 0. \tag{1.5}$$

We are interested in estimating the standard deviation of $\hat{\mu}_n$. For this purpose we present the following computer intensive method to generate a large number of re-calculated estimates $\hat{\mu}_n^*$. We will use the variability observed in these re-calculated estimates to assess the variance of $\hat{\mu}_n$.

It is easy to see that $\hat{\mu}_n$ can also be expressed as a weighted average of the observations:

$$\hat{\mu}_n = \frac{\sum_{i=1}^{n} \frac{\psi(x_i-\hat{\mu}_n)}{(x_i-\hat{\mu}_n)}\, x_i}{\sum_{i=1}^{n} \frac{\psi(x_i-\hat{\mu}_n)}{(x_i-\hat{\mu}_n)}} = \frac{\sum_{i=1}^{n} \omega_i\, x_i}{\sum_{i=1}^{n} \omega_i}, \tag{1.6}$$

where $\omega_i = \psi\left(x_i - \hat{\mu}_n\right) / \left(x_i - \hat{\mu}_n\right)$. This representation of $\hat{\mu}_n$ cannot be used directly to calculate $\hat{\mu}_n$ because the weights on the right hand side depend on the estimate.

Note that commonly used functions ψ (such as Huber's family ψ_c, see equation 2.5) yield weights $\omega\left(u\right) = \psi\left(u\right) / \, u$ that are decreasing functions of $|u|$. In this case, outlying observations that typically have a large residual $|x_i - \hat{\mu}_n|$ will have a small weight in (1.6).

Let $x_1^*, \ldots, x_n^*$ be a bootstrap sample of the data (i.e. a random sample taken from $x_1, \ldots, x_n$ with replacement). Recalculate $\hat{\mu}_n$ using equation (1.6):

$$\hat{\mu}_n^* = \frac{\sum_{i=1}^{n} \omega_i^*\, x_i^*}{\sum_{i=1}^{n} \omega_i^*}, \tag{1.7}$$

with $\omega_i^* = \psi\left(x_i^* - \hat{\mu}_n\right) / \left(x_i^* - \hat{\mu}_n\right)$. We have seen above that observations that are far from the bulk of the data will typically come into the bootstrap samples associated with small weights. Hence the influence of outliers in the bootstrapped estimate is bounded. Also note that we are not fully recalculating the estimate from each bootstrap sample.

The re-calculated $\hat{\mu}_n^*$'s in (1.7) may not reflect the actual variability of $\hat{\mu}_n$. Intuitively this happens because the weights ω_i are not re-computed with each bootstrap sample. Instead, we are using the weights obtained with the estimate $\hat{\mu}_n$ as calculated with the original data. To remedy this loss of variability in the $\hat{\mu}_n^*$'s we use an estimable correction factor. One way to derive this correction is to think of (1.6) as a fixed-point equation of the form $\hat{\mu}_n = f\left(\hat{\mu}_n\right)$. The first-order Taylor expansion of f around the limit μ of $\hat{\mu}_n$ suggests that we should multiply the re-weighted $\hat{\mu}_n$'s by $\left[1 - f'\left(\mu\right)\right]^{-1}$. With this notation, the correction factor we use is $\left[1 - f'\left(\hat{\mu}_n\right)\right]^{-1}$. Theorem 3.1 in Section 3.3 shows that the corrected $\hat{\mu}_n^*$'s have the same asymptotic distribution as the estimates $\hat{\mu}_n$.

Our method yields quantile estimates with a high breakdown point as defined by Singh (1998) (see Sections 3.4 and 5.4). This property means that a high proportion of outliers is needed to push the robust bootstrap quantile estimates above any bound. Classical bootstrap quantile estimates have a disappointingly low breakdown point, in spite of the robustness of the estimate being re-calculated (Singh, 1998). In Section 3 we study the robust bootstrap for the location model with unknown scale.

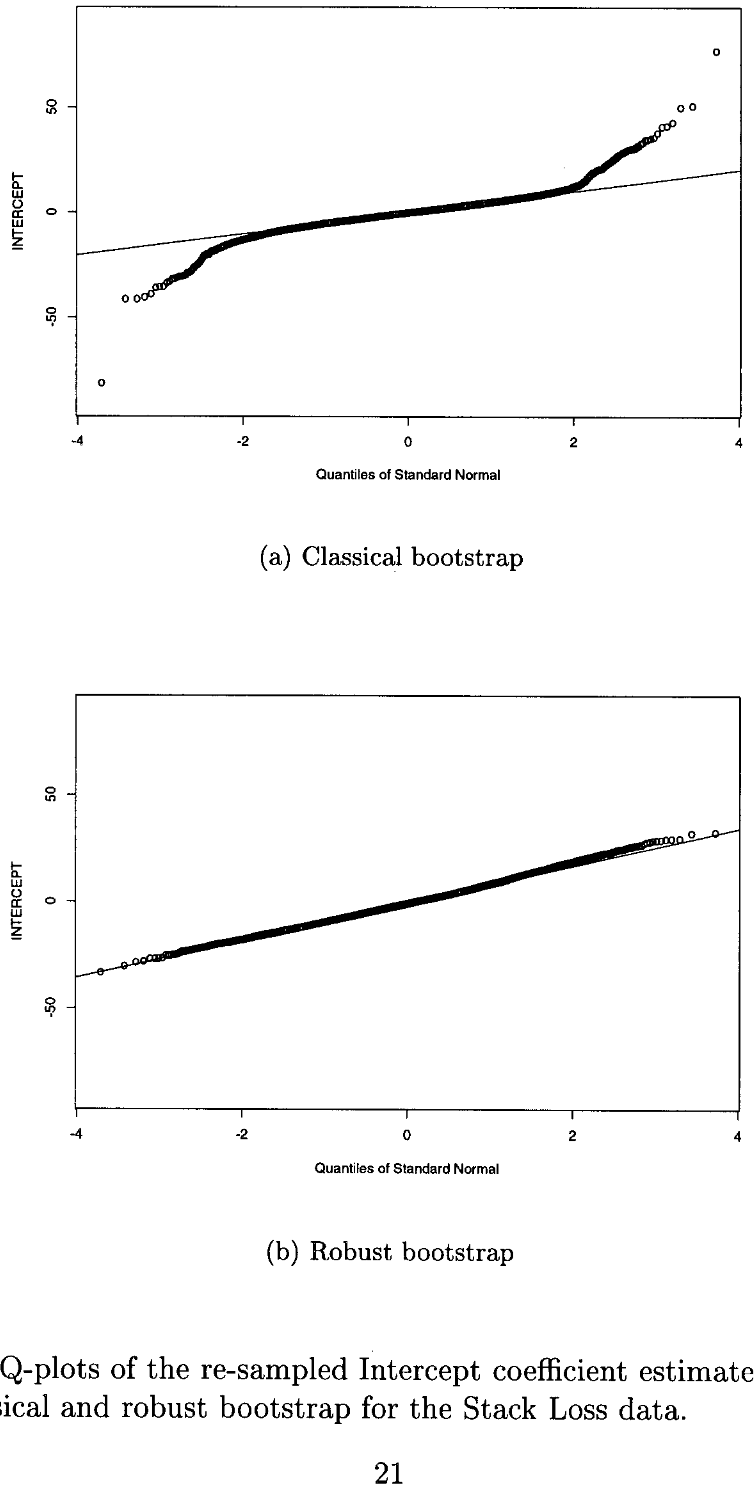
This new bootstrap method, which we call the robust bootstrap, is also computationally simple. In the linear regression context studied in Chapter 5, this property is very desirable. As opposed to the classical bootstrap that would need to solve a full multivariate optimization problem with each re-calculation, robust bootstrap evaluations only require solving a linear system of equations.

To compare the performance of our method with the classical bootstrap we ran 5,000 robust bootstrap iterations on the same artificial data set we used to illustrate the computational demands of the classical bootstrap (see page 17). Our method took 416 CPU seconds (approximately 7 minutes) to finish, while the classical bootstrap used 2.5 CPU hours. Both programs were written in C and called within Splus 3.4 for Unix.

To illustrate the stability of the distribution estimates obtained with the robust bootstrap, we applied our method to the MM-regression estimate for the Stack Loss data (see Chapter 4 for the definitions). We used both re-sampling methods to estimate the distribution of the 4-dimensional vector

$$\sqrt{n}\left[\left(\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3\right)' - \left(\beta_0,\beta_1,\beta_2,\beta_3\right)'\right],$$

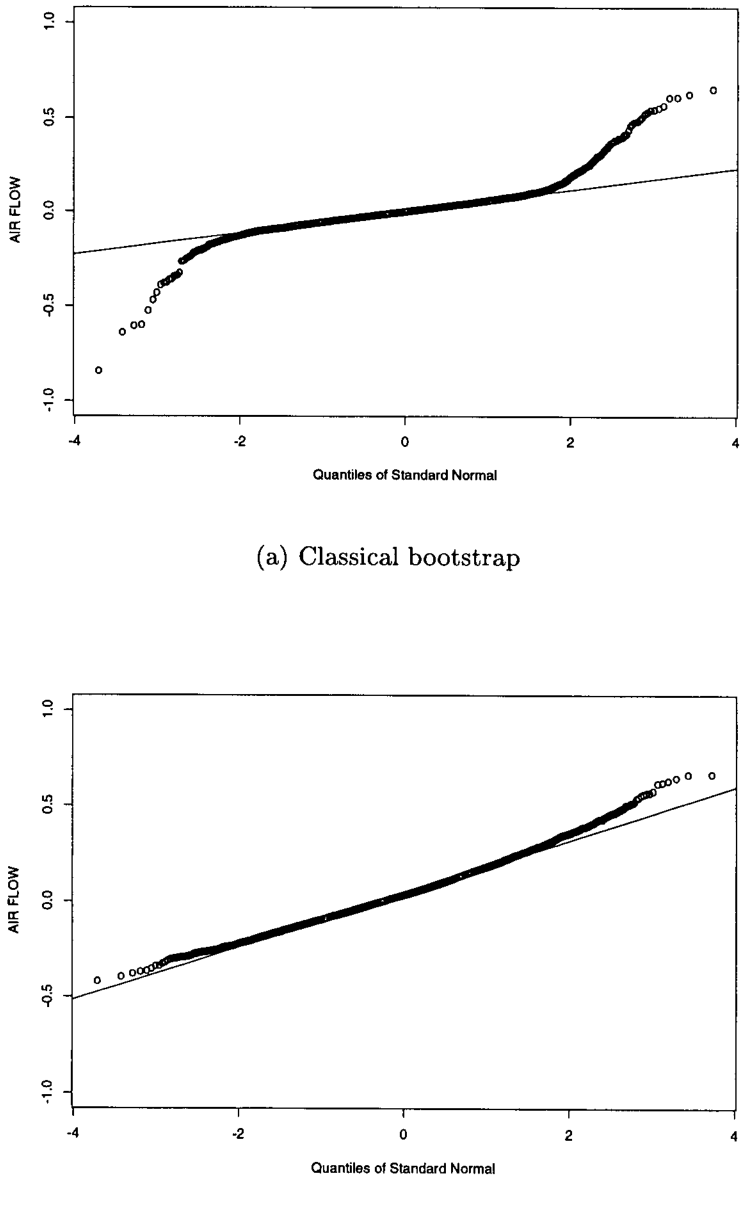
where $(\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3)$ is the MM-regression estimate. Figures 1.3 to 1.6 display the QQ-plots of the estimates of the distribution of the projections $(\hat{\beta}_i - \beta_i)$, $i = 1,\ldots,4$, obtained with each method. Note that in all cases the distribution estimates yielded by our method are closer to the limiting normal distribution and have lighter tails than the re-calculated estimates using the classical bootstrap.



(a) Classical bootstrap

(b) Robust bootstrap

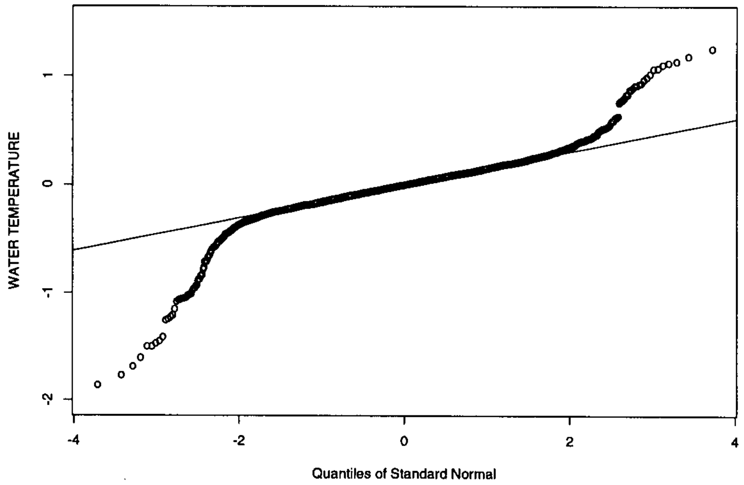
Figure 1.3: QQ-plots of the re-sampled Intercept coefficient estimates obtained with both the classical and robust bootstrap for the Stack Loss data.



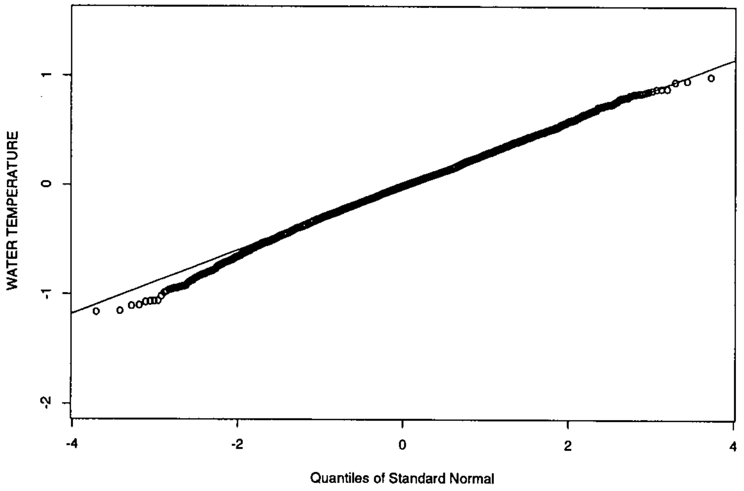
(a) Classical bootstrap

(b) Robust bootstrap

Figure 1.4: QQ-plots of the re-sampled Air Flow coefficient estimates obtained with both the classical and robust bootstrap for the Stack Loss data.

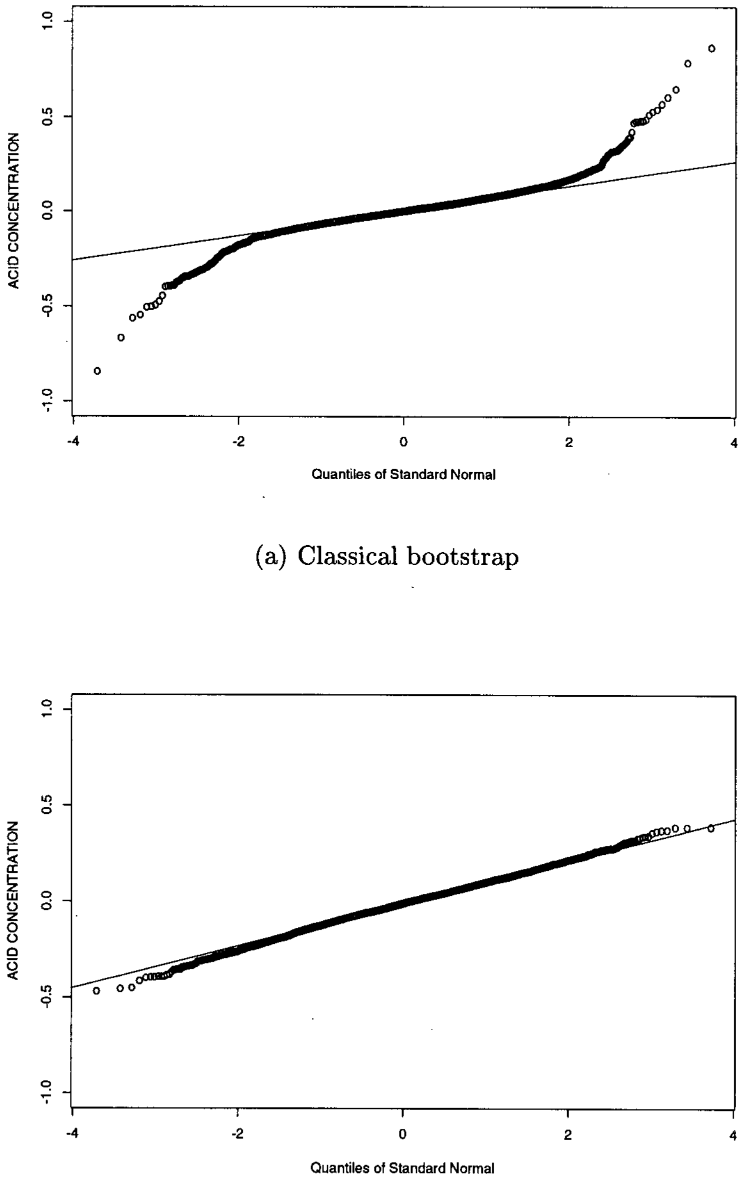


(a) Classical bootstrap



(b) Robust bootstrap

Figure 1.5: QQ-plots of the re-sampled Water Temperature coefficient estimates obtained with both the classical and robust bootstrap for the Stack Loss data.



(a) Classical bootstrap

(b) Robust bootstrap

Figure 1.6: QQ-plots of the re-sampled Acid Concentration coefficient estimates obtained with both the classical and robust bootstrap for the Stack Loss data.

1.6 Thesis outline

The rest of this thesis is organized as follows. Chapter 2 studies the asymptotic properties (consistency and asymptotic distribution) of some robust scale and location estimates. We introduce the location-scale model and the classes of S-scale, S-location and MM-location estimates. We study the asymptotic behaviour of these estimates when the distribution of the errors belongs to a contamination neighbourhood of the standard normal. We present consistency and asymptotic normality results that, under additional regularity conditions, hold uniformly on the distribution of the errors (see Davies, 1998). As a side result we derive a technique to determine the maximum asymptotic bias of M-location estimates with re-descending score functions.

In Chapter 3 we present a new computer intensive inference method for the location-scale model and we study its asymptotic properties. In particular we show that the resulting bootstrap distribution converges to the asymptotic distribution of the estimates of interest and that the derived quantile estimates have satisfactory robustness properties. Finally, we report the results of two Monte Carlo studies that compare the performance of this new method with other proposals in the literature. The first study compares several asymptotic variance estimates while the second compares the mean length and empirical coverage of confidence intervals for the parameters of interest in the model.

In Chapter 4 we extend the results of Chapter 2 to the linear regression model. Section 4.1 presents the model and defines the MM-regression estimates. Section 4.2

presents the contamination neighbourhood and the robustness properties of the MM-regression estimates. Section 4.3 studies the asymptotic properties of these estimates.

Chapter 5 extends the inference method presented in Chapter 3 to the linear regression model. We illustrate its use with two examples and we study the consistency of the distribution estimate and the robustness of the corresponding quantile estimates. Section 5.5 contains the results of a simulation study that investigates the finite sample size behaviour of the confidence intervals based on new method introduced here.

Chapter 6 contains a brief list of the results obtained in this thesis, the challenges that remain to be solved and the directions we forsee for future work.

The appendix in Chapter 7 contains most of the auxiliary results needed in the proofs. Proofs are presented for those results that could not be found in the literature.

Chapter 2

Global asymptotic properties of robust estimates for the location-scale model

In this chapter we study the asymptotic properties (consistency and asymptotic distribution) of some robust estimates of a location parameter when the observations may have an asymmetric distribution. First we define the classes of M-location estimates with general scale, S-location, S-scale estimates, and MM-location estimates. Most attention in the robustness literature has been paid to the asymptotic properties of robust estimates (in particular to their consistency and asymptotic distribution) when the data follow the non-contaminated model. In this chapter we study the properties of S- and MM-estimates in the full contamination neighbourhood. We show that the

S- and MM-estimates are consistent and asymptotically normal for any distribution in the gross-error neighbourhood of a symmetric distribution. We also discuss conditions that ensure these results hold uniformly over the contamination neighbourhood. As discussed by Davies (1998), uniformity is a reasonable property to expect in this context. Robust estimates have been proposed to deal with uncertainty in the model that generates the data, hence we expect their properties not to depend on a specific distribution in the neighbourhood. For example, the speed of convergence can depend on the distribution that generated the data. Our results guarantee that this is not the case with the estimates we consider in this chapter.

This chapter is organized as follows. Section 2.1 defines the classes of M-, S- and MM-location and scale estimates. Section 2.2 introduces the contamination neighbourhood $\mathcal{H}_\epsilon$ and briefly discusses the robustness properties of these estimates. Section 2.3 studies the asymptotic properties of S-location, S-scale and MM-location estimates when the distribution of the data belongs to $\mathcal{H}_\epsilon$. We provide conditions to obtain consistency and asymptotic normality of these estimates. We also obtain regularity conditions that ensure that the consistency and asymptotic normality results hold uniformly on the contamination neighbourhood. We show that a certain family of robust estimates satisfies these conditions.

2.1 Definitions

In this chapter we consider the following location-scale model. Let $x_1, \ldots, x_n$ be n observations on the real line satisfying

$$x_i = \mu + \sigma\,\epsilon_i \qquad i = 1, \ldots n, \tag{2.1}$$

where ϵ_i, $i = 1, \ldots n$ are independent and identically distributed (i.i.d.) observations with variance equal to 1. The interest is in estimating μ. The scale σ is a nuisance parameter.

Huber (1964) introduced the class of M-estimates. Suppose that $x_1, \ldots, x_n$ are i.i.d. observations with density function $f(x, \theta)$, $\theta \in \Theta$, with Θ some parameter space. The M-estimate of θ is

$$\hat{\mu}_n = \hat{\mu}_n(x_1, \ldots, x_n) = \arg\min_{\theta \in \Theta} \sum_{i=1}^{n} \rho(x_i, \theta), \tag{2.2}$$

where ρ is a loss function. When $\rho(x, \theta) = -\log f(x, \theta)$, $\hat{\mu}_n$ is the maximum likelihood estimate of θ. Under regularity conditions on ρ and Θ the estimate $\hat{\mu}_n$ also satisfies

$$\sum_{i=1}^{n} \psi(x_i, \hat{\mu}_n) = 0, \tag{2.3}$$

where $\psi = \partial\rho/\partial\theta$.

If the data follow model (2.1) and $\theta = \mu$ in (2.2) it is natural to choose the loss function ρ to be a function of the residuals, $\rho(x, \theta) = \rho(x - \theta)$. Then (2.3) becomes

$$\sum_{i=1}^{n} \psi(x_i - \hat{\mu}_n) = 0. \tag{2.4}$$

In the following we will assume that the function $\psi : \mathbb{R} \to \mathbb{R}$ satisfies:

P.1 $\psi(-u) = -\psi(u)$, $u \geq 0$, and bounded;

P.2 ψ is non-decreasing and $\lim_{u\to\infty} \psi(u) > 0$;

P.3 ψ is continuous.

Without loss of generality, if ψ satisfies P.1 we can assume that $|\psi(u)| \leq 1$, $u \in \mathbb{R}$.

Definition 2.1 - **M-location estimates** *(with known scale): Let $x_1, \ldots, x_n$ be a random sample following model (2.1). Let $\psi : \mathbb{R} \to \mathbb{R}$ satisfy P.1 to P.3 above. The solution $\hat{\mu}_n$ of (2.4) is called an M-location estimate.*

A widely used family of functions ψ_c was proposed by Huber (1964). Its members are given by

$$\psi_c(u) = \begin{cases} \operatorname{sgn}(u) & \text{if } |u| \geq c \\ u/c & \text{if } |u| < c, \end{cases} \tag{2.5}$$

where $c \in \mathbb{R}_+$ is a user-chosen constant and $\operatorname{sgn}(u)$ is the sign function. The constant c determines the asymptotic properties of the sequence $\hat{\mu}_n$ (see Definition 2.13 on page 39). One corresponding function ρ_c is given by

$$\rho_c(u) = \begin{cases} u - c/2 & \text{if } u > c \\ u^2/2c & \text{if } |u| \leq c \\ -u - c/2 & \text{if } u < -c\,. \end{cases}$$

Under certain regularity conditions the corresponding M-location estimates are asymptotically normal (Huber, 1967). The choice $c = 1.345$ yields an asymptotic efficiency of 95% when $\epsilon_i \sim N(0,1)$. For some asymptotic results we will need the function ψ_c to be twice continuously differentiable. We can easily construct functions that satisfy the regularity conditions of Definition 2.1 and that are twice continuously differentiable. For example, for a given $c > 0$ we can find constants a, b, d and e such that

$$f_c(u) = \begin{cases} \operatorname{sgn}(u) & \text{if } |u| \geq c \\ a\,u^7 + b\,u^5 + d\,u^3 + e\,u & \text{if } |u| \leq c \end{cases} \tag{2.6}$$

is twice continuously differentiable with $f_c(\pm c) = \pm 1$, $f_c'(\pm c) = 0$, $f_c'(0) = 1$ and $f_c''(\pm c) = 0$.

Beaton and Tukey (1974) proposed another family of functions ψ_d,

$$\psi_d(u) = \begin{cases} 0 & \text{if } |u| \geq d \\ (u/d)\left(1 - (u/d)^2\right)^2 & \text{if } |u| < d. \end{cases} \tag{2.7}$$

The constant d determines the asymptotic properties of these estimates. The associated family of functions ρ_d is given by

$$\rho_d(u) = \begin{cases} 3(u/d)^2 - 3(u/d)^4 + (u/d)^6 & \text{if } |u| \leq d \\ 1 & \text{if } |u| > d, \end{cases} \tag{2.8}$$

This family of functions ψ_d differs from Huber's in that its members vanish for large values of x. In terms of the estimate this feature means that outlying points will be ignored instead of down-weighted. In the robustness literature these functions are called re-descending.

A property that is natural to expect from an estimate for the location parameter in (2.1) is that it be equivariant under shifts in the center of the data.

Definition 2.2 - **Translation equivariance**: *We will say that an estimate $\hat{\mu}_n = \hat{\mu}_n(x_1, \ldots x_n)$ is translation equivariant if for any sample $x_1, \ldots, x_n$ and real number a we have*

$$\hat{\mu}_n(x_1 + a, \ldots, x_n + a) = \hat{\mu}_n(x_1, \ldots, x_n) + a.$$

It is easy to verify that estimates $\hat{\mu}_n$ that satisfy (2.4) are translation equivariant. Equivariance with respect to change of scale is also of interest.

Definition 2.3 - **Scale equivariance**: *We will say that an estimate $\hat{\mu}_n = \hat{\mu}_n(x_1, \ldots x_n)$ is scale equivariant if for any sample $x_1, \ldots, x_n$ and real number a we have*

$$\hat{\mu}_n(a\,x_1, \ldots, a\,x_n) = a\,\hat{\mu}_n(x_1, \ldots, x_n)\,. \tag{2.9}$$

The estimates $\hat{\mu}_n$ defined by equation (2.4) are not generally scale equivariant. To obtain equivariant estimates we introduce scale estimates.

Definition 2.4 - **Scale estimate**: *Let $x_1, \ldots, x_n$ be a sample of n real numbers. An estimate $\hat{\sigma}_n = \hat{\sigma}_n(x_1, \ldots, x_n)$ such that*

$$\hat{\sigma}_n(a\,x_1 + b, \ldots, a\,x_n + b) = |a|\;\hat{\sigma}_n(x_1, \ldots, x_n) \qquad \forall\; a, b \in \mathbb{R}, \tag{2.10}$$

will be called a scale estimate.

Equation (2.4) can incorporate the scale estimate $\hat{\sigma}_n$ as follows.

Definition 2.5 - **M-location estimates with general scale**: *Let $\psi : \mathbb{R} \to \mathbb{R}$ satisfy P.1 to P.3. Let $x_1, \ldots, x_n$ be a random sample of real numbers and let $\hat{\sigma}_n$ be* *a scale estimate. The M-location estimate with general scale is the solution $\hat{\mu}_n$ of*

$$\frac{1}{n}\sum_{i=1}^{n}\psi\left((x_i-\hat{\mu}_n)/\,\hat{\sigma}_n\right)=0\,. \tag{2.11}$$

Let ρ_ψ be a real function such that $\rho'_\psi=\psi$, then $\hat{\mu}_n$ can also be defined by

$$\hat{\mu}_n=\arg\min_{t\in\mathbb{R}}\frac{1}{n}\sum_{i=1}^{n}\rho_\psi\left((x_i-t)/\,\hat{\sigma}_n\right)\,. \tag{2.12}$$

If ψ is not continuous in the above definition, then the solution of (2.11) may not exist. We can still define the M-location estimate in this situation as

$$\hat{\mu}_n=\inf\left\{\quad t\in\mathbb{R}:\sum_{i=1}^{n}\psi\left((x_i-t)/\,\hat{\sigma}_n\right)\leq 0\quad\right\},$$

where $\inf\mathcal{A}$ denotes the infimum of the set $\mathcal{A}$ (see Huber, 1981, page 46).

It is easy to verify that the M-location estimates with general scale as defined in Definition 2.5 are translation and scale equivariant.

Different scale estimates $\hat{\sigma}_n$ generate different classes of M-location estimates. Definitions 2.6 and 2.7 consider two particular classes: the M-scale and S-scale estimates respectively.

In the following we will assume that the real function $\rho:\mathbb{R}\to\mathbb{R}_+$, satisfies $\rho\,(0)=0$ and

R.1 $\rho\,(-u)=\rho\,(u)$, $u\geq 0$, and $\sup_{u\in\mathbb{R}}\rho\,(u)=1$;

R.2 $\rho\,(u)$ is non-decreasing in $u\geq 0$;

R.3 ρ is continuous.

Note that without loss of generality, any symmetric and bounded function ρ that is not constantly equal to 0 can be adjusted to satisfy R.1 above.

For an arbitrary measurable function f and a random variable X with distribution function F, let $E_F f(X)$ denote the expected value of the random variable $f(X)$ when X has distribution F, if this expectation exists.

Definition 2.6 - M-scale estimates *(Huber, 1964): Let $\rho : \mathbb{R} \to \mathbb{R}$ satisfy R.1 to R.3 above. Let $b \in (0, 1/2]$. Let $x_1, \ldots, x_n$ be a random sample and let $\hat{\mu}_n$ be a scale- and translation-equivariant estimate. Define the residuals $r_1 = x_1 - \hat{\mu}_n, \ldots, r_n = x_n - \hat{\mu}_n$. The M-scale $\hat{\sigma}_n$ is implicitly defined by*

$$\frac{1}{n}\sum_{i=1}^{n} \rho\left(r_i / \hat{\sigma}_n\right) = b. \tag{2.13}$$

The choices of the function ρ and the constant b in (2.13) determine the properties of the resulting scale estimate. For example, to ensure consistency of $\hat{\sigma}_n$ when the residuals r_i's in (2.13) are standard normal random variables we choose $b = E_\Phi \rho(Z)$, where $Z \sim N(0, 1)$. The constant b will also characterize the robustness properties of the sequence $\hat{\sigma}_n$ (see Section 2.2).

A widely used family of ρ functions is given by

$$\rho_k(u) = \begin{cases} (u/k)^2 & \text{if } |u| \leq k \\ 1 & \text{if } |u| > k, \end{cases} \tag{2.14}$$

where k is a user-chosen constant. For a given $b \in (0, 1/2]$ we can choose k to obtain $E_\Phi \rho_k(Z) = b$. $k = 1.04086$ satisfies $E_\Phi \rho_k(Z) = 1/2$.

Another family of scale estimates is that of the S-scales (Rousseeuw and Yohai, 1984).

Definition 2.7 - **S-scale estimates**: *Let $\rho : \mathbb{R} \to \mathbb{R}_+$ and $b \in \mathbb{R}$ as in Definition 2.6. Let $x_1, \ldots, x_n$ be a random sample. For every $t \in \mathbb{R}$ consider the residuals $x_1 - t, \ldots, x_n - t$ and their M-scale $s_n(t)$ satisfying*

$$\frac{1}{n} \sum_{i=1}^{n} \rho\left((x_i - t)/\, s_n(t)\right) = b. \tag{2.15}$$

The S-scale $\hat{\sigma}_n$ is defined by

$$\hat{\sigma}_n(x_1, \ldots, x_n) = \inf_{t \in \mathbb{R}} s_n(t). \tag{2.16}$$

Naturally associated with this family are the S-location estimates.

Definition 2.8 - **S-location estimates**: *Let $x_1, \ldots, x_n$ be a random sample, and for each $t \in \mathbb{R}$ let $s_n(t)$ be as in (2.15). The S-location estimate $\tilde{\mu}_n$ is*

$$\tilde{\mu}_n(x_1, \ldots, x_n) = \arg\inf_{t \in \mathbb{R}} s_n(t). \tag{2.17}$$

It is easy to see that if the function ρ is continuously differentiable, the pair $(\tilde{\mu}_n, \hat{\sigma}_n)$ in (2.16) and (2.17) satisfies the following system of equations

$$\frac{1}{n} \sum_{i=1}^{n} \rho\left((x_i - \tilde{\mu}_n)/\, \hat{\sigma}_n\right) = b \tag{2.18}$$

$$\frac{1}{n} \sum_{i=1}^{n} \rho'\left((x_i - \tilde{\mu}_n)/\, \hat{\sigma}_n\right) = 0, \tag{2.19}$$

where ρ' denotes the derivative of ρ.

In analogy with Yohai (1987) we will refer to the M-location estimates calculated with an S-scale as MM-location estimates.

Definition 2.9 - **MM-location estimates**: *Let $x_1, \ldots, x_n$ be a random sample following model (2.1). Let $\psi : \mathbb{R} \to \mathbb{R}$ satisfy P.1 to P.3. Let $\hat{\sigma}_n$ be an S-scale estimate as in (2.16). The solution $\hat{\mu}_n$ of*

$$\sum_{i=1}^{n} \psi\left((x_i - \hat{\mu}_n) / \hat{\sigma}_n \right) = 0\,, \tag{2.20}$$

will be called the MM-location estimate of $x_1, \ldots, x_n$.

Definition 2.10 - **Simultaneous M-location and scale estimates** *(Huber, 1964): Let $\psi : \mathbb{R} \to \mathbb{R}$ satisfy P.1 to P.3, and let $\rho : \mathbb{R} \to \mathbb{R}_+$ satisfy R.1 to R.3. Let $x_1, \ldots, x_n$ be a random sample and let $b = E_\Phi \rho(Z)$. The simultaneous M-location and scale estimates $\hat{\mu}_n$ and $\hat{\sigma}_n$ are given by the solution of the following system of equations*

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^{n} \psi\left((x_i - \hat{\mu}_n) / \hat{\sigma}_n \right) &= 0, \\ \frac{1}{n} \sum_{i=1}^{n} \rho\left((x_i - \hat{\mu}_n) / \hat{\sigma}_n \right) &= b, \end{aligned} \tag{2.21}$$

2.2 Robustness properties

The asymptotic properties of the M-location estimates given by (2.12) are well-known when the distribution of the errors is symmetric (Huber, 1964, 1967, 1981; Boos and Serfling, 1980; Clarke, 1983, 1984).

We will assume that the distribution of the errors belongs to the following gross error neighbourhood

$$\mathcal{H}_\epsilon = \left\{ F \in \mathcal{D} \,:\, F(x) = (1 - \epsilon)\; F_0\left((x - \mu) / \sigma \right) + \epsilon\, H(x) \right\}, \tag{2.22}$$

where $\mathcal{D}$ denotes the set of all distribution functions, F_0 is a fixed symmetric distribution, $\epsilon \in (0, 1/2)$, and H is an arbitrary distribution function. Intuitively ϵ is the fraction of outliers that are expected to be present in the sample. We should mention here that (2.22) does not constitute a topological neighbourhood. See Huber (1981, page 10) for a discussion of different neighbourhoods.

We need to introduce some notation. Let $\mathcal{E}$ be a subset of the set of all distribution functions. We will assume that all possible empirical distribution functions belong to $\mathcal{E}$. Let ρ and b be as in the definition of M-scale estimates, Definition 2.6. Define the functional $\boldsymbol{\sigma} : \mathcal{E} \to \mathbb{R}$ as follows. Let $F \in \mathcal{E}$. For every $t \in \mathbb{R}$, let $\sigma\,(F, t)$ satisfy

$$E_F \left[\, \rho\left(\, (X - t) \,/\, \sigma\,(F, t)\,\right)\,\right] = b. \tag{2.23}$$

For $F \in \mathcal{E}$ the value of the functional $\boldsymbol{\sigma}\,(F)$ is

$$\boldsymbol{\sigma}\,(F) = \inf_{t \in \mathbb{R}} \sigma\,(F, t)\,. \tag{2.24}$$

Define the S-location functional as

$$\tilde{\boldsymbol{\mu}}\,(F) = \arg \inf_{t \in \mathbb{R}} \sigma\,(F, t)\,. \tag{2.25}$$

Clearly, we have $\boldsymbol{\sigma}\,(F) = \sigma\,(F, \tilde{\boldsymbol{\mu}}\,(F))$. Similarly define the functional $\boldsymbol{\mu} : \mathcal{E} \to \mathbb{R}$ by the equation

$$E_F \left[\, \psi\left(\, (X - \boldsymbol{\mu}\,(F)) \,/\, \boldsymbol{\sigma}\,(F)\,\right)\,\right] = 0, \tag{2.26}$$

or, equivalently

$$\boldsymbol{\mu}\,(F) = \arg \min_{t \in \mathbb{R}} E_F \left[\, \rho_\psi \left(\, (X - \boldsymbol{\mu}\,(F)) \,/\, \boldsymbol{\sigma}\,(F)\,\right)\,\right],$$

where $\rho'_\psi = \psi$. It is easy to see that if F_n denotes the empirical distribution function of the sample, then $\boldsymbol{\mu}(F_n) = \hat{\mu}_n$ and $\boldsymbol{\sigma}(F_n) = \hat{\sigma}_n$ where $\hat{\mu}_n$ and $\hat{\sigma}_n$ are given by (2.12) and (2.16) respectively.

In what follows we will assume that $\mathcal{H}_\epsilon \subseteq \mathcal{E}$.

One measure of robustness of an estimate is given by its asymptotic bias. If F is not symmetric we will typically have $\boldsymbol{\mu}(F) \neq \mu$, where $\boldsymbol{\mu}(F)$ is defined in (2.26) and μ is the actual location parameter in (2.1). The supremum of the absolute value of this difference as F ranges over the neighbourhood $\mathcal{H}_\epsilon$ (see 2.22) measures the worst asymptotic deviation we can have. This quantity is called the maximum asymptotic bias.

Definition 2.11 - Maximum asymptotic bias: *Let $\boldsymbol{\mu}$ be a statistic defined by a functional as in (2.26). The maximum asymptotic bias of $\boldsymbol{\mu}$ over $\mathcal{H}_\epsilon$ is given by*

$$\mathbf{B}(\epsilon) = \sup_{F \in \mathcal{H}_\epsilon} |\boldsymbol{\mu}(F) - \boldsymbol{\mu}(F_0)| / \boldsymbol{\sigma}(F_0),$$

where F_0 is the central distribution of the neighborhood $\mathcal{H}_\epsilon$.

Another measure of robustness for estimates is the breakdown point. This concept was defined by Hampel (1971). Intuitively the breakdown point is the smallest proportion of contamination ϵ^* such that the maximum asymptotic bias is unbounded.

Definition 2.12 - Asymptotic breakdown point: *Let $\boldsymbol{\mu}$ be a statistic as before and let $\mathbf{B}(\epsilon)$ be its maximum asymptotic bias function given in Definition (2.11). The* *asymptotic breakdown point of* $\boldsymbol{\mu}$ *is defined by*

$$\epsilon^* = \epsilon^*(F_0) = \sup\{\epsilon \in (0,1) \,:\, \mathbf{B}(\epsilon) < \infty\}.$$

In many cases ϵ^* does not depend on F_0 (Huber, 1981, page 13). Donoho and Huber (1983) introduced the following modified version for finite samples. Let $\boldsymbol{\mu}(F_n)$ be an estimate of the parameters of interest, where F_n denotes the empirical distribution of the sample. Let $b(m, \boldsymbol{\mu}, F_n)$ be the maximum disturbance you can cause to the estimate for this data set if you arbitrarily change m observations. Formally, let

$$b(m, \boldsymbol{\mu}, F_n) = \sup_{F_n'} |\boldsymbol{\mu}(F_n) - \boldsymbol{\mu}(F_n')|,$$

where F_n' is an empirical distribution that differs from F_n in that m data points have been replaced by arbitrary values. The finite sample breakdown point (BP) is the minimum m that yields an unbounded $b(m, \boldsymbol{\mu}, F_n)$.

Definition 2.13 - Finite sample breakdown point*: Let* F_n *be the empirical distribution function of a sample of size n, and let* $\boldsymbol{\mu}$ *be a statistic defined by a functional as above. The breakdown point of* μ *at the sample* F_n *is*

$$\epsilon_n^*(\boldsymbol{\mu}, F_n) = \min\left\{\frac{m}{n}, \quad m \in \mathbb{N}, \quad \textit{such that} \quad b(m, \boldsymbol{\mu}, F_n) < \infty\right\}$$

There is an asymptotically equivalent version of this definition where $b(m, \boldsymbol{\mu}, F_n)$ is calculated by adding m points to the sample instead of replacing them (see Donoho and Huber, 1983).

When the scale σ is known, M-location estimates can be defined as in (2.4).

In this case we have (Huber, 1981, page 53) that

$$\epsilon^* = \frac{\eta}{1+\eta}$$

with $\eta = \min\{-\psi(-\infty)/\psi(\infty), -\psi(\infty)/\psi(-\infty)\}$ and $\psi(\infty) = \lim_{x\to\infty}\psi(x)$. Hence, $\epsilon^* = 0.50$ when $\psi(\infty) = -\psi(-\infty)$. Note that if ψ satisfies P.1 and P.2 then $\epsilon^* = 0.50$. When σ is unknown and is estimated simultaneously as in (2.21) the breakdown point ϵ^* of $\hat{\mu}_n$ decreases. To avoid this effect, we can use an estimator $\hat{\sigma}_n$ with $\epsilon^* = 0.50$ and that does not use $\hat{\mu}_n$ as its centering statistic (such as the median of the absolute deviations from the median, MAD). In this way, if ψ satisfies P.1 we can achieve $\epsilon^* = 1/2$ for $\hat{\mu}_n$ when σ is unknown (see Huber, 1981, page 144).

On the other hand, M-scale estimates (2.13) have $\epsilon^* = \min(b, 1-b)$ (Huber, 1981). To obtain a consistent scale estimate $\hat{\sigma}_n$ with $\epsilon^* = 0.50$ when we use a function ρ_k in (2.14) we set $k = 1.04086$. When ρ_d belongs to Tukey's family (2.8) we need $d = 1.54764$. More generally, to construct a consistent M-scale estimate with breakdown point $\delta \in (0, 1/2)$, the tuning constant $d = d(\delta)$ can be determined by solving $E_\Phi \rho_d(Z) = \delta$ for d. S-estimates of scale and location as defined in (2.16) and (2.17) respectively, also have $\epsilon^* = 0.50$ when $b = 0.50$ (Rousseeuw and Yohai, 1984).

2.3 Asymptotic properties

The objective of this section is to determine the conditions under which the MM-location estimates (2.20) are consistent and asymptotically normal when the distribution of the data $F \in \mathcal{H}_\epsilon$ is not necessarily symmetric. We are also interested in obtaining uniform properties over the contamination neighbourhood $\mathcal{H}_\epsilon$.

We will see that if F is asymmetric the asymptotic distribution of MM-location estimates depends heavily on that of the S-scale (2.16) and associated S-location (2.17) estimates. Hence we begin by studying their asymptotic behaviour.

The rest of this section is organized as follows. Section 2.3.1 considers the consistency of the S-scale estimate. Sections 2.3.2 and 2.3.3 deal with the S-location estimates. Finally, Sections 2.3.4 to 2.3.7 study the consistency and asymptotic distribution of MM-location estimates.

2.3.1 Uniform consistency of the S-scale estimate

The following theorem shows that the S-scale estimate is consistent for the asymptotic value defined in (2.24), and that this convergence holds uniformly for $F \in \mathcal{H}_\epsilon$. Let $g_\rho(s,t) = E_{F_0}\rho((X-t)/s)$ and $h_\rho(s,t) = (\partial/\partial s)\, g_\rho(s,t)$.

Theorem 2.1 - *(Martin and Zamar, 1993)* - **Uniform consistency of the S-scale***:* *Let ρ and b satisfy the conditions in the definition of M-scale estimates, Definition 2.6. Let $\hat{\sigma}_n$ be the S-scale as in Definition 2.7 and $\boldsymbol{\sigma}(F)$ its asymptotic value as in (2.24). Let h_ρ be as above. Assume that h_ρ is continuous and that $h_\rho(s,t) < 0$ for all $s \in \mathbb{R}_+$ and $t \in \mathbb{R}$. Then, for any $\delta > 0$*

$$\lim_{m\to\infty} \sup_{F\in\mathcal{H}_\epsilon} P_F\left[\sup_{n\geq m} |\hat{\sigma}_n - \boldsymbol{\sigma}(F)| > \delta\right] = 0\,.$$

The following lemma shows that if we use a function ρ_d in Tukey's family to obtain an S-scale estimate, and the contamination neighbourhood $\mathcal{H}_\epsilon$ is centered around a distribution function with strictly positive density on $\mathbb{R}$ (such as the standard normal) then the conditions of Theorem 2.1 are met.

Lemma 2.1 *Let ρ_d belong to Tukey's class (2.8) and let $\mathcal{H}_\epsilon$ be a contamination neighbourhood as in (2.22) around a distribution F_0 with density function f_0 that is strictly positive on $\mathbb{R}$. Let h_{ρ_d} be defined as above. Then, ρ_d and h_{ρ_d} are continuous, and $h_{\rho_d}(s,t) < 0$, $\forall s \in \mathbb{R}_+$, $\forall t \in \mathbb{R}$.*

Proof: The continuity of ρ_d follows from its definition. Note that for ρ_d in Tukey's family we have

$$\rho_d'(u)\ u = \begin{cases} 0 & \text{if } |u| > d \\ \left(u^2/d\right)\left(1 - \left(u^2/d^2\right)\right)^2 & \text{if } |u| \leq d. \end{cases}$$

The continuity of h_{ρ_d} is a consequence of the Dominated Convergence Theorem. Note that $\rho_d'(u)\ u \geq 0$ for all $u \in \mathbb{R}$. Finally, that $h_{\rho_d}(s,t) < 0$, $\forall s \in \mathbb{R}_+$, $\forall t \in \mathbb{R}$ follows by noting that for any pair $(s,t) \in \mathbb{R}_+ \times \mathbb{R}$ there exists a set $\mathcal{K}_{s,t} \in \mathbb{R}$ such that

$$\rho_d'\left(\frac{X-t}{s}\right)\left(\frac{X-t}{s}\right) > 0 \qquad X \in \mathcal{K}_{s,t}\,.$$

By hypothesis we have $F_0\left(\mathcal{K}_{s,t}\right) > 0$. Then

$$E_{F_0}\rho_d'\left(\frac{X-t}{s}\right)\left(\frac{X-t}{s}\right) \geq \int_{\mathcal{K}_{s,t}} \rho_d'\left(\frac{X-t}{s}\right)\left(\frac{X-t}{s}\right) f_0(X) > 0\,.$$

■

2.3.2 Consistency of the S-location estimate

Our next theorem shows that under certain regularity conditions the S-location estimate is consistent for distributions $F \in \mathcal{H}_\epsilon$. Conditions under which this convergence holds uniformly on $\mathcal{H}_\epsilon$ are discussed in the next section.

We need to introduce the following notation. Let $\rho(x,t,s) = \rho((x-t)/s)$. Denote the set of positive real numbers $(0,\infty)$ by $\mathbb{R}_+$. For each $t \in \mathbb{R}$ and $s \in \mathbb{R}_+$ let

$$\gamma(F,t,s) = E_F \rho(X,t,s), \tag{2.27}$$

$$\gamma_n(t,s) = \gamma(F_n,t,s) = \frac{1}{n}\sum_{i=1}^{n} \rho(x_i,t,s), \tag{2.28}$$

where F_n denotes the empirical distribution function of the random sample $x_1, \dots, x_n$. As in Martin and Zamar (1993), for each $\epsilon \in [0,1/2)$ let $s^- = s^-(\epsilon)$ and $s^+ = s^+(\epsilon)$ be such that

$$0 < s^- \leq \inf_{F\in\mathcal{H}_\epsilon} \boldsymbol{\sigma}(F) < \sup_{F\in\mathcal{H}_\epsilon} \boldsymbol{\sigma}(F) \leq s^+ < \infty, \tag{2.29}$$

where $\boldsymbol{\sigma}(F)$ is given by (2.24).

Theorem 2.2 - Consistency of the S-location estimate*: Let ρ satisfy R.1 to R.3, and assume that ρ is not constant. Let $\tilde{\mu}_n$ be the associated S-location estimate as in Definition 2.8. Let $F \in \mathcal{H}_\epsilon$ and let $\tilde{\boldsymbol{\mu}}(F)$ be as in (2.25). Assume the following:*

1. *$\gamma(F,t,s) = E_F\rho((X-t)/s)$ is continuous in s uniformly in t;*

2. *$g(t) = \gamma(F,t,\boldsymbol{\sigma}(F))$ has a unique minimum $\tilde{\boldsymbol{\mu}}$;*

3. for any neighbourhood $B(t_0)$ we have

$$E_F\left[\inf_{t'\in B(t_0)} \rho(X,t',\boldsymbol{\sigma})\right] \xrightarrow[B(t_0)\searrow\{t_0\}]{} E_F\Big[\rho(X,t_0,\boldsymbol{\sigma})\Big],$$

as $B(t_0)$ shrinks to $\{t_0\}$;

4. for any bounded neighbourhood $B(t_0)$ we have

$$\frac{1}{n}\sum_{i=1}^{n} \inf_{t'\in B(t_0)} \rho(x_i,t',\hat{\sigma}_n) \xrightarrow[n\to\infty]{a.s.} E_F\left[\inf_{t'\in B(t_0)} \rho(X,t',\boldsymbol{\sigma})\right].$$

Then

$$\tilde{\mu}_n \xrightarrow[n\to\infty]{P} \tilde{\boldsymbol{\mu}}.$$

Proof: First note that by the previous result we know that $\hat{\sigma}_n \to \boldsymbol{\sigma}$ almost surely. Hence, with high probability, $\hat{\sigma}_n \in \mathcal{K}$, a fixed compact set, for n sufficiently large.

Note that $g(t) = \gamma(F,t,\boldsymbol{\sigma})$ is a continuous, bounded function, and by hypothesis it has a unique minimum. Denote the value of t where this minimum is attained by t_0, i.e. $g(t_0) < g(t)$ for all $t \neq t_0$. We will show that there exist sets $I_1 \subset I_2$ with $t_0 \in I_1$ such that

$$\sup_{t\in I_1} \gamma(F,t,\boldsymbol{\sigma}) < \inf_{t\notin I_2} \gamma(F,t,\boldsymbol{\sigma}). \tag{2.30}$$

Because ρ is not constant, we have $\lim_{|t|\to\infty} g(t) = 1 > g(t_0)$. Let $\epsilon_1 = 1 - g(t_0) > 0$. Choose a_1 such that $|t| > a_1$ implies $1 - g(t) < \epsilon_1/2$. Then we have $\inf_{|t|>a_1} g(t) \geq 1 - \epsilon_1/2 = g(t_0) + \epsilon_1/2$. Hence, $A = \{|x| \leq a_1\}$ satisfies $\inf_{t\notin A} g(t) - g(t_0) > 0$. Also note that necessarily $|t_0| \leq a_1$. By continuity of g there exists a neighbourhood of t_0,

$B(t_0)$ such that $g(t) - g(t_0) < \epsilon_1/4$ for all $t \in B(t_0)$. It follows that $\sup_{B(t_0)} g(t) \leq g(t_0) + \epsilon_1/4$. We will now show that $B(t_0) \subset \{|t| \leq a_1\}$. Let $t \in B(t_0)$. Then, $g(t) < g(t_0) + \epsilon_1/4$. If $|t| > a_1$ then, $g(t) \geq g(t_0) + \epsilon_1/2$, which is a contradiction. Hence, the above inclusion holds. Next note that $\inf_{|t|>a_1} g(t) \geq g(t_0) + \epsilon_1/2 > g(t_0) + \epsilon_1/4 \geq \sup_{B(t_0)} g(t)$. Hence $I_1 = B(t_0) \subset \{|t| \leq a_1\} = I_2$ satisfy (2.30).

We now show that, with high probability, $\tilde{\mu}_n$ eventually lies in the compact set I_2. It is enough to prove that with high probability, there exists n_1 such that for $n \geq n_1$ we have

$$\sup_{t \in I_1} \gamma_n(t, \hat{\sigma}_n) < \inf_{t \notin I_2} \gamma_n(t, \hat{\sigma}_n). \tag{2.31}$$

Let

$$\alpha = \sup_{t \in I_1} \gamma(F, t, \boldsymbol{\sigma}), \quad \text{and} \quad \eta = \inf_{t \notin I_2} \gamma(F, t, \boldsymbol{\sigma}).$$

Let $0 < \epsilon' < (\eta - \alpha)/2$. Note that for any $\delta > 0$, $t \in \mathbb{R}$ and $s \in \mathbb{R}_+$, Chebychev's inequality yields

$$P_F\left(|\gamma_n(t, s) - \gamma(F, t, s)| > \delta\right) \leq \frac{1}{n}\frac{1}{\delta^2}, \qquad \forall F.$$

It follows that there exists an integer $n_0(\epsilon')$ such that, with high probability,

$$|\gamma_n(t, s) - \gamma(F, t, s)| < \epsilon'/2, \qquad \forall\, n \geq n_0(\epsilon'),\ \forall\, t,\ \forall\, s,\ \forall F. \tag{2.32}$$

Note that because of the uniformity in (2.32) we have for each fixed t and $n \geq n_0$

$$|\gamma_n(t, \hat{\sigma}_n) - \gamma(F, t, \hat{\sigma}_n)| < \epsilon'/2.$$

Let $n_1(\epsilon')$ be such that with high probability $\hat{\sigma}_n$ and $\boldsymbol{\sigma}$ are close enough so that by the continuity of γ we have

$$|\gamma(F, t, \hat{\sigma}_n) - \gamma(F, t, \boldsymbol{\sigma})| < \epsilon'/2 \qquad \text{for } n \geq n_1.$$

Note that n_1 does not depend on t. It follows that $\gamma_n(t, \hat{\sigma}_n) < \gamma(F, t, \boldsymbol{\sigma}) + \epsilon'$ and $\sup_{t \in I_1} \gamma_n(t, \hat{\sigma}_n) \leq \alpha + \epsilon'$ for $n \geq n_1$. Similarly we have $\eta - \epsilon' \leq \inf_{t \notin I_2} \gamma_n(t, \hat{\sigma}_n)$ for $n \geq n_1$ and (2.31) holds. It follows that, with high probability, there exists an integer $\tilde{n}$ such that $n \geq \tilde{n} \Rightarrow \tilde{\mu}_n \in I_2$.

Having proved that $\tilde{\mu}_n$ is ultimately in a compact set I_2, and that the unique minimum of the asymptotic equation belongs to I_2, we now adapt a standard argument (Huber, 1967) to show that $\tilde{\mu}_n$ converges almost surely to $\tilde{\boldsymbol{\mu}}$.

Let $B(\tilde{\boldsymbol{\mu}})$ be an arbitrary neighbourhood of $\tilde{\boldsymbol{\mu}}$. We will restrict our attention to the compact set I_2. We have

$$\inf_{t \notin B(\tilde{\boldsymbol{\mu}})} \gamma(F, t, \boldsymbol{\sigma}) \geq \gamma(F, \tilde{\boldsymbol{\mu}}, \boldsymbol{\sigma}) + 4\,\epsilon,$$

for some $\epsilon > 0$. By hypothesis, for each $t \notin B(\tilde{\boldsymbol{\mu}})$ there exists a neighbourhood of t such that

$$E\left[\inf_{t' \in B(t)} \rho(X, t', \boldsymbol{\sigma})\right] \geq \gamma(F, \tilde{\boldsymbol{\mu}}, \sigma) + 3\,\epsilon. \tag{2.33}$$

The collection of open sets $B(t)$ covers the compact set $I_2 \cap B(\tilde{\boldsymbol{\mu}})^c$. Pick a finite number of them such that

$$\bigcup_{i=1}^{k} B(t_i) \supset I_2 \cap B(\tilde{\boldsymbol{\mu}})^c. \tag{2.34}$$

For each of them we have that if n is large enough

$$\begin{aligned}\inf_{t \in B(t_i)} \frac{1}{n} \sum_{i=1}^{n} \rho(x_i, t, \hat{\sigma}_n) &\geq \frac{1}{n} \sum_{i=1}^{n} \inf_{t \in B(t_i)} \rho(x_i, t, \hat{\sigma}_n) \\ &\geq \gamma(F, \tilde{\boldsymbol{\mu}}, \boldsymbol{\sigma}) + 2\,\epsilon. \end{aligned} \tag{2.35}$$

Also

$$\frac{1}{n}\sum_{i=1}^{n}\rho\left(x_i,\tilde{\boldsymbol{\mu}},\hat{\sigma}_n\right)\leq\gamma\left(F,\tilde{\boldsymbol{\mu}},\boldsymbol{\sigma}\right)+\epsilon. \tag{2.36}$$

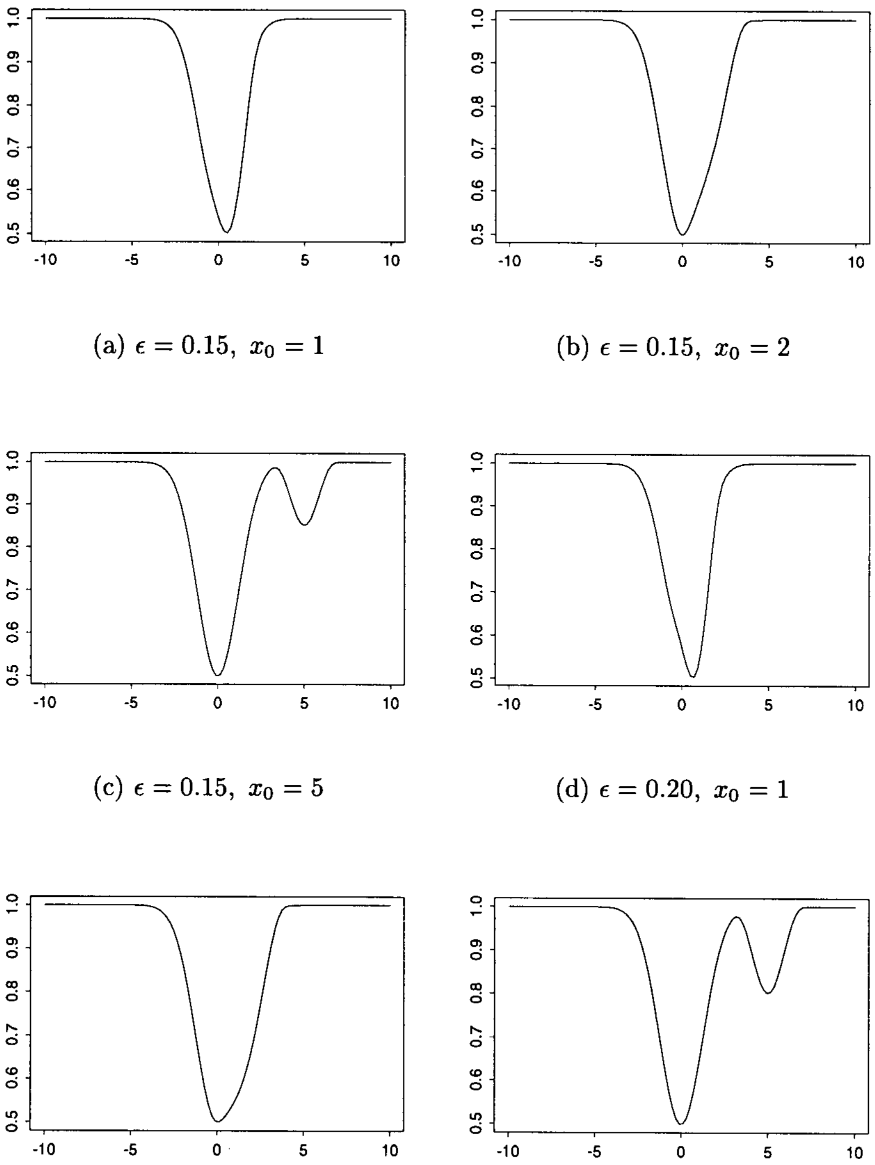
It follows that for n large enough

$$\begin{aligned}\inf_{t\notin B(\tilde{\mu})}\frac{1}{n}\sum_{i=1}^{n}\rho\left(x_i,t,\hat{\sigma}_n\right) &\geq \frac{1}{n}\sum_{i=1}^{n}\rho\left(x_i,\tilde{\boldsymbol{\mu}},\hat{\sigma}_n\right)+\epsilon \\ &\geq \inf_{t\in B(\tilde{\mu})}\frac{1}{n}\sum_{i=1}^{n}\rho\left(x_i,t,\hat{\sigma}_n\right)+\epsilon,\end{aligned}$$

and hence $\tilde{\mu}_n\in B\left(\tilde{\boldsymbol{\mu}}\right)$. ■

Application of Theorem 2.2 to Tukey's family of functions ρ_d

Lemmas 7.7 to 7.11 in the Appendix show that if ρ_d belongs to Tukey's family (2.8) then assumptions 1, 3 and 4 of Theorem 2.2 are met. Assumption 2 is particularly difficult to verify in general. Under certain regularity conditions and for $\epsilon\leq 0.10$ we can show that it holds uniformly on $F\in\mathcal{H}_\epsilon$ (see next Section). We conjecture that it holds for any $F\in\mathcal{H}_\epsilon$. The following plots illustrate the behaviour of the family of functions $f_\epsilon\left(t\right)=\gamma\left(F_\epsilon,t,\boldsymbol{\sigma}\left(F_\epsilon\right)\right)$ for F_ϵ in the contamination neighbourhood of the standard normal distribution and ρ_d in Tukey's family (2.8). We considered $F_\epsilon\left(x\right)=\left(1-\epsilon\right)\Phi\left(x\right)+\epsilon\Phi\left(\left(x-x_0\right)/0.1\right)$ for $\epsilon=0.15, 0.20, 0.25, 0.30, 0.40, 0.45$ and $x_0=1, 2$, and 5. We see that in all cases $\gamma\left(F_\epsilon,t,\boldsymbol{\sigma}\left(F_\epsilon\right)\right)$ has a unique global minimum. Note that as x_0 increases a second local minimum appears. Only for $\epsilon\geq 0.40$ this local minimum approaches the global minimum of the function, but these functions seem to always have a unique global minimum for $\epsilon<0.50$ for the contaminations considered here.



(a) $\epsilon = 0.15,\ x_0 = 1$

(b) $\epsilon = 0.15,\ x_0 = 2$

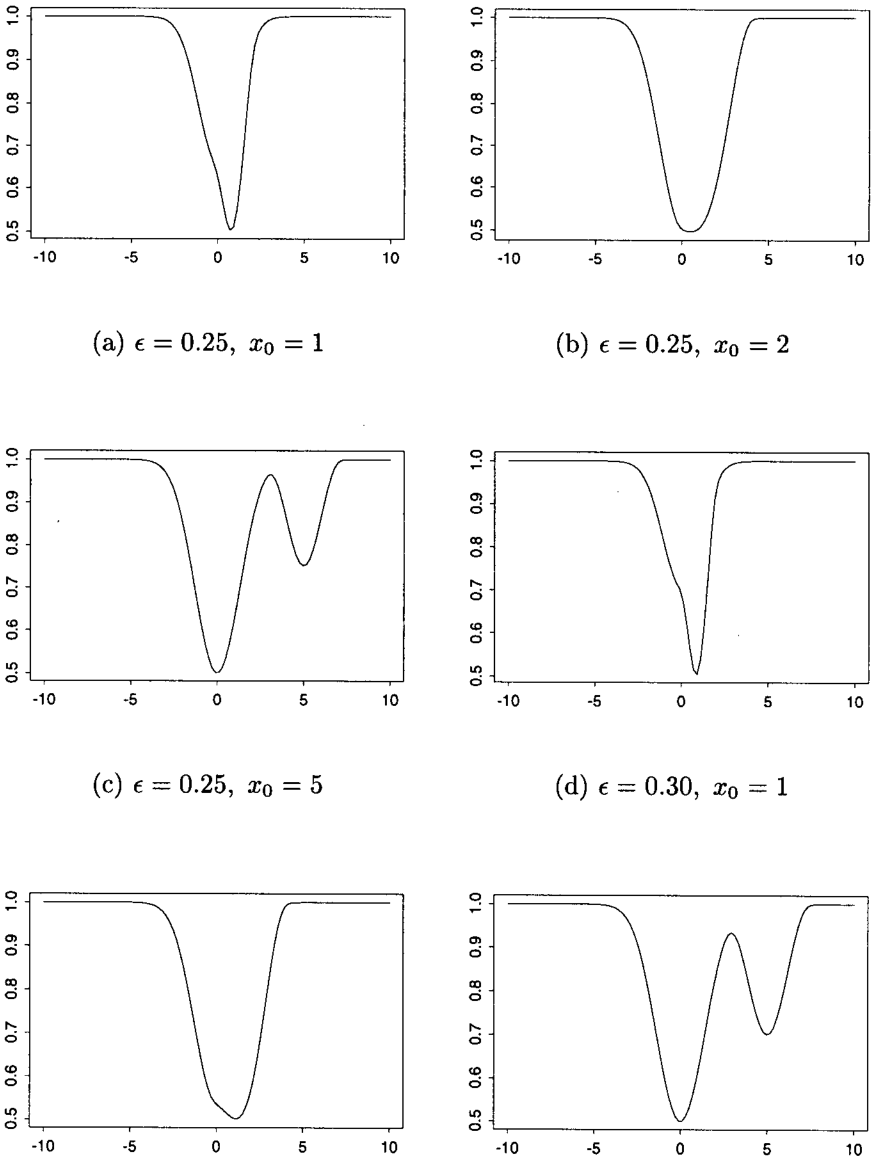
(c) $\epsilon = 0.15,\ x_0 = 5$

(d) $\epsilon = 0.20,\ x_0 = 1$

(e) $\epsilon = 0.20,\ x_0 = 2$

(f) $\epsilon = 0.20,\ x_0 = 5$

Figure 2.1: Plots of $f(t) = \gamma(F_\epsilon, t, \boldsymbol{\sigma}(F_\epsilon))$ with $F_\epsilon(x) = (1-\epsilon)\,\Phi(x) + \epsilon\,\Phi((x-x_0)/\,0.1)$, for different values of ϵ and x_0.



(a) $\epsilon = 0.25,\ x_0 = 1$

(b) $\epsilon = 0.25,\ x_0 = 2$

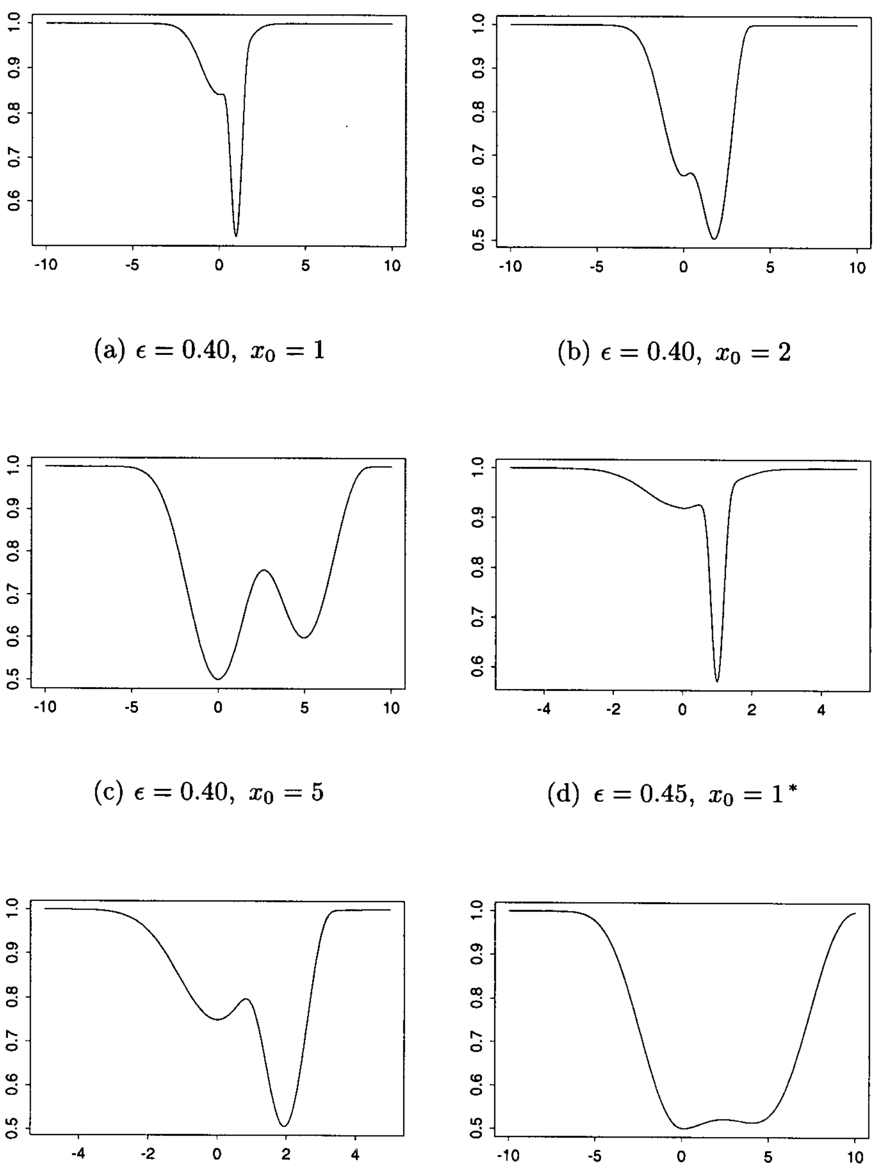
(c) $\epsilon = 0.25,\ x_0 = 5$

(d) $\epsilon = 0.30,\ x_0 = 1$

(e) $\epsilon = 0.30,\ x_0 = 2$

(f) $\epsilon = 0.30,\ x_0 = 5$

Figure 2.2: Plots of $f\,(t) = \gamma\,(\,F_\epsilon,\, t,\, \boldsymbol{\sigma}\,(F_\epsilon)\,)$ with $F_\epsilon\,(x) = (1-\epsilon)\;\Phi\,(x) + \epsilon\,\Phi\,(\,(x - x_0)/\,0.1)$, for different values of ϵ and x_0.



(a) $\epsilon = 0.40, \ x_0 = 1$

(b) $\epsilon = 0.40, \ x_0 = 2$

(c) $\epsilon = 0.40, \ x_0 = 5$

(d) $\epsilon = 0.45, \ x_0 = 1^*$

(e) $\epsilon = 0.45, \ x_0 = 2^*$

(f) $\epsilon = 0.45, \ x_0 = 5$

Figure 2.3: Plots of $f(t) = \gamma(F_\epsilon, t, \boldsymbol{\sigma}(F_\epsilon))$ with $F_\epsilon(x) = (1-\epsilon)\,\Phi(x) + \epsilon\,\Phi((x-x_0)/\,0.1)$, for different values of ϵ and x_0. *Note the different scale on the x-axis for $\epsilon = 0.45, \ x_0 = 1$ and 2.

2.3.3 Uniform consistency of the S-location estimate

In this section we show that under stronger conditions on ρ than those of Theorem 2.2 we obtain uniform consistency of $\tilde{\mu}_n$. These new regularity assumptions are basically the uniform counterparts of those of Theorem 2.2.

We first state the theorem and its proof. We then obtain sufficient conditions to meet the required uniform assumptions and we verify them for Tukey's family of functions ρ_d in a range of values of the proportion of contamination ϵ.

Theorem 2.3 - **Uniform consistency of the S-location estimate**: *Let ρ satisfy R.1 to R.3, and assume that ρ is not constant. Let $b \in (0, 1/2]$, $\tilde{\mu}_n$ as in Definition 2.8 and $\tilde{\boldsymbol{\mu}}(F)$ as in (2.25). If*

U.1 $\gamma(F, t, s)$ is continuous in s uniformly in $t \in \mathbb{R}$ and $F \in \mathcal{H}_\epsilon$, that is, for any $\tilde{\epsilon} > 0$ there exists a $\delta = \delta(\tilde{\epsilon}) > 0$ such that if $|s_1 - s_2| < \delta$ then

$$\Big|\gamma(F, t, s_1) - \gamma(F, t, s_2)\Big| < \tilde{\epsilon}, \qquad \forall\, t \in \mathbb{R},\ \forall\, F \in \mathcal{H}_\epsilon\,; \tag{2.37}$$

U.2 for each $F \in \mathcal{H}_\epsilon$, $f_F(t) = \gamma\big(F, t, \boldsymbol{\sigma}(F)\big)$ has a unique minimum $\tilde{\boldsymbol{\mu}}(F)$;

U.3 there exists sets $I_1 \subset I_2$ with I_2 compact that do not depend on $F \in \mathcal{H}_\epsilon$, and such that

$$\sup_{t\in I_1} \gamma\big(F, t, \boldsymbol{\sigma}(F)\big) < \inf_{t\notin I_2} \gamma\big(F, t, \boldsymbol{\sigma}(F)\big), \quad \forall\, F \in \mathcal{H}_\epsilon\,; \tag{2.38}$$

U.4 the convergence in assumption 3 of Theorem 2.2 holds uniformly in $F \in \mathcal{H}_\epsilon$, *i.e. for every $\tilde{\epsilon} > 0$ and $t_0 \in \mathbb{R}$ there exists $\delta = \delta\left(\tilde{\epsilon}, t_0\right)$ such that*

$$\left| E_F\Big[\inf_{t' \in B_\delta(t_0)} \rho\left(X, t', \boldsymbol{\sigma}(F)\right)\Big] - E_F\Big[\rho\left(X, t_0, \boldsymbol{\sigma}(F)\right)\Big] \right| < \tilde{\epsilon}, \quad \forall\, F \in \mathcal{H}_\epsilon\,, \tag{2.39}$$

where the ball $B_\delta\left(t_0\right)$ has diameter δ;

U.5 the convergence in 4 of Theorem 2.2 holds uniformly in $F \in \mathcal{H}_\epsilon$, that is, if

$$Y_i = \inf_{t' \in B_\delta(t_0)} \rho\left(X_i, t', \hat{\sigma}_n\right) \quad \text{and} \quad Y\left(F\right) = E_F\left[\inf_{t' \in B_\delta(t_0)} \rho\left(X, t_0, \boldsymbol{\sigma}(F)\right)\right],$$

then for any $\delta > 0$

$$\lim_{m\to\infty} \sup_{F \in \mathcal{H}_\epsilon} P_F\Big(\sup_{n \geq m} \left|\overline{Y}_n - Y\left(F\right)\right| > \delta \Big) = 0\,; \tag{2.40}$$

U.6 for every $\delta > 0$, $\tilde{\epsilon}\left(\delta, F\right)$ defined by the property

$$\inf_{t \notin B_\delta(\tilde{\boldsymbol{\mu}}(F))} \gamma\big(F, t, \boldsymbol{\sigma}(F)\big) \;\geq\; \gamma\big(F, \tilde{\boldsymbol{\mu}}(F), \boldsymbol{\sigma}(F)\big) \;+\; \tilde{\epsilon}\left(\delta, F\right)\,, \tag{2.41}$$

where $\tilde{\boldsymbol{\mu}}\left(F\right)$ is the global minimum of $\gamma\big(F, t, \boldsymbol{\sigma}(F)\big)$, satisfies

$$\tilde{\epsilon}\left(\delta\right) = \inf_{F \in \mathcal{H}_\epsilon} \tilde{\epsilon}\left(\delta, F\right) > 0\,; \tag{2.42}$$

then

$$\lim_{m\to\infty} \sup_{F \in \mathcal{H}_\epsilon} P_F\Big(\sup_{n \geq m} \left|\tilde{\mu}_n - \boldsymbol{\mu}\left(F\right)\right| > \delta \Big) = 0\,. \tag{2.43}$$

Proof: Fix an arbitrary neighbourhood $B\big(\tilde{\boldsymbol{\mu}}(F)\big)$ of $\tilde{\boldsymbol{\mu}}\left(F\right)$ and let $\tilde{\epsilon} > 0$ be given by (2.42). Let I_1 and I_2 be as in U.3. Note that by assumption U.4 the finite coverage of I_2 in (2.33) and (2.34) associated with $\tilde{\epsilon}\left(\delta\right)$ in U.6 does not depend on F. Let Y_i and Y be as in U.5. Consider the events

$$A_m\left(F\right) = \left\{\sup_{n \geq m} \left|\overline{Y}_n - Y\left(F\right)\right| \leq \tilde{\epsilon}\right\}, \qquad m \in \mathbb{N}\,.$$

Choose an arbitrary $\delta > 0$. By U.5 we have that there exists $m_0(\delta)$ such that

$$P_F\Big(A_m(F)\Big) > 1-\delta, \qquad \forall m \geq m_0(\delta) \quad \forall F \in \mathcal{H}_\epsilon .$$

Now note that

$$A_m(F) \subseteq \left\{ \frac{1}{n} \sum_{i=1}^{n} \inf_{t \in B(t_j)} \rho(x_i, t, \hat{\sigma}_n) \geq \gamma\big(F, \tilde{\boldsymbol{\mu}}(F), \boldsymbol{\sigma}(F)\big) + 2\tilde{\epsilon}, \forall n \geq m \right\} = C_m(F) .$$

Let

$$D_m(F) = \left\{ \frac{1}{n} \sum_{i=1}^{n} \rho\big(x_i, \tilde{\boldsymbol{\mu}}(F), \boldsymbol{\sigma}(F)\big) \ \leq\ \gamma\big(F, \tilde{\boldsymbol{\mu}}(F), \boldsymbol{\sigma}(F)\big) + \tilde{\epsilon},\ \forall n \geq m \right\} .$$

We also have that there exists $m_1 = m_1(\delta)$ such that for $m \geq m_1$ we have

$$P_F\big(D_m(F)\big) > 1-\delta \qquad \forall F \in \mathcal{H}_\epsilon .$$

Take $m_2 = \max(m_0, m_1)$. We have

$$P_F\Big[C_m(F) \cap D_m(F)\Big] \geq 1-2\delta \qquad \forall m \geq m_2, \quad \forall F \in \mathcal{H}_\epsilon .$$

We also have

$$C_m(F) \cap D_m(F) \quad \subseteq \quad \Big[\tilde{\mu}_m \in B\big(\tilde{\boldsymbol{\mu}}(F)\big), m \geq m_2\Big] .$$

Hence, for each $\delta > 0$ there exists $m_2(\delta)$ such that

$$P_F\Big[\tilde{\mu}_m \in B\big(\tilde{\boldsymbol{\mu}}(F)\big),\ \forall m \geq m_2\Big] \geq 1-2\delta, \qquad \forall F \in \mathcal{H}_\epsilon ,$$

that is, for each neighbourhood $B\big(\tilde{\boldsymbol{\mu}}(F)\big)$ we have

$$\lim_{m\to\infty} \inf_{F \in \mathcal{H}_\epsilon} P_F\Big[\tilde{\mu}_n \in B\big(\tilde{\boldsymbol{\mu}}(F)\big),\ \forall n \geq m\Big] = 1 ,$$

or equivalently, if $d_B > 0$ is the diameter of $B\big(\tilde{\boldsymbol{\mu}}(F)\big)$,

$$\lim_{m\to\infty} \sup_{F \in \mathcal{H}_\epsilon} P_F\Big[\sup_{n \geq m} |\tilde{\mu}_n - \tilde{\boldsymbol{\mu}}(F)| > d_B\Big] = 0 .$$

■

Application of Theorem 2.3 for Tukey's family of functions ρ_d

Here we discuss sufficient conditions for U.1 to U.6 to hold for S-location estimates obtained with functions ρ_d in Tukey's family (2.8).

Assumption U.1 holds by Lemmas 7.8 and 7.9. To see that assumption U.4 holds use Lemmas 7.7 and 7.10.

We now show that U.5 holds. Let Y_i and $Y(F)$ be as in U.5 and

$$V_i(F) = \inf_{t' \in B(t_0)} \rho\left(X_i, t', \boldsymbol{\sigma}(F)\right), \qquad i = 1, \ldots, n.$$

Then $Y(F) = E_F[V_i]$. We have to show that for any $\delta > 0$ and $\tilde{\epsilon} > 0$ there exists m_0 such that

$$P_F\left[\sup_{n \geq m} \left|\overline{Y}_n - Y(F)\right| > \delta\right] < \tilde{\epsilon} \qquad \forall m \geq m_0.$$

We cannot use Lemma 7.2 (Bernstein's inequality) on $Y_i - Y(F)$ because these random variables do not have mean zero nor are they independent. We have

$$P_F\left[\sup_{n \geq m} \left|\overline{Y}_n - Y(F)\right| > \delta\right] \leq P_F\left[\sup_{n \geq m} \left|\overline{V}_n(F) - Y(F)\right| > \delta/2\right] + \\ + P_F\left[\sup_{n \geq m} \left|\overline{Y}_n - \overline{V}_n(F)\right| > \delta/2\right] \quad (2.44)$$

for some $\epsilon'(\delta) > 0$ that depends on δ. We have

$$P_F\left[\sup_{n \geq m} \left|\overline{Y}_n - \overline{V}_n(F)\right| > \delta/2\right] \leq P_F\left[\sup_{n \geq m} \left|\hat{\sigma}_n - \boldsymbol{\sigma}(F)\right| > \epsilon'\right], \qquad (2.45)$$

for some $\epsilon' = \epsilon'(\delta)$. To prove inequality (2.45) note that $\overline{Y}_n = 1/n \sum_{i=1}^n g(x_i, \hat{\sigma}_n)$ and $\overline{V}_n = 1/n \sum_{i=1}^n g(x_i, \boldsymbol{\sigma})$. In the proof of Lemma 7.11 we see that $g(x, s)$ is continuous in s uniformly on x. Hence, for a given $\delta/2$ there exists a positive ϵ' such that $|\hat{\sigma}_n - \boldsymbol{\sigma}| < \epsilon'$ implies $\left|\overline{Y}_n - \overline{V}_n\right| < \delta/2$. Hence, for each n we have

$$\left\{\left|\overline{Y}_n - \overline{V}_n\right| > \delta/2\right\} \subset \left\{|\hat{\sigma}_n - \boldsymbol{\sigma}| > \epsilon'\right\},$$

and then note that for any sequence of random variables $\{X_n\}_{n\in\mathbb{N}}$ if a is a real number, we have $\left\{\sup_{n\geq m} X_n > a\right\} = \bigcup_{n\geq m}\{X_n > a\}$. Together with (2.45) this bounds the second term in (2.44). To control the first term, note that the sequence of random variables $W_i = V_i - E(V_i) = V_i - Y(F)$ satisfies the assumptions of Bernstein's Lemma (Lemma 7.2 in the Appendix) with $c = 2\sup_u \rho(u)$ and $s_n = n\,\sigma_w^2$, where σ_w^2 denotes the variance of W_i. Hence for any $\delta > 0$ we have

$$\begin{aligned} P_F\Big(\left|\overline{V}_n - E(V)\right| > \delta\Big) &= P_F\Big(\left|\overline{W}_n\right| > \delta\Big) \\ &\leq 2\exp\left(\frac{-n\,\delta^2}{2\,(\sigma_w^2 + c\,\delta)}\right) \leq 2\ \exp\left(\frac{-n\,\delta^2}{2\,(k^2 + c\,\delta)}\right) \\ &= 2\ \Big[\exp\Big(-a(\delta)\Big)\Big]^n, \end{aligned} \tag{2.46}$$

where $\sigma_w^2 \leq k^2 < \infty$ for all $F \in \mathcal{H}_\epsilon$ and $a(\delta) > 0$. Note that they do not depend on F. Use Theorem 2.1 to find m_0 large enough such that

$$\sup_{F\in\mathcal{H}_\epsilon} P_F\left[\sup_{n\geq m}\left|\hat{\sigma}_n - \boldsymbol{\sigma}(F)\right| > \epsilon'\right] < \tilde{\epsilon}/2, \tag{2.47}$$

and use (2.46) together with the Borel-Cantelli Lemma (Lemma 7.3) and a standard argument to find m_1 large enough such that

$$\sup_{F\in\mathcal{H}_\epsilon} P_F\left[\sup_{n\geq m}\left|\overline{V}_n - Y(F)\right| > \delta/2\right] < \tilde{\epsilon}/2. \tag{2.48}$$

Equations (2.44), (2.47) and (2.48) show U.5.

Assumptions U.2, U.3 and U.6 are closely related and we consider them together. Let s^+ and s^- be as in (2.29) and suppose that there exists $t^* \in \mathbb{R}$ such that

$$\inf_{s^- \leq s \leq s^+} \left[E_\Phi \rho \left(\frac{X-t}{s} \right) - E_\Phi \rho \left(\frac{X}{s} \right) \right] > \frac{\epsilon}{1-\epsilon}, \qquad \forall \, |t| \geq t^*. \tag{2.49}$$

The above condition, together with the hypothesis of existence of a unique minimum for each $F \in \mathcal{H}_\epsilon$ suffices to show U.3 (i.e., that the set I_2 in (2.30) does not depend on $F \in \mathcal{H}_\epsilon$). Equation (2.49) is hard to verify analytically. Numerical evaluation shows that (2.49) holds for $\epsilon \leq 0.25$ for estimates calculated with ρ_d in Tukey's family (2.8) and $d = 1.54764$ (i.e., estimates with breakdown point 50%). Assumption U.2 seems hard to prove. Together with (2.49) a condition to ensure that for all $F \in \mathcal{H}_\epsilon$ the function $\gamma \left(F, \cdot, \boldsymbol{\sigma} \right)$ has a unique minimum is

$$\inf_{\substack{-t^* \leq t \leq t^* \\ s^- \leq s \leq s^+}} E_\Phi \rho'' \left(\frac{X-t}{s} \right) > \frac{\epsilon}{1-\epsilon} \sup_x \rho'' \left(x \right)^- \ , \tag{2.50}$$

where t^* is given in (2.49). The above condition suffices to show that the functions $\gamma \left(F, \cdot, \boldsymbol{\sigma} \left(F \right) \right)$, with $F \in \mathcal{H}_\epsilon$, are uniformly convex on $(-t^*, t^*)$. We have the following implication:

$$(2.49) \text{ and } (2.50) \quad \Rightarrow \quad \text{U.2, U.3 and U.6} \, .$$

It is easy to see that (2.49) implies that $\gamma \left(F, 0, \boldsymbol{\sigma} \left(F \right) \right) < \gamma \left(F, t, \boldsymbol{\sigma} \left(F \right) \right)$ for all $|t| \geq t^*$ and thus U.3 holds. Equation (2.50) implies

$$\inf_{\substack{-t^* \leq t \leq t^* \\ s^- \leq s \leq s^+}} \gamma'' \left(F, t, s \right) \geq \eta > 0, \qquad \forall \, F \in \mathcal{H}_\epsilon,$$

where η does not depend on F. Equation (2.50) ensures that the functions are strictly convex on that interval, and hence have a unique global minimum in this fixed interval (assumption U.2). To verify U.6 note that for any $t \notin B_\delta\big(\tilde{\boldsymbol{\mu}}(F)\big)$ we have

$$\begin{aligned}\gamma\big(F, t, \boldsymbol{\sigma}(F)\big) - \gamma\left(F, \tilde{\boldsymbol{\mu}}(F), \boldsymbol{\sigma}(F)\right) &= \frac{1}{2}\gamma''\left(F, \bar{t}, \boldsymbol{\sigma}(F)\right)\left(t - \tilde{\boldsymbol{\mu}}(F)\right)^2 \\ &\geq \eta\ \left(t - \tilde{\boldsymbol{\mu}}(F)\right)^2 \\ &> \eta\ \delta^2 \qquad\qquad \forall\, F \in \mathcal{H}_\epsilon,\end{aligned}$$

where $\bar{t} \notin B_\delta\big(\tilde{\boldsymbol{\mu}}(F)\big)$ and η does not depend on F.

Condition (2.50) is quite strong. The S-estimates obtained with ρ functions in Tukey's family having breakdown point 50% do not satisfy (2.50) for $\epsilon = 0.10$. That is, if there is 10% contamination we cannot guarantee uniform convergence over the neighbourhood. The main problem seems to be that the threshold t^* found in (2.49) is unnecessarily large. We can adjust the choice of this constant as follows. For each $t \in \mathbb{R}$ define the set

$$\mathcal{A}(t) = \left\{ s_H(t)\ :\ E_H\rho\left(\frac{X - t}{s_H(t)}\right) = b,\ H \in \mathcal{H}_\epsilon \right\},$$

let $s^-(t) = \inf \mathcal{A}(t)$ and $s^+(t) = \sup \mathcal{A}(t)$. If we choose t^* as the solution of

$$\inf_{s^-(t^*) \leq s \leq s^+(t^*)} \left[-E_\Phi \rho'\left(\frac{X - t^*}{s}\right)\right] = \frac{\epsilon}{1-\epsilon} \sup_x \rho'(x)\ , \tag{2.51}$$

then (2.50) holds for a larger range of values of ϵ.

We will now show that (2.51) and (2.50) are sufficient conditions to ensure that $\gamma(F, \cdot, \boldsymbol{\sigma})$ has its unique global minimum in the interval $(-t^*, t^*)$ for any $F \in \mathcal{H}_\epsilon$, where t^* is given by (2.51). The reasoning is as follows. If t is a minimum of $\gamma(F, \cdot, \boldsymbol{\sigma})$ then it solves the equation

$$(1-\epsilon)\, E_\Phi \rho_d'\left(\frac{X-t}{s(t)}\right) + \epsilon E_H \rho_d'\left(\frac{X-t}{s(t)}\right) = 0, \tag{2.52}$$

where $s(t) = \boldsymbol{\sigma}$. Hence, t solves

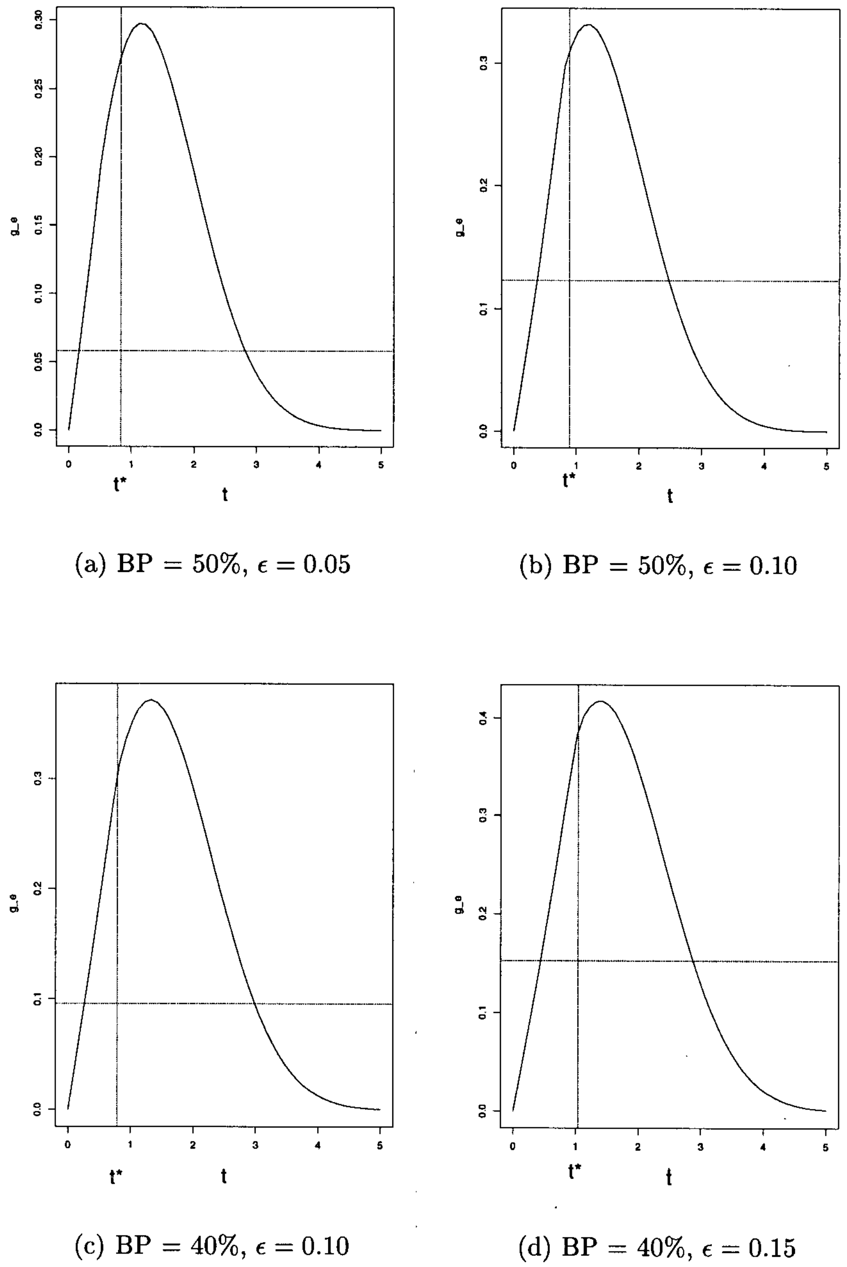
$$-E_{\Phi}\rho_d'\left(\frac{X-t}{s(t)}\right) = \frac{\epsilon}{(1-\epsilon)}E_H\rho_d'\left(\frac{X-t}{s(t)}\right) . \tag{2.53}$$

For each $\epsilon \in (0, 1/2]$ the largest solution t of (2.53) is determined by solving

$$g_\epsilon(t) = \inf_{s^-(t)\leq s\leq s^+(t)}\left[-E_\Phi\rho_d'\left(\frac{X-t}{s}\right)\right] =$$
$$= \sup_{H\in\mathcal{H}_\epsilon}\frac{\epsilon}{(1-\epsilon)}E_H\rho_d'\left(\frac{X-t}{s(t)}\right) = \frac{\epsilon}{1-\epsilon}\sup_x \rho_d'(x) \ , \tag{2.54}$$

that is, equation (2.51). In Figure 2.4 we plot the function $g_\epsilon(t)$ for estimates with breakdown point 50% and 40% and different values of ϵ. We include the threshold t^* obtained in (2.49). We see that the largest solution of (2.51) (or 2.54) is larger than the mentioned threshold, and hence this solution corresponds to a local minimum of $\gamma(F, \cdot, \boldsymbol{\sigma})$. The smallest solution t^* of (2.51) is then the largest possible value of t satisfying (2.52) that corresponds to a global minimum. Equation (2.50) guarantees that every function $\gamma(F, \cdot, \boldsymbol{\sigma})$ is strictly convex in $(-t^*, t^*)$. It follows that there only exists one global minimum, and that it belongs to this interval.

We evaluated conditions (2.50) and (2.51) for S-location estimates obtained with Tukey's ρ_d functions. We considered estimates with breakdown point 50% and 40%. Details are presented in Tables 2.1 and 2.2. For estimates with 50% breakdown point equation (2.50) holds for $\epsilon \leq 0.10$. When we lower the breakdown to 40%, (2.50) holds for $\epsilon \leq 0.15$. We see that there is a trade-off between high breakdown point and the uniform convexity condition in (2.50).



(a) BP = 50%, $\epsilon = 0.05$

(b) BP = 50%, $\epsilon = 0.10$

(c) BP = 40%, $\epsilon = 0.10$

(d) BP = 40%, $\epsilon = 0.15$

Figure 2.4: Plots of $g_\epsilon(t) = \inf_{s^-(t)\leq s\leq s^+(t)} \left[-E_\Phi \rho_d'\left(\frac{X-t}{s}\right)\right]$ for estimates with breakdown point 50 and 40%. The threshold t^* is given by (2.49). The horizontal line is at $\epsilon/(1-\epsilon)\sup_x \rho_d'(x)$.

BP	d	ϵ	s^-	s^+	sup ρ'_d	sup $[\rho''_d]^-$
0.50	1.54764	0.05	0.933919	1.069247	1.1096251	2.0040167
		0.10	0.863700	1.150487		
		0.11	0.849103	1.168490		
		0.12	0.834307	1.187162		
		0.13	0.819307	1.206545		
		0.15	0.788662	1.247639		
		0.20	0.707933	1.366741		
0.40	1.987967	0.05	0.949208	1.081607	0.86384744	1.214571
		0.10	0.895659	1.181595		
		0.15	0.838933	1.308399		
		0.16	0.827163	1.338136		
		0.17	0.815240	1.369688		
		0.20	0.778506	1.477396		

Table 2.1: Numerical parameters for Tukey's family of functions ρ_d. BP = Breakdown Point. $d =$ tunning constant. s^- and s^+ are defined in (2.29). $\sup \rho'_d = \sup_{x\in\mathbb{R}} \rho'_d\,(x)$. $\sup \rho''_d = \sup_{x\in\mathbb{R}} \rho''_d\,(x)$.

BP	ϵ	t^* as in	t^* as in	satisfies
		(2.49)	(2.51)	(2.50)
0.50	0.10	0.835	0.37205	Yes
	0.11	0.891	0.41940	No
	0.15	1.229	0.62620	No
0.40	0.10	0.786	0.26313	Yes
	0.15	1.038	0.43912	Yes
	0.16	1.092	0.47839	No
	0.20	1.203	0.65011	No

Table 2.2: Numerical evaluation of regularity conditions required for uniform consistency of S-location estimates with Tukey's family of functions ρ_d.

2.3.4 Consistency of the MM-location estimate

The next theorem shows that under certain regularity conditions, if $\hat{\sigma}_n$ is a consistent S-scale estimate, then the MM-location estimates that satisfy

$$\sum_{i=1}^{n} \psi\left(\left(x_i - \hat{\mu}_n\right) / \hat{\sigma}_n\right) = 0, \tag{2.55}$$

are also consistent. In Section 2.3.5 we study regularity conditions that suffice for these estimates to be uniformly consistent over the contamination neighbourhood $\mathcal{H}_\epsilon$.

Theorem 2.4 - Consistency of the MM-location estimate - *Let $x_1, \ldots, x_n$ be a random sample of i.i.d. random variables with distribution function $F \in \mathcal{H}_\epsilon$. Let $\hat{\sigma}_n = \hat{\sigma}_n\left(x_1, \ldots, x_n\right)$ be an S-scale estimate. Let $\psi : \mathbb{R} \to \mathbb{R}$ satisfy P.1 to P.3 and let $\hat{\mu}_n$ be a sequence of MM-estimates that solve (2.55) above. Then*

i) if $\hat{\sigma}_n \xrightarrow{P} \sigma\left(F\right)$ then $\hat{\mu}_n \xrightarrow[n\to\infty]{P} \boldsymbol{\mu}\left(F\right)$;

ii) if $\hat{\sigma}_n \xrightarrow{a.s.} \sigma\left(F\right)$ then $\hat{\mu}_n \xrightarrow[n\to\infty]{a.s.} \boldsymbol{\mu}\left(F\right)$.

Proof: Fraiman, Yohai and Zamar (2000) show that if ψ satisfies P.1 to P.3 and the central distribution of $\mathcal{H}_\epsilon$ has a density function $f_0\left(u\right)$ that satisfies $f_0'\left(u\right) < 0$ for all $u \geq 0$, then there exists a unique solution $\boldsymbol{\mu}\left(F\right)$ of

$$E_F\left[\psi\left(\frac{X - \boldsymbol{\mu}\left(F\right)}{\sigma}\right)\right] = 0\,.$$

They also show that under the same conditions $\gamma\left(F, t, \sigma\right) = E_F\left[\psi\left(\left(X - t\right)/\sigma\right)\right]$ is strictly monotone as a function of $t \in \mathbb{R}$. Also note that P.1 to P.3 imply that there exists $0 < L < \infty$ such that $\lim_{|x|\to\infty} \psi\left(x\right) = L$. The proof of Lemma 7.13 can be easily modified to show that these conditions imply that ψ is uniformly continuous on $\mathbb{R}$. We now adapt a classical argument (Huber, 1981, page 46) to obtain the desired results. Let $\hat{\mu}_n$ satisfy (2.55). To simplify the notation, in what follows let μ denote $\boldsymbol{\mu}(F)$.

To prove (i) we will show that for any $\epsilon > 0$

$$P\Big(\hat{\mu}_n \leq \mu - \epsilon\Big) \longrightarrow 0 \qquad \text{and} \qquad P\Big(\hat{\mu}_n \geq \mu + \epsilon\Big) \longrightarrow 0.$$

By Lemma 7.1 we know that for each $t \in \mathbb{R}$,

$$\frac{1}{n}\sum_{i=1}^{n} \psi\left(\frac{x_i - t}{\hat{\sigma}_n}\right) \xrightarrow[n\to\infty]{P} E_F\psi\left(\frac{X-t}{\sigma}\right).$$

In particular, if $t < \mu$ then the monotonicity of $E_F\left[\psi\left((X-t)/\sigma\right)\right]$ as a function of $t \in \mathbb{R}$ implies

$$\frac{1}{n}\sum_{i=1}^{n} \psi\left(\frac{x_i - t}{\hat{\sigma}_n}\right) \xrightarrow[n\to\infty]{P} E_F\psi\left(\frac{X-t}{\sigma}\right) > 0,$$

and hence

$$P\left(\frac{1}{n}\sum_{i=1}^{n} \psi\left(\frac{x_i - t}{\hat{\sigma}_n}\right) \leq 0\right) \longrightarrow 0\,, \qquad \forall \quad t < \mu\,. \tag{2.56}$$

Similarly we can show that

$$P\left(\frac{1}{n}\sum_{i=1}^{n} \psi\left(\frac{x_i - t}{\hat{\sigma}_n}\right) < 0\right) \longrightarrow 1\,, \qquad \forall \quad t > \mu\,. \tag{2.57}$$

Note the following inclusion

$$\Big\{t : \hat{\mu}_n < t\Big\} \subset \left\{t : \frac{1}{n}\sum_{i=1}^{n} \psi\left((x_i - t)/\,\hat{\sigma}_n\right) \leq 0\right\} \subset \Big\{t : \hat{\mu}_n \leq t\Big\}$$

We have

$$\Big\{\hat{\mu}_n \leq \mu - \epsilon\Big\} \subset \Big\{\hat{\mu}_n < \mu - \epsilon/2\Big\} \subset \left\{\frac{1}{n}\sum_{i=1}^{n} \psi\left((x_i - (\mu - \epsilon/2))/\,\hat{\sigma}_n\right) \leq 0\right\}.$$

Now (2.56) yields

$$P\Big(\hat{\mu}_n \leq \mu - \epsilon\Big) \leq P\left(\frac{1}{n}\sum_{i=1}^{n} \psi\left((x_i - (\mu - \epsilon/2))/\,\hat{\sigma}_n\right) \leq 0\right) \longrightarrow 0, \qquad (2.58)$$

so that $\overline{\lim}_{n\to\infty} P\left(\hat{\mu}_n \leq \mu - \epsilon\right) = 0$, and hence $\lim_{n\to\infty} P\left(\hat{\mu}_n \leq \mu - \epsilon\right) = 0$. For the second result, the inclusion

$$\Big\{\hat{\mu}_n \leq \mu + \epsilon\Big\} \supset \left\{\sum_{i=1}^{n} \psi\left((x_i - (\mu + \epsilon))/\,\hat{\sigma}_n\right) < 0\right\},$$

together with (2.57) yields

$$P\Big(\hat{\mu}_n \leq \mu + \epsilon\Big) \geq P\left(\frac{1}{n}\sum_{i=1}^{n} \psi\left((x_i - (\mu + \epsilon))/\,\hat{\sigma}_n\right) < 0\right) \longrightarrow 1\,.$$

Thus $\underline{\lim}_{n\to\infty} P\left(\hat{\mu}_n \leq \mu + \epsilon\right) = 1$ and $\lim_{n\to\infty} P\left(\hat{\mu}_n \geq \mu + \epsilon\right) = 0$. The result follows by noting that for any $\epsilon > 0$

$$P\Big(\mu - \epsilon \leq \hat{\mu}_n \leq \mu + \epsilon\Big) \xrightarrow[n\to\infty]{} 1.$$

This proves part (i) of the Theorem.

The proof of part (ii) follows the same lines. Now Lemma 7.1 yields

$$\frac{1}{n}\sum_{i=1}^{n} \psi\left(\frac{x_i - t}{\hat{\sigma}_n}\right) \xrightarrow[n\to\infty]{\text{a.s.}} E_F\psi\left(\frac{X - t}{\sigma}\right).$$

Hence, for each $\epsilon > 0$ there exists a null set $\mathcal{N}_\epsilon$ such that if $\omega \notin \mathcal{N}_\epsilon$ there exists $n_0 = n_0\,(\omega)$ with

$$\frac{1}{n}\sum_{i=1}^{n} \psi\left(\frac{x_i - (t - \epsilon)}{\hat{\sigma}_n}\right) < 0\,, \qquad (2.59)$$

for all $n \geq n_0$. Consider the null set $\mathcal{N} = \bigcup_{k\in\mathrm{N}} \mathcal{N}_{1/k}$. For any $\epsilon > 0$ and any $\omega \notin \mathcal{N}$ there exists $n_1 = n_1\,(\omega, \epsilon)$ such that (2.59) holds for $n \geq n_1$. This null set does not depend on ϵ. If $\mathcal{M}$ denotes the corresponding null set where

$$\frac{1}{n}\sum_{i=1}^{n}\psi\left(\frac{x_i-(t+\epsilon)}{\hat{\sigma}_n}\right)>0 \tag{2.60}$$

holds for large enough n, then for any $\omega \notin \mathcal{N}\bigcup\mathcal{M}$ the same reasoning as in the previous proof yields that there exists $n_2 = n_2(\omega,\epsilon)$ such that for any $n \geq n_2$, $|\hat{\mu}_n - \mu| < \epsilon$. ■

2.3.5 Uniform consistency of the MM-location estimate

In this section we show that if the scale estimates $\hat{\sigma}_n$ are uniformly consistent over the distributions in $\mathcal{H}_\epsilon$, and we impose more regularity conditions on the function ψ, then the MM-location estimates (2.20) are also uniformly consistent.

We need the following additional regularity condition:

P.4 ψ is continuously differentiable.

Theorem 2.5 - Uniform consistency of the M-location estimate with general scale: *Let $x_1,\ldots,x_n$ be i.i.d. observations following the location model (2.1). Let $\hat{\sigma}_n$ be a scale estimate that satisfies (2.29) and the conclusion of Theorem 2.1. Let ψ satisfy P.1 to P.4. Let $\hat{\mu}_n$ be the solution of (2.20) and let $\boldsymbol{\mu}(F)$ be the asymptotic value of $\hat{\mu}_n$ as defined in (2.26). Then for any $\delta > 0$*

$$\lim_{m\to\infty}\sup_{F\in\mathcal{H}_\epsilon}P_F\left[\sup_{n\geq m}\left|\hat{\mu}_n-\boldsymbol{\mu}(F)\right|>\delta\right]=0\,.$$

Proof: For any $t \in \mathbb{R}$ and $F \in \mathcal{H}_\epsilon$ let

$$\mu_\psi(t, F) = E_F \psi\left(\frac{X - t}{\boldsymbol{\sigma}(F)}\right) .$$

Let $\epsilon > 0$ be arbitrary. We first show that

$$a(\epsilon) = \inf_{F \in \mathcal{H}_\epsilon} \mu_\psi\big(\boldsymbol{\mu}(F) - \epsilon/2, F\big) > 0 , \tag{2.61}$$

and

$$b(\epsilon) = \inf_{F \in \mathcal{H}_\epsilon} -\mu_\psi\big(\boldsymbol{\mu}(F) + \epsilon/2, F\big) > 0 . \tag{2.62}$$

Equations (2.61) and (2.62) can be expressed as: the family of functions $\mu_\psi(t, F)$ has "uniform minimum slope" at $\boldsymbol{\mu}(F)$. Bounding $\partial\mu_\psi / \partial t\big|_{\boldsymbol{\mu}}$ uniformly over $F \in \mathcal{H}_\epsilon$ will be enough for these conditions to hold. Let $\lambda_F(\epsilon)$ be

$$\lambda_F(\epsilon) = E_F \psi\left(\frac{X - \boldsymbol{\mu}(F) + \epsilon}{\boldsymbol{\sigma}(F)}\right) ,$$

then $a(\epsilon) = \inf_{F \in \mathcal{H}_\epsilon} \lambda_F(\epsilon)$. Note that $\lambda_F(0) = 0$; hence

$$\lambda_F(\epsilon) = \epsilon\, \lambda'_F(\tilde{\epsilon}_F) ,$$

where $\tilde{\epsilon}_F \in (0, \epsilon)$. By (2.29) we have $0 < s^- \leq \boldsymbol{\sigma}(F) \leq s^+ < \infty$. Then

$$\begin{aligned}\lambda'_F(\tilde{\epsilon}_F) &= E_F \psi'\left(\frac{X - \boldsymbol{\mu}(F) + \tilde{\epsilon}_F}{\boldsymbol{\sigma}(F)}\right) \frac{1}{\boldsymbol{\sigma}(F)} \\ &\geq \frac{1}{s^+}\left(1 - \epsilon_{\mathcal{H}_\epsilon}\right) E_{F_0} \psi'\left(\frac{X - \boldsymbol{\mu}(F) + \tilde{\epsilon}_F}{\boldsymbol{\sigma}(F)}\right) ,\end{aligned}$$

where $\epsilon_{\mathcal{H}_\epsilon}$ is the proportion of contamination in $\mathcal{H}_\epsilon$. It is easy to see that the last term in the above equation is a decreasing function of $\tilde{\epsilon}_F$. Hence $\tilde{\epsilon}_F \leq \epsilon$ implies

$$\lambda_F(\epsilon) = \epsilon\, \lambda'_F(\tilde{\epsilon}_F) \geq \frac{\epsilon}{s^+}\left(1 - \epsilon_{\mathcal{H}_\epsilon}\right) E_{F_0} \psi'\left(\frac{X - \boldsymbol{\mu}(F) + \epsilon}{\boldsymbol{\sigma}(F)}\right) .$$

The Dominated Convergence Theorem shows that the above expression is continuous as a function of $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$. It is also positive and hence a sufficient condition to obtain a positive lower bound is that $\boldsymbol{\mu}(F)$ and $\boldsymbol{\sigma}(F)$ be bounded for any $F \in \mathcal{H}_\epsilon$. A similar argument can be applied to show that equation (2.62) holds.

Let $\boldsymbol{\sigma} = \boldsymbol{\sigma}(F)$ and $\boldsymbol{\mu} = \boldsymbol{\mu}(F)$. To simplify the notation let $\psi(X,t,s) = \psi((X-t)/s)$. For each t it is easy to see that $Y_i(t) = \psi(X_i, t, \hat{\sigma}_n)$ and $Y(F,t) = E_F \psi(X,t,\boldsymbol{\sigma})$ have the same properties as those in U.5, hence the proof on page 54 holds. Let $\overline{\psi}_n(t) = \frac{1}{n}\sum_{i=1}^n Y_i(t)$ and $\mu_\psi(t,F) = E_F(\psi(X,t,\boldsymbol{\sigma}))$. For each $\tau > 0$ and $t \in \mathbb{R}$ we have

$$\lim_{m\to\infty} \sup_{F\in\mathcal{H}_\epsilon} P_F\Big(\sup_{n\geq m} \left|\overline{\psi}_n(t) - \mu_\psi(t,F)\right| > \tau \Big) = 0\,. \tag{2.63}$$

For each $m \in \mathbb{N}$, $t \in \mathbb{R}$, $F \in \mathcal{H}_\epsilon$ and $\tau > 0$ let

$$\mathcal{A}_m(F,t,\tau) = \left\{ \sup_{n\geq m} \left| \overline{\psi}_n(t) - \mu_\psi(t,F) \right| > \tau \right\}\,;$$

then (2.63) can be written as

$$\lim_{m\to\infty} \sup_{F\in\mathcal{H}_\epsilon} P_F\Big(\mathcal{A}_m(F,t,\tau)\Big) = 0\,. \tag{2.64}$$

Now note that $\mu_\psi(\boldsymbol{\mu}(F), F) = 0$ and that $\mu_\psi(t,F)$ is a non-increasing function in t. We also have

$$\begin{aligned}\left\{ \hat{\mu}_n < \boldsymbol{\mu} - \epsilon \right\} &\subseteq \left\{ \frac{1}{n}\sum_{i=1}^n \psi(x_i, \boldsymbol{\mu} - \epsilon/2, \hat{\sigma}_n) \leq 0 \right\} \\ &\subseteq \left\{ \left| \overline{\psi}_n(\boldsymbol{\mu} - \epsilon/2) - \mu_\psi(\boldsymbol{\mu} - \epsilon/2, F) \right| > \mu_\psi(\boldsymbol{\mu} - \epsilon/2, F) \right\} \\ &\subseteq \left\{ \left| \overline{\psi}_n(\boldsymbol{\mu} - \epsilon/2) - \mu_\psi(\boldsymbol{\mu} - \epsilon/2, F) \right| > a(\epsilon) \right\} = A_n(F,\epsilon)\,,\end{aligned}$$

where $a(\epsilon)$ is given by (2.61). Similarly

$$\begin{aligned}\left\{\hat{\mu}_n > \boldsymbol{\mu} + \epsilon\right\} &\subseteq \left\{\frac{1}{n}\sum_{i=1}^{n}\psi\left(x_i, \boldsymbol{\mu} + \epsilon/2, \hat{\sigma}_n\right) \geq 0\right\} \\ &\subseteq \left\{\left|\overline{\psi}_n\left(\boldsymbol{\mu} + \epsilon/2\right) - \mu_\psi\left(\boldsymbol{\mu} + \epsilon/2, F\right)\right| > -\mu_\psi\left(\boldsymbol{\mu} + \epsilon/2, F\right)\right\} \\ &\subseteq \left\{\left|\overline{\psi}_n\left(\boldsymbol{\mu} - \epsilon/2\right) - \mu_\psi\left(\boldsymbol{\mu} - \epsilon/2, F\right)\right| > b(\epsilon)\right\} = B_n(F, \epsilon)\ .\end{aligned}$$

It follows that $\left\{\left|\hat{\mu}_n - \boldsymbol{\mu}\right| > \epsilon\right\} \subseteq A_n(F,\epsilon) \bigcup B_n(F,\epsilon)$. Hence,

$$\bigcup_{n=m}^{\infty}\left\{\left|\hat{\mu}_n - \boldsymbol{\mu}\right| > \epsilon\right\} \subseteq \bigcup_{n=m}^{\infty} A_n(F,\epsilon) \bigcup \bigcup_{n=m}^{\infty} B_n(F,\epsilon)\ .$$

Immediately

$$\begin{aligned}\mathcal{M}_m(F,\epsilon) &= \left\{\sup_{n\geq m}\left|\hat{\mu}_n - \boldsymbol{\mu}\right| > \epsilon\right\} \\ &\subseteq \left\{\sup_{n\geq m}\left|\overline{\psi}_n\left(\boldsymbol{\mu} - \epsilon/2\right) - \mu_\psi\left(\boldsymbol{\mu} - \epsilon/2, F\right)\right| > \mu_\psi\left(\boldsymbol{\mu} - \epsilon/2, F\right)\right\} \\ &\qquad \bigcup\left\{\sup_{n\geq m}\left|\overline{\psi}_n\left(\boldsymbol{\mu} + \epsilon/2\right) - \mu_\psi\left(\boldsymbol{\mu} + \epsilon/2, F\right)\right| > -\mu_\psi\left(\boldsymbol{\mu} + \epsilon/2, F\right)\right\} \\ &\subseteq \mathcal{A}_m\Big(F, \boldsymbol{\mu} - \epsilon/2, a(\epsilon)\Big) \bigcup \mathcal{A}_m\Big(F, \boldsymbol{\mu} + \epsilon/2, b(\epsilon)\Big)\ .\end{aligned}$$

We have

$$P_F\left[\mathcal{M}_m(F,\epsilon)\right] \leq P_F\left[\mathcal{A}_m\left(F, \boldsymbol{\mu} - \epsilon/2, a(\epsilon)\right)\right] + P_F\left[\mathcal{A}_m\left(F, \boldsymbol{\mu} + \epsilon/2, b(\epsilon)\right)\right]\ ,$$

and then

$$\begin{aligned}\sup_{F\in\mathcal{H}_\epsilon} P_F\left[\mathcal{M}_m(F,\epsilon)\right] \leq \sup_{F\in\mathcal{H}_\epsilon} P_F\left[\mathcal{A}_m\left(F, \boldsymbol{\mu} - \epsilon/2, a(\epsilon)\right)\right] \\ + \sup_{F\in\mathcal{H}_\epsilon} P_F\left[\mathcal{A}_m\left(F, \boldsymbol{\mu} + \epsilon/2, b(\epsilon)\right)\right]\ ,\end{aligned}$$

so that

$$\overline{\lim_{m\to\infty}} \sup_{F\in\mathcal{H}_\epsilon} P_F\left[\mathcal{M}_m\left(F,\epsilon\right)\right] \leq \lim_{m\to\infty} \sup_{F\in\mathcal{H}_\epsilon} P_F\left[\mathcal{A}_m\left(F,\boldsymbol{\mu}-\epsilon/2,a\left(\epsilon\right)\right)\right]$$
$$+\lim_{m\to\infty} \sup_{F\in\mathcal{H}_\epsilon} P_F\left[\mathcal{A}_m\left(F,\boldsymbol{\mu}+\epsilon/2,b\left(\epsilon\right)\right)\right] = 0\,,$$

and the proof is complete. ■

2.3.6 Asymptotic distribution of the MM-location estimate

Having shown the consistency of the estimates for any distribution F in the contamination neighbourhood, we turn our attention to their asymptotic distribution. The following argument will show the basic idea behind the proof of Theorem 2.6 and will also illustrate why we concentrate on location estimates calculated with an S-scale. Let $\hat{\mu}_n$ be an M-location estimate and $\hat{\sigma}_n$ a general scale estimate. To simplify the notation denote their asymptotic values $\boldsymbol{\mu}\left(F\right)$ and $\boldsymbol{\sigma}\left(F\right)$ by μ and σ respectively. We will consider M-location estimates with general scale, i.e. $\hat{\mu}_n$ satisfies

$$\frac{1}{n}\sum_{i=1}^{n}\psi\left(\left(x_i-\hat{\mu}_n\right)/\,\hat{\sigma}_n\right)=0\,. \tag{2.65}$$

Under certain regularity conditions a Taylor expansion of the above equation around (μ,σ) yields

$$0=\sum_{i=1}^{n}\psi\left(\left(x_i-\hat{\mu}_n\right)/\,\hat{\sigma}_n\right)=\sum_{i=1}^{n}\psi\left(\left(x_i-\mu\right)/\,\sigma\right)-\frac{\left(\hat{\mu}_n-\mu\right)}{\sigma}\sum_{i=1}^{n}\psi'\left(\left(x_i-\mu\right)/\,\sigma\right)$$
$$-\frac{\left(\hat{\sigma}_n-\sigma\right)}{\sigma}\sum_{i=1}^{n}\left[\psi'\left(\left(x_i-\mu\right)/\,\sigma\right)\left(x_i-\mu\right)/\,\sigma\right]+R_n, \tag{2.66}$$

where $R_n = o_P\left(1/\sqrt{n}\right)$ is the residual term. From here we obtain

$$\sqrt{n}\left(\hat{\mu}_n - \mu\right) = \sqrt{n}\, A_n\left(\mu, \sigma\right) + \sqrt{n}\left(\hat{\sigma}_n - \sigma\right)\, B_n\left(\mu, \sigma\right) + \tilde{R}_n, \tag{2.67}$$

where

$$A_n\left(\mu, \sigma\right) = \sigma \sum_{i=1}^{n} \psi\left(\left(x_i - \mu\right)/\sigma\right) \Big/ \sum_{i=1}^{n} \psi'\left(\left(x_i - \mu\right)/\sigma\right),$$

and

$$B_n\left(\mu, \sigma\right) = \sum_{i=1}^{n} \left[\psi'\left(\left(x_i - \mu\right)/\sigma\right)\left(x_i - \mu\right)/\sigma\right] \Big/ \sum_{i=1}^{n} \psi'\left(\left(x_i - \mu\right)/\sigma\right).$$

Assume that F is symmetric and $\psi\left(u\right)$ is odd (and hence $\psi'\left(u\right)\, u$ is also odd). Let $U = \left(X - \mu\right)/\sigma$. It is easy to see that $E_F\left[\psi'\left(U\right)U\right] = 0$ and that in this case $B_n\left(\mu, \sigma\right)$ converges almost surely to zero. If in addition $\sqrt{n}\left(\hat{\sigma}_n - \sigma\right) = O_P\left(1\right)$ (see Definition 7.1) we immediately obtain

$$\sqrt{n}\,\left(\hat{\mu}_n - \mu\right) \xrightarrow[n\to\infty]{\mathcal{W}} N\left(0, V\right),$$

where $\xrightarrow{\mathcal{W}}$ denotes weak convergence and

$$V = \sigma^2 \frac{E_F\left[\psi^2\left(\left(X - \mu\right)/\sigma\right)\right]}{\left\{E_F\left[\psi'\left(\left(X - \mu\right)/\sigma\right)\right]\right\}^2}.$$

In the case of asymmetric F we typically have $E_F\left[\psi'\left(U\right)U\right] \neq 0$. Hence, the term involving $\sqrt{n}\left(\hat{\sigma}_n - \sigma\right)$ in (2.67) will not vanish. Thus we need the corresponding Taylor expansion for $\hat{\sigma}_n$. Assume that $\hat{\sigma}_n$ is a M-scale, that is, it is given by an equation of the form

$$\sum_{i=1}^{n} \rho\left(\left(x_i - T_n\right)/\hat{\sigma}_n\right) = b', \tag{2.68}$$

for some function ρ not necessarily related with ψ in (2.65). As in Definition 2.6, T_n in (2.68) is some arbitrary location estimate and $b' \in (0, 1/2]$. A Taylor expansion of equation (2.68) yields

$$\sqrt{n}\,(\hat{\sigma}_n - \sigma) = \frac{1}{\sqrt{n}} C_n\,(T, \sigma) + \frac{1}{n} D_n\,(T, \sigma)\ \sqrt{n}\,(T_n - T) + R'_n, \tag{2.69}$$

where R'_n is the remaining term, T is the asymptotic value of T_n, and C_n and D_n are sums of independent random variables.

To be able to obtain a Taylor expansion of (2.68) that can be used in (2.67) we need an estimate T_n which is at the same time linearizable (i.e., it accepts a Taylor expansion) and does not depend on another scale estimate. If we use the same location estimate in the scale equation, that is, if we set $T_n = \hat{\mu}_n$, we are solving a system of two simultaneous equations as in (2.21). Estimates obtained in this way do not have satisfactory robustness properties (Martin and Zamar, 1993). As far as we know there is no robust estimate that simultaneously satisfies: (i) is location and scale equivariant; (ii) admits a Taylor expansion of first order; and (iii) does not depend on an scale estimate.

Our next result shows that this problem can be avoided if we use an S-scale $\hat{\sigma}_n$ in (2.65). The basic idea is that in this case the expansion (2.69) asymptotically does not depend on the distribution of $\sqrt{n}\,(T_n - T)$.

We need the following additional regularity conditions:

R.4 ρ is twice continuously differentiable;

P.5 ψ is twice continuously differentiable;

Theorem 2.6 - **Asymptotic normality** - *Assume the regularity conditions of Theorems 2.1 and 2.4. Assume that ψ satisfies P.1 to P.3 and P.5. Assume that ρ satisfies R.1 to R.4. Let $\hat{\mu}_n$ be the M-estimate of location given by (2.12), $\tilde{\mu}_n$ be the S-estimate of location and $\hat{\sigma}_n$ the corresponding S-estimate of scale, as defined in (2.17) and (2.16). Denote the almost sure finite limits of the sequences $\hat{\mu}_n$, $\tilde{\mu}_n$ and $\hat{\sigma}_n$ by $\mu(F)$, $\tilde{\mu}(F)$ and $\sigma(F)$ respectively. Assume that ψ and ρ also satisfy the following regularity conditions for any $F \in \mathcal{H}_\epsilon$:*

A1: $E_F \psi'(u) > 0$ *and finite;*

A2: $E_F [\psi'(u) \;\; u]$ *is finite;*

A3: $E_F [\rho'(\tilde{u}) \;\; u] \neq 0$ *and finite;*

A4: $E_F [\psi''(u)]$ *is finite;*

A5: $E_F [\rho''(\tilde{u})]$ *is finite;*

A6: $E_F [\psi''(u)\, u^2 + 2\, \psi'(u)\, u]$ *is finite;*

A7: $E_F [\rho''(\tilde{u})\, \tilde{u}^2 + 2\, \rho'(\tilde{u})\, \tilde{u}]$ *is finite;*

where $u = (x - \mu(F))/\sigma(F)$ and $\tilde{u} = (x - \tilde{\mu}(F))/\sigma(F)$. Then

$$\sqrt{n}\,(\hat{\mu}_n - \mu(F)) \longrightarrow N(0\,,\, V(\mu, \sigma, F)) \tag{2.70}$$

where

$$V(\mu,\sigma,F) = \sigma(F)^2 H(F)^2 E_F \Bigg\{ \Bigg[\psi\left(\frac{X-\mu(F)}{\sigma(F)}\right) - J(F) \times \left(\rho\left(\frac{X-\tilde{\mu}(F)}{\sigma(F)}\right) - b \right) \Bigg]^2 \Bigg\}, \tag{2.71}$$

$$H(F) = 1/E_F\left\{\psi'\left(\left(X-\mu(F)\right)/\sigma(F)\right)\right\},$$

and

$$J(F) = \frac{E_F\left\{\psi'\left(\left(X-\mu(F)\right)/\sigma(F)\right)\left(X-\mu(F)\right)/\sigma(F)\right\}}{E_F\left\{\rho'\left(\left(X-\tilde{\mu}(F)\right)/\sigma(F)\right)\left(X-\tilde{\mu}(F)\right)/\sigma(F)\right\}}$$

Proof: To simplify the notation let $\mu = \mu(F)$, $\sigma = \sigma(F)$, $\tilde{\mu} = \tilde{\mu}(F)$ and

$$u_i = (x_i - \mu)/\sigma.$$

A second order Taylor expansion of (2.12) around the limit values (μ, σ) yields

$$\left\{\frac{1}{n}\sum_{i=1}^{n}\psi'(u_i)\right\}\frac{1}{\sigma}\sqrt{n}\left(\hat{\mu}_n-\mu\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi(u_i) - \sqrt{n}\frac{(\hat{\sigma}_n-\sigma)}{\sigma}\frac{1}{n}\sum_{i=1}^{n}\psi'(u_i)\,u_i \tag{2.72}$$

$$+\frac{1}{2}\frac{1}{\sigma^2}\sqrt{n}\left(\hat{\mu}_n-\mu\right)^2\frac{1}{n}\sum_{i=1}^{n}\psi''(\tilde{u}_i)+ \tag{2.73}$$

$$+\frac{1}{2}\frac{1}{\sigma^2}\sqrt{n}\left(\hat{\sigma}_n-\sigma\right)^2\frac{1}{n}\sum_{i=1}^{n}\Big[\psi''(\tilde{u}_i)\,\tilde{u}_i^2 \tag{2.74}$$

$$+\,2\psi'(\tilde{u}_i)\,\tilde{u}_i\Big] \tag{2.75}$$

$$+\frac{1}{n}\frac{1}{\sigma^2}\sum_{i=1}^{n}\Big[\psi''(\tilde{u}_i)\,\tilde{u}_i+\psi'(\tilde{u}_i)\Big]\left(\hat{\sigma}_n-\sigma\right)\left(\hat{\mu}_n-\mu\right) \tag{2.76}$$

where $\tilde{u}_i = (x_i - \tilde{\mu})/\tilde{\sigma}$ and $(\tilde{\mu}, \tilde{\sigma})$ lies between $(\hat{\mu}_n, \hat{\sigma}_n)$ and (μ, σ). Let

$$B_n = \frac{1}{2}\frac{1}{\sigma}\left(\hat{\mu}_n - \mu\right)\frac{1}{n}\sum_{i=1}^{n}\psi''\left(\tilde{u}_i\right) , \tag{2.77}$$

$$C_n = \frac{1}{2}\frac{1}{\sigma}\left(\hat{\sigma}_n - \sigma\right)\ \frac{1}{n}\sum_{i=1}^{n}\left[\psi''\left(\tilde{u}_i\right)\tilde{u}_i^2 + 2\psi'\left(\tilde{u}_i\right)\tilde{u}_i\right] , \tag{2.78}$$

and

$$D_n = \frac{1}{n}\frac{1}{\sigma^2}\sum_{i=1}^{n}\left[\psi''\left(\tilde{u}_i\right)\tilde{u}_i + \psi'\left(\tilde{u}_i\right)\right]\left(\hat{\sigma}_n - \sigma\right) . \tag{2.79}$$

By hypothesis $\hat{\mu}_n - \mu = o_P(1)$ and, from assumption A4,

$$\frac{1}{n}\sum_{i=1}^{n}\psi''\left(\tilde{u}_i\right) = O_P(1) .$$

Then (7.3) implies that $B_n = o_P(1)$. Similarly we can show that $C_n = o_P(1)$ and $D_n = o_P(1)$. From (2.72)-(2.76) we have

$$\begin{aligned}\frac{1}{\sigma}\sqrt{n}\left(\hat{\mu}_n - \mu\right)\ &\left(\frac{1}{n}\sum_{i=1}^{n}\psi'\left(u_i\right) - B_n - D_n\right) \\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi\left(u_i\right) - \frac{1}{\sigma}\sqrt{n}\left(\hat{\sigma}_n - \sigma\right)\ \left(\frac{1}{n}\sum_{i=1}^{n}\psi'\left(u_i\right)u_i - C_n\right) .\end{aligned} \tag{2.80}$$

From equation (2.18) we get

$$\begin{aligned}\frac{1}{\sigma}\sqrt{n}\left(\hat{\sigma}_n - \sigma\right)&\left[\frac{1}{n}\sum_{i=1}^{n}\rho'\left(v_i\right)v_i - B_n' - D_n'\right] \\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\rho\left(v_i\right) - b - \frac{1}{\sigma}\sqrt{n}\left(\hat{\mu}_n - \tilde{\mu}\right)\left(\frac{1}{n}\sum_{i=1}^{n}\rho'\left(v_i\right) - C_n'\right) ,\end{aligned} \tag{2.81}$$

where, as before, $B_n' = o_P(1)$, $C_n' = o_P(1)$ and $D_n' = o_P(1)$. Note that

$$\frac{1}{n}\sum_{i=1}^{n}\rho'\left(v_i\right) = o_P(1) ,$$

and hence

$$\frac{1}{n}\sum_{i=1}^{n}\rho'(v_i) - C_n' = o_P(1). \tag{2.82}$$

From (2.19) we have

$$\frac{1}{\sigma}\sqrt{n}\left(\tilde{\mu}_n - \tilde{\mu}\right)\left(\frac{1}{n}\sum_{i=1}^{n}\rho''(v_i)\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\rho'(v_i) - \frac{1}{\sigma}\sqrt{n}\left(\hat{\sigma}_n - \sigma\right)\left(\frac{1}{n}\sum_{i=1}^{n}\rho''(v_i)\,v_i\right)$$

$$= O_P(1) - \frac{1}{\sigma}\sqrt{n}\left(\hat{\sigma}_n - \sigma\right) + O_P(1). \tag{2.83}$$

From (2.81), (2.83) and Lemma 7.12 we have

$$\frac{1}{\sigma}\sqrt{n}\left(\hat{\sigma}_n - \sigma\right) = \frac{1/\sqrt{n}\sum_{i=1}^{n}\rho(v_i) - b}{1/n\sum_{i=1}^{n}\rho'(v_i)\,v_i} + o_P(1)\ . \tag{2.84}$$

The theorem now follows from (2.80), (2.84) and Lemma 7.12. ■

2.3.7 Uniform asymptotic distribution for MM-location estimates

In this section we will show that the MM-location estimates $\hat{\mu}_n$ converge weakly to a normal distribution uniformly over $F \in \mathcal{H}_\epsilon$. Our main result is the following Theorem.

Theorem 2.7 *Suppose that all the assumptions of Theorems 2.1, 2.3, 2.5 and 2.6 hold. Then*

$$\sup_{F\in\mathcal{H}_\epsilon}\sup_{x\in\mathbb{R}}\left|P_F\left\{\sqrt{n}\,\frac{(\hat{\mu}_n - \mu)}{\sqrt{V}} < x\right\} - \Phi(x)\right| = o(1)\ ,$$

where $V = V(F)$ is given by (2.71).

To prove Theorem 2.7 we need to define uniform versions of $o_P(a_n)$, $O_P(a_n)$ and asymptotic normality, and that these quantities have analogous convergence properties to the usual (non-uniform) ones.

Definition 2.14 - **Uniform big O in probability**: *Let a_n, $n \geq 1$, be a sequence of real numbers and let X_n, $n \geq 1$, be a sequence of random variables. We say that $X_n = UO_P(a_n)$ over the set of distribution functions $\mathcal{H}_\epsilon$ if*

$$\lim_{k\to\infty} \sup_{F\in\mathcal{H}_\epsilon} \lim_{n\to\infty} P_F\left[\,\left|\frac{X_n}{a_n}\right| > k\,\right] = 0.$$

Definition 2.15 - **Uniform small o in probability**: *Let a_n, $n \geq 1$, be a sequence of real numbers and let X_n, $n \geq 1$, be a sequence of random variables. We say that $X_n = Uo_P(a_n)$ over the set of distribution functions $\mathcal{H}_\epsilon$ if $\forall\,\delta > 0$*

$$\lim_{n\to\infty} \sup_{F\in\mathcal{H}_\epsilon} P_F\left[\,\left|\frac{X_n}{a_n}\right| > \delta\,\right] = 0.$$

Definition 2.16 - **Uniformly asymptotically normal**: *We say that a sequence X_n, $n \in \mathbb{N}$ is uniformly asymptotically normal (UAN) over the set of distribution functions $\mathcal{H}_\epsilon$ if*

$$\sup_{F\in\mathcal{H}_\epsilon} \sup_{x\in\mathbb{R}} \Big| P_F(X_n \leq x) - \Phi(x) \Big| = o(1)\ . \tag{2.85}$$

Lemma 2.2 *Let X_n, $n \in \mathbb{N}$, be sequence of random variables that are uniformly asymptotically normal as in Definition 2.16. Then $X_n = UO_P(1)$.*

Proof: For any $K > 0$ we have

$$\begin{aligned} P_F\big(|X_n| > K\big) &= P_F\big(X_n > K\big) + P_F\big(X_n < -K\big) \\ &= 1 - P_F\big(X_n \leq K\big) + P_F\big(X_n < -K\big)\,. \end{aligned}$$

Fix $\tilde{\epsilon} > 0$. By (2.85) there exists $n_0 = n_0(\tilde{\epsilon})$ such that for all $n \geq n_0$

$$\left| P_F(X_n \leq x) - \Phi(x) \right| < \epsilon, \quad \forall x, \quad \forall F \in \mathcal{H}_\epsilon.$$

Hence, $P_F\big(|X_n| > K\big) \leq 1 - \Phi(K) + \Phi(-K) + 2\tilde{\epsilon}$, for all $n \geq n_0$ and for all $F \in \mathcal{H}_\epsilon$. Similarly we obtain $P_F\big(|X_n| > K\big) \geq 1 - \Phi(K) + \Phi(-K) - 2\tilde{\epsilon}$. Hence, given $\tilde{\epsilon} > 0$ we find $n_0(\tilde{\epsilon})$ such that for all $n \geq n_0$

$$\left| P_F\big(|X_n| > K\big) - \big[1 - \Phi(K) + \Phi(-K)\big] \right| < 2\tilde{\epsilon},$$

or, equivalently, $\lim_{n\to\infty} P_F\big(|X_n| > K\big) = 1 - \Phi(K) + \Phi(-K)$. It follows then that $\lim_{K\to\infty} \sup_{F\in\mathcal{H}_\epsilon} \lim_{n\to\infty} P_F\big(|X_n| > K\big) = 0$. ■

Lemma 2.3 *Let a_n, $n \in \mathbb{N}$, be a sequence of random variables such that $a_n = UO_P(1)$, and let b_n, $n \in \mathbb{N}$, be another sequence such that $b_n = Uo_P(1)$, then $a_n \times b_n = Uo_P(1)$.*

Proof: The Lemma follows easily from the following inequality, valid for any $\delta > 0$ and $K > 0$.

$$\begin{aligned} P_F(|a_n b_n| > \delta) &\leq P_F(|b_n| > \delta/|a_n|\,,\, |a_n| \leq K) + P_F(|a_n| > K) \\ &\leq P_F(|b_n| > \delta/K) + P_F(|a_n| > K)\,. \end{aligned}$$

■

Lemma 2.4 *Let a_n, $n \in \mathbb{N}$, be a sequence of random variables such that $a_n = UO_P(1)$, and let b_n, $n \in \mathbb{N}$, be another sequence such that there exists $b \neq 0$ with*

$b_n - b = Uo_P(1)$, *then*

$$\frac{a_n}{b_n} = \frac{a_n}{b} + Uo_P(1) \; .$$

Proof: We have

$$\frac{a_n}{b_n} - \frac{a_n}{b} = \frac{a_n}{b}\left(\frac{b}{b_n} - 1\right) \; . \tag{2.86}$$

We now show that

$$\frac{b}{b_n} - 1 = Uo_P(1) \; . \tag{2.87}$$

For simplicity assume that $b > 0$. The same argument, with appropriate modifications can be applied to the case $b < 0$. Fix $\delta > 0$. For any $0 < \tilde{\epsilon} < b < K$ we have

$$\begin{aligned} P_F\left(\left|\frac{b}{b_n} - 1\right| > \delta\right) &= P_F\left(|b - b_n| > |b_n|\,\delta\right) \\ &\leq P_F\left(|b - b_n| > |b_n|\,\delta\,, |b_n| \leq K\right) + P_F\left(|b_n| > K\right) \\ &\leq P_F\left(|b_n| \leq \tilde{\epsilon}\right) + P_F\left(|b - b_n| > |b_n|\,\delta\,, \tilde{\epsilon} < |b_n| \leq K\right) \\ &\quad + P_F\left(|b_n| > K\right) \\ &\leq P_F\left(|b_n| \leq \tilde{\epsilon}\right) + P_F\left(|b - b_n| > \tilde{\epsilon}\,\delta\right) + P_F\left(|b_n| > K\right) \; . \end{aligned}$$

Now note that $|b - b_n| > b - |b_n|$ and $|b - b_n| > |b_n| - b$ imply

$$P_F\left(|b_n| \leq \tilde{\epsilon}\right) \leq P_F\left(|b - b_n| > b - \tilde{\epsilon}\right) \; ,$$

and $P_F\left(|b_n| > K\right) \leq P_F\left(|b - b_n| > K - b\right)$. Choose an arbitrary $\tau > 0$. For fixed δ, $\tilde{\epsilon}$ and K choose $n_0(\tau)$ such that

$$P_F\left(|b - b_n| > b - \tilde{\epsilon}\right) < \tau/3\,,$$

$$P_F\left(|b - b_n| > K - b\right) < \tau/3\,,$$

and

$$P_F\left(|b - b_n| > \tilde{\epsilon}\,\delta\right) < \tau/3 .$$

for all $F \in \mathcal{H}_\epsilon$ and $n \geq n_0$. It follows that for $n \geq n_0$ we have

$$P_F\left(\left|\frac{b}{b_n} - 1\right| > \delta\right) < \tau \qquad \forall\, F \in \mathcal{H}_\epsilon .$$

Hence (2.87) holds. The result now follows from (2.86) and Lemma 2.3. ■

Lemma 2.5 *Let a_n, $n \in \mathbb{N}$, be a sequence of random variables such that $a_n = UO_P(1)$, and let b_n, $n \in \mathbb{N}$, be another sequence such that there exists b with $b_n - b = Uo_P(1)$, then $a_n\, b_n = a_n\, b + Uo_P(1)$.*

Proof: This follows immediately by noting that $a_n\, b_n - a_n\, b = a_n\, (b_n - b) = UO_P(1) \times Uo_P(1)$ and applying Lemma 2.3. ■

Lemma 2.6 *Let $D_1, \ldots, D_n$ be n independent and identically distributed random variables and let $\overline{D}_n = 1/n \sum_{i=1}^n D_i$. Assume that $E_F\left[D_i^2\right] \leq c < \infty$, for all $F \in \mathcal{H}_\epsilon$. Then $\overline{D}_n = UO_P(1)$ and $\overline{D}_n - E_F(D_i) = Uo_P(1)$.*

Proof: Note that the assumption on the second moment of D_i implies that $E_F\,|D_i| \leq 1 + c$ for all $F \in \mathcal{H}_\epsilon$. To simplify the notation, let $d = 1 + c$. Then we have

$$P_F\left[\left|\overline{D}_n\right| > 2\,d\right] \leq P_F\left[\left|\overline{D}_n - E_F D_i\right| > d\right] \leq \frac{1}{d^2}\frac{1}{n}\mathrm{Var}_F\left(D_i\right) \leq \frac{1}{d^2}\frac{1}{n} E_F D_i^2 .$$

Hence, $\lim_n P_F\left[\left|\overline{D}_n\right| > 2\,d\right] = 0$ for all $F \in \mathcal{H}_\epsilon$, where d does not depend on F. It follows that

$$\lim_{k\to\infty} \sup_{F\in\mathcal{H}_\epsilon} \lim_{n\to\infty} P_F\left[\left|\overline{D}_n\right| > k\right] = 0\,,$$

that is, $\overline{D}_n = UO_P\,(1)$. A similar argument shows that $\overline{D}_n - E_F\,(D_i) = Uo_P\,(1)$. ■

Lemma 2.7 *If $a_n = Uo_P\,(1)$ and X_n is UAN (see Definition 2.16) then $X_n + a_n$ is UAN. That is:*

$$\sup_{F\in\mathcal{H}_\epsilon} \sup_{x\in\mathbb{R}} \left|P_F\left\{X_n + a_n < x\right\} - \Phi\,(x)\right| = o\,(1)\ . \tag{2.88}$$

Proof: First note that for any $x \in \mathbb{R}$, $\delta > 0$ and $F \in \mathcal{H}_\epsilon$

$$\begin{aligned} P_F\left[X_n + a_n < x\right] &\leq P_F\left[X_n + a_n < x\,,\ |a_n| \leq \delta\right] + P_F\left[|a_n| > \delta\right] \\ &\leq P_F\left[X_n < x - \delta\right] + P_F\left[|a_n| > \delta\right]\ . \end{aligned} \tag{2.89}$$

Similarly we have

$$\begin{aligned} P_F\left[X_n + a_n \geq x\right] &\leq P_F\left[|a_n| > \delta\right] + P_F\left[X_n + a_n \geq x\,,\ |a_n| \leq \delta\right] \\ &\leq P_F\left[X_n \geq x - \delta\right] + P_F\left[|a_n| > \delta\right]\ , \end{aligned}$$

which yields

$$P_F\left[X_n + a_n < x\right] \geq P_F\left[X_n < x - \delta\right] - P_F\left[|a_n| > \delta\right]\ . \tag{2.90}$$

Equations (2.89) and (2.90) together yield

$$-P_F\left[|a_n| > \delta\right] \leq P_F\left[X_n + a_n < x\right] - P_F\left[X_n < x - \delta\right] \leq P_F\left[|a_n| > \delta\right]\ . \tag{2.91}$$

To simplify the notation, let $u_n(\delta, F) = P_F[|a_n| > \delta]$. We have

$$-u_n(\delta, F) \leq P_F[X_n + a_n < x] - \Phi(x) + \Phi(x - \delta) - P_F[X_n < x - \delta] \\ + \Phi(x) - \Phi(x - \delta) \leq u_n(\delta, F) \,. \quad (2.92)$$

Let $\epsilon > 0$ be arbitrary. Choose $\delta = \delta(\epsilon) > 0$ such that

$$\sup_{x \in \mathbb{R}} |\Phi(x - \delta) - \Phi(x)| < \epsilon/3 \,. \quad (2.93)$$

For this δ choose $n_0 = n_0(\delta)$ such that $\sup_{F \in \mathcal{H}_\epsilon} |u_n(\delta, F)| < \epsilon/3$. Choose $n_1 \doteq n_1(\epsilon)$ such that

$$\sup_{F \in \mathcal{H}_\epsilon} \sup_{x \in \mathbb{R}} |\Phi(x - \delta) - P_F[X_n < x - \delta]| = \sup_{F \in \mathcal{H}_\epsilon} \sup_{x \in \mathbb{R}} |\Phi(x) - P_F[X_n < x]| < \epsilon/3 \,. \quad (2.94)$$

Let $b_n(x, F) = \Phi(x - \delta) - P_F[X_n < x - \delta]$ and $c(x) = \Phi(x - \delta) - \Phi(x)$. Then, equation (2.92) can be written as

$$-\epsilon/3 < P_F[X_n + a_n < x] - \Phi(x) + b_n(x, F) + c(x) < \epsilon/3 \,. \quad (2.95)$$

We know that for $n \geq \max(n_0, n_1)$, $\sup_F \sup_x |b_n(x, F)| < \epsilon/3$, $\sup_x |c(x)| < \epsilon/3$. Let $d_n(x, F) = P_F[X_n + a_n < x] - \Phi(x)$. Equation (2.95) implies

$$-\epsilon/3 - b_n(x, F) - c(x) < d_n(x, F) < \epsilon/3 - b_n(x, F) - c(x) \;, \quad (2.96)$$

and we immediately obtain that for n sufficiently large (not depending on x or F), $-\epsilon < d_n(x, F) < \epsilon$. ■

Remark 2.1 Let f be a real function such that

$$E_{F_0}[f(X, t, s)] = \int f(X, t, s) \, dF_0(X) > 0$$

for any $t \in \mathbb{R}$ and $s > 0$, where F_0 denotes the central distribution of the contamination neighbourhood $\mathcal{H}_\epsilon$. It is easy to see that if $E_{F_0}\left[f\left(X,t,s\right)\right]$ is a continuous function of (t,s) and $\mathcal{K}_t$ and $\mathcal{K}_s$ are compact sets in the real line such that $\mathcal{K}_s \subset (0,\infty)$ then we have

$$\inf_{F\in\mathcal{H}_\epsilon,\, t\in\mathcal{K}_t,\, s\in\mathcal{K}_s} E_F\left[f\left(X,t,s\right)\right] > 0\,.$$

In particular, if $\sigma\left(F\right)$ denotes a scale estimate that satisfies (2.29) and $\mu\left(F\right)$ is an M-location with general scale calculated with a function ψ_c in Huber's family (2.5) then

$$\inf_{F\in\mathcal{H}_\epsilon} \operatorname{Var}_F\left[\psi_c\left(\left(X-\mu\left(F\right)\right)/\,\sigma\left(F\right)\right)\right] > 0\,. \tag{2.97}$$

Proof of Theorem 2.7: To simplify the notation, in what follows let $\mu = \mu\left(F\right)$, $\tilde{\mu} = \tilde{\mu}\left(F\right)$ and $\sigma = \sigma\left(F\right)$. The idea of the proof is to show that $\sqrt{n}\left(\hat{\mu}_n - \mu\right)$ can be represented as a linear term plus a uniformly small remainder. We use the Berry Esseen Theorem to show that the linear part is UAN (see Definition 2.16) and Lemma 2.7 to show that the sum of these terms is also UAN.

We now show that

$$\sqrt{n}\frac{\left(\hat{\mu}_n-\mu\right)}{\sqrt{V}} = \sqrt{n}\frac{\overline{W}_n}{\sqrt{V}} + Uo_P\left(1\right)\,. \tag{2.98}$$

where

$$\begin{aligned} W_i &= \Big(\psi\left(\left(x_i-\mu\right)/\,\sigma\right) - d\,\left(\rho\left(\left(x_i-\tilde{\mu}\right)/\,\sigma\right)-b\right)\Big)\Big/\,e\,, \\ d &= \frac{E_F\left\{\psi'\left(\left(X-\mu\right)/\sigma\right)\left(X-\mu\right)/\sigma\right\}}{E_F\left\{\rho'\left(\left(X-\tilde{\mu}\right)/\sigma\right)\left(X-\tilde{\mu}\right)/\sigma\right\}} \\ e &= E_F\left\{\psi'\left(\left(X-\mu\right)/\sigma\right)\right\}\,. \end{aligned} \tag{2.99}$$

Theorems 2.1, 2.3 and 2.5 show that that $\hat{\sigma}_n - \sigma = Uo_P(1)$, $\tilde{\mu}_n - \tilde{\mu} = Uo_P(1)$ and $\hat{\mu}_n - \mu = Uo_P(1)$ respectively. From (2.72) to (2.76) and (2.77) to (2.79), using Lemmas 2.3 and 2.6 we have

$$\frac{1}{\sigma}\sqrt{n}\left(\hat{\mu}_n - \mu\right) \left(\frac{1}{n}\sum_{i=1}^{n}\psi'\left(u_i\right) + Uo_P\left(1\right)\right)$$
$$= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi\left(u_i\right) - \frac{1}{\sigma}\sqrt{n}\left(\hat{\sigma}_n - \sigma\right)\left(\frac{1}{n}\sum_{i=1}^{n}\psi'\left(u_i\right)u_i + Uo_P\left(1\right)\right). \quad (2.100)$$

From (2.81) and (2.82) we have

$$\frac{1}{\sigma}\sqrt{n}\left(\hat{\sigma}_n - \sigma\right)\left[\frac{1}{n}\sum_{i=1}^{n}\rho'\left(v_i\right)v_i + Uo_P\left(1\right)\right]$$
$$= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\rho\left(v_i\right) - b - \frac{1}{\sigma}\sqrt{n}\left(\tilde{\mu}_n - \tilde{\mu}\right) \times Uo_P\left(1\right).$$

Similarly, from equation (2.83) we have

$$\frac{1}{\sigma}\sqrt{n}\left(\tilde{\mu}_n - \tilde{\mu}\right)\left(\frac{1}{n}\sum_{i=1}^{n}\rho''\left(v_i\right)\right) = UO_P\left(1\right) - \frac{1}{\sigma}\sqrt{n}\left(\hat{\sigma}_n - \sigma\right) + UO_P\left(1\right).$$

From the last two equations we obtain

$$\frac{1}{\sigma}\sqrt{n}\left(\hat{\sigma}_n - \sigma\right)\Big[a + Uo_P\left(1\right)\Big] = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\rho\Big(\left(v_i\right) - b\Big) + Uo_P\left(1\right), \quad (2.101)$$

where $a = E_F\left\{\rho'\left(\left(X - \tilde{\mu}\right)/\sigma\right)\left(X - \tilde{\mu}\right)/\sigma\right\}$. From (2.100) and (2.101) we have

$$\frac{1}{\sigma}\sqrt{n}\left(\hat{\mu}_n - \mu\right)\Big[c + Uo_P\left(1\right)\Big]$$
$$= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi\left(u_i\right) - \left[\frac{1}{a}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big(\rho\left(v_i\right) - b\Big) + Uo_P\left(1\right)\right]\Big[d + Uo_P\left(1\right)\Big],$$

where $c = E_F\left\{\psi'\left(\left(X - \mu\right)/\sigma\right)\right\}$, and $d = E_F\left\{\psi'\left(\left(X - \mu\right)/\sigma\right)\left(X - \mu\right)/\sigma\right\}$. Hence,

$$\frac{1}{\sigma}\sqrt{n}\left(\hat{\mu}_n - \mu\right)\Big[c + Uo_P\left(1\right)\Big] = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi\left(u_i\right) - \frac{d}{a}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big(\rho\left(v_i\right) - b\Big) + Uo_P\left(1\right).$$

From the last equation and Lemma 2.4 we obtain (2.98).

Note that $|W_i|$ are bounded (see (2.99)), and hence their moments are bounded uniformly for $F \in \mathcal{H}_\epsilon$. Using (2.97) we see that their variance is bounded away from zero uniformly on $F \in \mathcal{H}_\epsilon$. The Berry Esseen Theorem (Chow and Teicher, 1988, page 305) yields

$$\sup_{F\in\mathcal{H}_\epsilon} \sup_{x\in\mathbb{R}} \left| P_F \left\{ \frac{\sqrt{n}\,\overline{W}_n}{\sqrt{V}} < x \right\} - \Phi\left(x\right) \right| = o\left(1\right) .$$

Hence we have

$$\sqrt{n}\frac{(\hat{\mu}_n - \mu)}{\sqrt{V}} = V_n + Uo_P\left(1\right) ,$$

where V_n is UAN. Lemma 2.7 completes the proof. ■

Validity of the required regularity conditions to obtain uniform asymptotic normality

The required conditions to obtain uniform consistency and asymptotic normality are satisfied by estimates with breakdown point 50% obtained with functions ρ_d in Tukey's family (2.8) when the proportion of contamination ϵ is up to 10%. If we consider estimates in the same family with breakdown point 40% the conditions hold for $\epsilon \leq 0.15$. There seems to be a trade-off between the breakdown point of the estimate and the extent to which the uniform consistency holds. Uniform results for contamination neighbourhoods with $\epsilon \leq 0.10$ are nevertheless of practical interest. There are reports in the literature suggesting that most data sets have fractions of contamination ranging between 0 and 10%. (Hampel, *et al.*, 1986).

Chapter 3

Robust bootstrap for the location-scale model

In this chapter we introduce a new computer intensive method of inference based on Efron's bootstrap (Efron, 1979; Efron, 1982; Hall, 1992; Efron and Tibshirani, 1993). To distinguish Efron's bootstrap from the method presented here we will refer to the former as "classical bootstrap". Efron's method applies to the problem of estimating the sampling distribution of a complex statistic. It is based on the following principle. Suppose we are interested in the sampling distribution of the statistic $\hat{\mu}_n$ when the data have distribution function F. If we knew F we could obtain either the exact distribution of $\hat{\mu}_n$ or an approximation to it via Monte Carlo simulations. The idea behind Efron's bootstrap is to use an estimate of F in order to estimate the distribution of $\hat{\mu}_n$. If F is assumed to belong to a parametric family, $F = F_\theta$, then, given an estimate $\hat{\theta}$ for θ, we can use $\hat{F} = F_{\hat{\theta}}$ to estimate F_θ. This method is called parametric bootstrap. If we do not assume an underlying parametric family for F, then a natural non-parametric estimate is the sample distribution function $\hat{F} = F_n$. In this case we call the method non-parametric bootstrap. Throughout this thesis we only assume that the distribution F that generated the data belongs to a contamination neighborhood $\mathcal{H}_\epsilon$ of a certain central distribution (see Section 2.2). Hence, we will focus on non-parametric bootstrap methods.

For both types of bootstrap (parametric and non-parametric) we need to calculate the distribution of $\hat{\mu}_n$ assuming data generated by $\hat{F}$. However, in most cases it is very difficult to obtain an explicit expression for this distribution of $\hat{\mu}_n$. In those cases we can use computer simulations to get an estimate of this law. One generates several thousand samples from $\hat{F}$ and re-calculates $\hat{\mu}_n$ for each of these samples. The empirical distribution of those replicated values of $\hat{\mu}_n$ gives an estimate of the desired sampling distribution. In what follows we will also refer to this empirical distribution as the "bootstrap distribution estimate".

When each evaluation of $\hat{\mu}_n$ is computationally demanding, re-calculating the statistic many times can make the method too slow to be of practical interest. Robust estimates are not easy to calculate. For example, S-scales (2.16) solve an optimization problem where the function being minimized is implicitly defined. The MM-location estimate is calculated by solving the non-linear equation in (2.12). The numerical complexity of robust estimates for the linear model is significantly greater (see Chapter 4).

Another potential problem when using the classical bootstrap on data that might contain outliers arises with the tails of the distribution estimate. Intuitively, the problem is that the outliers may appear in the bootstrap samples in larger proportions than in the original sample. For example, if there is one outlier among 10 data points, we expect that over 3% of the bootstrap samples of size 10 will contain the outlier 5 or more times. This means that over 3% of the re-sampled statistics may be severely affected by the single outlier. In other words, the corresponding tail of the empirical distribution of the re-calculated estimates might be unreliable. A symmetric 95% confidence interval might then be affected because it uses the estimated 2.5% and 97.5% quantiles, and at least one of them can be influenced by the outlier. This problem has been quantified by Singh (1998). He defined the breakdown point of bootstrap quantile estimates and showed that even robust estimates do not yield bootstrap quantile estimates with satisfactory breakdown points. To remedy this problem he proposes to re-sample from the Winsorized data (see Section 3.6.1). Unfortunately this variant of the bootstrap method does not reduce the computational requirements of the re-calculation scheme.

The large amount of data that businesses and government agencies can collect and store with the available information technology makes large-scale applications a reality. Hence, it is of practical interest to study the feasibility of estimation methods when applied to moderate and large data sets. Compared to the classical bootstrap, our method, which we call "robust bootstrap", is significantly faster, more stable when the data are contaminated and produces comparable results when applied to data that do not contain outliers.

To illustrate the gain in speed of calculation of our proposal consider the simple case of constructing a 95% confidence interval for the location parameter μ when the observations x_i satisfy $x_i = \mu + \epsilon_i$. We assume that the errors ϵ_i are independent observations with unknown variance. We generated an artificial data set of 1,000 independent standard normal observations (i.e. we set $\mu = 0$ in the model above) and built a 95% confidence interval for the population mean based on an MM-location estimate (see Definition 2.9). We used the function $\psi_{1.345}$ in Huber's family. The S-scale $\hat{\sigma}_n$ was calculated with the function $\rho_{1.04086}$ in (2.14). The basic percentile confidence interval based on the classical bootstrap with 3,000 bootstrap samples was $(-0.05341, 0.07601)$ (for the definition of these bootstrap confidence intervals see Davison and Hinkley, 1997, page 194). It took 1545 CPU seconds (that is around 25 CPU minutes) to finish on a dual-CPU Sun Sparc Ultra 4 (each CPU a 296 Megahertz SUNW UltraSPARC-II) with 1.1 Gigabytes of RAM memory running SunOS 5.7. On the other hand, the basic percentile confidence interval using the robust bootstrap based on the same number of bootstrap samples took less than 3 CPU seconds. The 95% confidence interval was $(-0.05309, 0.07574)$. Figure 3.1 displays a comparison of the 3,000 re-calculated $\hat{\mu}_n^*$'s with each method. Note that the boxplots look very similar. This example illustrates that when the data do not contain outliers both methods are comparable, and the robust bootstrap is significantly faster to compute.

The improved stability of the robust bootstrap can be illustrated with the following simple example. Consider a random sample of size 100 containing 65 observations that follow a standard normal distribution and 35 observations with dis-

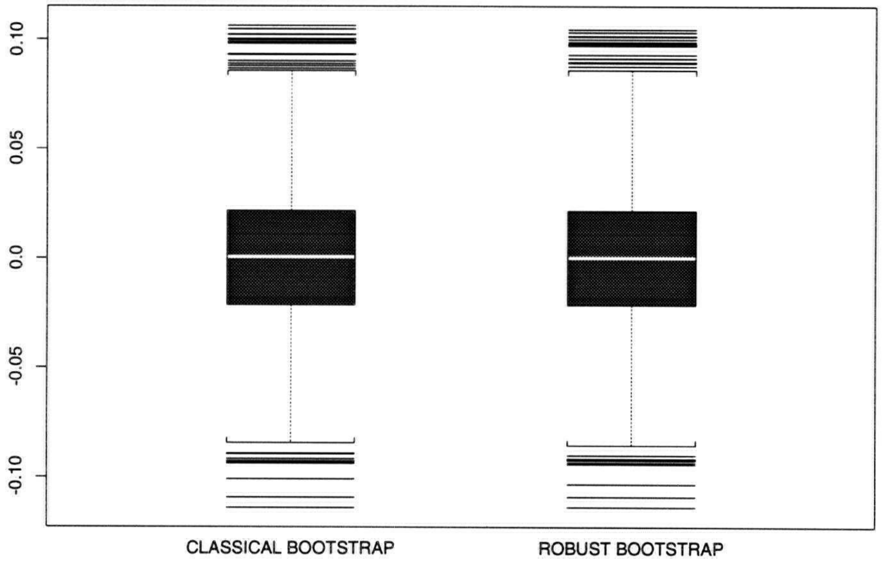


Figure 3.1: Boxplots of 3,000 re-calculated MM-location estimates with the classical and robust bootstrap. The artificial data set contains 1,000 independent standard normal observations without contamination.

tribution function $G(u) = \Phi\left((u-10)/\sqrt{0.5}\right)$, where Φ denotes the standard normal cumulative distribution function. In other words, this sample contains 35% of outliers centered around 10. We used the same MM-location estimate as above. Figure 3.2 contains the boxplots of the 3,000 re-calculated MM-location estimates for each bootstrap method, and a simulated data set from the actual asymptotic distribution of $\hat{\mu}_n$. Note that the tails of the robust bootstrap re-calculated MM-location estimates are more stable than those corresponding to the classical bootstrap method. The new method presented here also gives quantile estimates with higher breakdown points than those obtained with the classical bootstrap (see Section 3.4). The intuitive reason is that we use weights based on the robust estimate to re-calculate the statistic. Hence outlying points will typically be associated with small weights and have small impact on the bootstrapped estimate.

The rest of this chapter is organized as follows. Section 3.1 presents the resampling method for the location-scale model. Section 3.2 contains a one-sample and a two-sample example of statistical inference performed with the robust bootstrap. Section 3.3 studies the asymptotic behaviour of the robust bootstrap. Section 3.4 discusses the breakdown point of the quantiles estimates obtained with the robust bootstrap. Section 3.5 proposes a way to studentize the robust bootstrap in order to improve on its order of convergence. Finally, Section 3.6 contains two Monte Carlo comparison studies on the performance of the robust bootstrap to estimate asymptotic variances and to build confidence intervals.

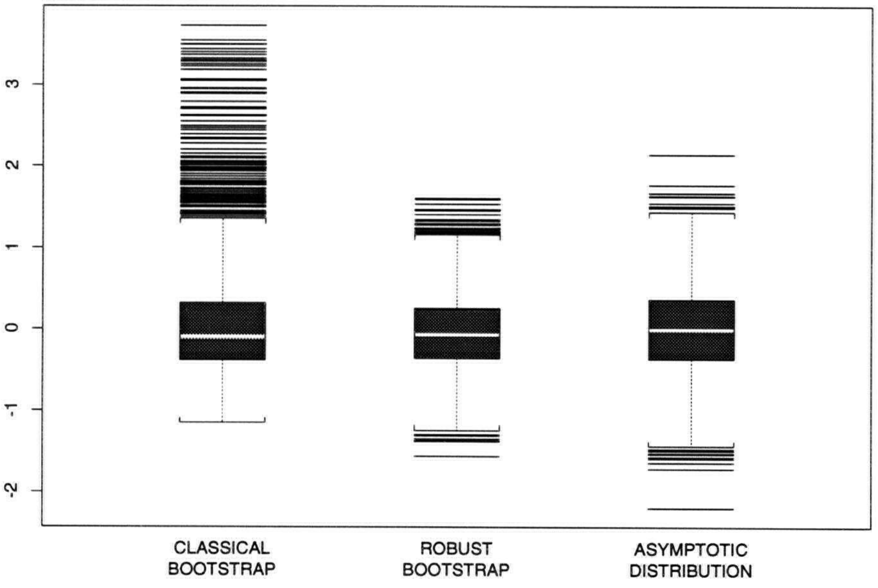


Figure 3.2: Boxplots of 3,000 re-calculated MM-location estimates with the classical and robust bootstrap. The artificial data set contains 100 independent observations; 65 of them follow a standard normal distribution, while the remaining 35 have distribution function $G(u) = \Phi\left((u-10)/\sqrt{0.5}\right)$, where Φ denotes the standard normal cumulative distribution function. The boxplot labeled "Asymptotic Distribution" contains a sample of 3,000 observations simulated from the actual asymptotic distribution of $\hat{\mu}_n$.

3.1 Definitions

The intuitive idea behind our method was briefly discussed in Section 1.5 for the simple case when the scale parameter σ is known. Here we present the method for the case of unknown σ. We use MM-location estimates (see Definition 2.9).

Let $x_1, \ldots, x_n$ be i.i.d. observations following model (2.1). Assume that the x_i's have distribution function F belonging to the contamination neighborhood $\mathcal{H}_\epsilon$ defined in (2.22). Let $\hat{\mu}_n$ be an MM-location calculated with an S-scale $\hat{\sigma}_n$, and let $\tilde{\mu}_n$ be the associated S-location estimate.

As in Section 1.5, we are interested in making statistical inferences about the location parameter μ. We consider two methods to achieve this goal. The first alternative is to use the result of Theorem 2.6 on the asymptotic normality of the sequence $\sqrt{n}\,(\hat{\mu}_n - \mu)$. To use this method we only have to estimate the variance of $\hat{\mu}_n$. The second option is to directly estimate the distribution function of $\sqrt{n}\,(\hat{\mu}_n - \mu)$. We can then use this distribution estimate to approximate the quantiles needed to construct confidence intervals (see for example, Efron and Tibshirani, 1993, and Davison and Hinkley, 1997).

We propose to use the following computer intensive method to generate a large number of re-calculated $\hat{\mu}_n^*$'s. These re-computed statistics can be used to estimate both the variance and the distribution function of the sequence $\hat{\mu}_n$. For the first objective we can use the empirical variance of the $\hat{\mu}_n^*$'s. A natural estimate of the distribution function F_n of $\hat{\mu}_n$ is given by the empirical distribution function of the re-computed statistics.

Recall that in the discussion in Section 2.3.6 we noted that for data generated by an arbitrary distribution F in the contamination neighborhood $\mathcal{H}_\epsilon$, the location and scale estimates are not necessarily asymptotically independent. Hence, to estimate the distribution of $\hat{\mu}_n$ we have to take into account the behaviour of the scale estimate $\hat{\sigma}_n$. Intuitively this is the reason why in what follows we re-calculate both estimates to incorporate the information obtained from the re-computed $\hat{\sigma}_n^*$'s into the final re-calculated $\hat{\mu}_n^*$'s.

For each $1 \leq i \leq n$ define the residuals $r_i = x_i - \hat{\mu}_n$ and $\tilde{r}_i = x_i - \tilde{\mu}_n$ associated with the MM- and the S-location estimates respectively. We first write $\hat{\mu}_n$ and $\hat{\sigma}_n$ as weighted averages. Define the weights ω_i and v_i as

$$\begin{aligned} \omega_i &= \psi\left(r_i / \hat{\sigma}_n\right) / \, r_i \,, \\ v_i &= \frac{1}{n\, b}\, \rho\left(\tilde{r}_i / \hat{\sigma}_n\right) / \, \tilde{r}_i \,, \qquad 1 \leq i \leq n \,. \end{aligned} \tag{3.1}$$

Simple computations yield

$$\hat{\mu}_n = \sum_{i=1}^n \omega_i \, x_i \Bigg/ \sum_{i=1}^n \omega_i \,, \tag{3.2}$$

$$\hat{\sigma}_n = \sum_{i=1}^n v_i \; \left(x_i - \tilde{\mu}_n\right) \,. \tag{3.3}$$

Clearly this representation does not help in calculating the estimates because the right-hand side depends on $\hat{\mu}_n$ and $\hat{\sigma}_n$, but it motivates the robust bootstrap procedure.

Let x_i^*, $i = 1, \ldots, n$ be a non-parametric bootstrap sample from the observations. That is: x_i^* are i.i.d. random variables with distribution function assigning probability $1/n$ to each of the points in the original sample. Define the random variables $\hat{\mu}_n^*$ and $\hat{\sigma}_n^*$ by

$$\hat{\mu}_n^* = \sum_{i=1}^{n} \omega_i^* \, x_i^* \Big/ \sum_{i=1}^{n} \omega_i^* \,, \tag{3.4}$$

$$\hat{\sigma}_n^* = \sum_{i=1}^{n} v_i^* \left(x_i^* - \tilde{\mu}_n \right) \,, \tag{3.5}$$

where $\omega_i^* = \psi\,(r_i^* / \hat{\sigma}_n) / \, r_i^*$, $v_i^* = \rho\,(\tilde{r}_i^* / \hat{\sigma}_n) / \, (n \; b \; \tilde{r}_i^*)$, $r_i^* = x_i^* - \hat{\mu}_n$, and $\tilde{r}_i^* = x_i^* - \tilde{\mu}_n$ for $1 \leq i \leq n$. Note that the estimates $\hat{\mu}_n$, $\tilde{\mu}_n$ and $\hat{\sigma}_n$ involved in ω_i^* and v_i^* are based on the original data, not the bootstrap samples.

The re-calculated $\hat{\mu}_n^*$ and $\hat{\sigma}_n^*$ obtained in (3.4) and (3.5) may not reflect the actual variability of the random vector $(\hat{\mu}_n, \hat{\sigma}_n)'$ due to the fact that the estimates $\hat{\mu}_n$, $\hat{\sigma}_n$ and $\tilde{\mu}_n$ used in the weights ω_i and v_i are those evaluated from the original sample. We now give an intuitive argument on how to derive a correction factor to remedy this unwanted phenomenon. Think of (3.2) and (3.3) as a fixed-point equation of the form $(\hat{\mu}_n, \hat{\sigma}_n)' = \mathbf{f}\,(\hat{\mu}_n, \hat{\sigma}_n)$ for certain $\mathbf{f} : \mathbb{R}^2 \to \mathbb{R}^2$. Let μ and σ be the almost sure limits of $\hat{\mu}_n$ and $\hat{\sigma}_n$ respectively. A first-order Taylor expansion of $\mathbf{f}$ around μ and σ suggests that we should multiply the re-calculated pairs $(\hat{\mu}_n^*, \hat{\sigma}_n^*)'$ by the matrix $[\mathbf{I} - \nabla \mathbf{f}\,(\mu, \sigma)]^{-1}$, where $\nabla \mathbf{f}$ denotes the matrix of first derivatives of $\mathbf{f}$. We estimate this factor with $[\mathbf{I} - \nabla \mathbf{f}\,(\hat{\mu}_n, \hat{\sigma}_n)]^{-1}$. Hence, the re-calculated $\hat{\mu}_n^{R*} - \hat{\mu}_n$ with the robust bootstrap is a linear combination of $\hat{\mu}_n^*$ and $\hat{\sigma}_n^*$ obtained in (3.4) and (3.5).

Recall that our interest lies in estimating the sampling distribution of $\sqrt{n}\,(\hat{\mu}_n - \mu)$.

The robust bootstrap "replicates" of this expression, $\hat{\mu}_n^{R*} - \hat{\mu}_n$, are given by

$$\hat{\mu}_n^{R*} - \hat{\mu}_n = a_n \left(\hat{\mu}_n^* - \hat{\mu}_n \right) + b_n \left(\hat{\sigma}_n^* - \hat{\sigma}_n \right), \tag{3.6}$$

where $\hat{\mu}_n$ and $\hat{\sigma}_n$ are the location and scale estimates obtained with the original sample,

$$\begin{aligned} a_n &= \hat{\sigma}_n \left[\sum_{i=1}^n \psi\left(r_i / \hat{\sigma}_n \right) / r_i \right] \Big/ \sum_{i=1}^n \psi'\left(r_i / \hat{\sigma}_n \right) , \\ b_n &= c_n^{-1} \left[\sum_{i=1}^n \psi'\left(r_i / \hat{\sigma}_n \right) r_i \right] \Big/ \sum_{i=1}^n \psi'\left(r_i / \hat{\sigma}_n \right) , \\ c_n &= \hat{\sigma}_n^2 \, \frac{1}{n} \frac{1}{b} \sum_{i=1}^n \left[\rho'\left(\tilde{r}_i / \hat{\sigma}_n \right) \tilde{r}_i / \hat{\sigma}_n \right] , \end{aligned} \tag{3.7}$$

and ψ' and ρ' denote the derivatives of ψ and ρ respectively.

Remark 3.1 - *Computational Ease*: To estimate the distribution of $\hat{\mu}_n - \mu$ with the classical bootstrap we have to re-calculate $\hat{\sigma}_n$ as well as $\hat{\mu}_n$. With each bootstrap sample, to determine $\hat{\sigma}_n$ we have to minimize the function $s_n(t)$, $t \in \mathbb{R}$, implicitly defined as the solution of (see (2.15))

$$\frac{1}{n} \sum_{i=1}^n \rho\left((x_i - t) / s_n(t) \right) = b .$$

Once the minimum $\hat{\sigma}_n^*$ of $s_n(t)$ is found we have to solve for $\hat{\mu}_n^*$

$$\sum_{i=1}^n \psi\left((x_i - \hat{\mu}_n^*) / \hat{\sigma}_n^* \right) = 0 .$$

These optimization problems have to be solved several thousand times. On the other hand, to re-calculate $\hat{\mu}_n^{R*} - \hat{\mu}_n$ with the robust bootstrap we use the weighted averages (3.4) and (3.5). The correction factors a_n, b_n and c_n are computed only once with the full sample. In the location-scale examples we considered, the robust bootstrap required less than 0.2% of the CPU time needed to compute the same number of classical bootstrap re-calculations.

Remark 3.2 - *Robustness*: Note that both Huber's and Tukey's families of functions ψ's yield weights $\omega_i(u) = \psi(u)/u$ that are decreasing functions of $|u|$. Outlying points will typically have large residuals and hence be associated with small weights in equations (3.4) and (3.5). Note that when using a function ψ_d from Tukey's family, extreme outliers (those with a residual $|r_i| > d\,\hat{\sigma}_n$) receive a zero weight, and hence have no effect on the re-calculated estimate. For ρ_d in Tukey's family, the resulting weights v_i used in re-calculating the scale are also decreasing in the absolute value of the residuals. This makes outlying points less influential in the re-calculated $\hat{\sigma}_n^*$. See Section 3.4 for a formal discussion of the robustness properties of this method.

3.2 Examples

3.2.1 One sample location-scale: Blood pressure

Consider the following data set obtained from a hypertension screening program (Rosner, 1977). Ten monthly observations were obtained for each patient. The data for a particular individual are: 40, 75, 80, 83, 86, 88, 90, 92, 93 and 95. We wish to estimate the individual's mean blood-pressure μ_0.

Let $\hat{\mu}_n$ denote the MM-location estimate for μ_0 calculated with the function $\psi_{1.345}$ in Huber's family and S-scale $\hat{\sigma}_n$ obtained with $\rho_{1.04086}$ in (2.14). With the above data set we obtain $\hat{\mu}_n = 86.03$. In order make inferences about μ_0 we would like to have an estimation of the distribution of $\hat{\mu}_n - \mu_0$ for this patient. To this end we use both the classical bootstrap and the robust bootstrap and compare their performance. We generated 50,000 bootstrap samples from the data and calculated $\hat{\mu}_n^{C*} - \hat{\mu}_n$, where $\hat{\mu}_n^{C*}$ denotes the classical bootstrap re-calculated $\hat{\mu}_n$. We also computed 50,000 robust bootstrap $\hat{\mu}_n^{R*} - \hat{\mu}_n$.

Figure 3.3 contains boxplots for both $\hat{\mu}_n^{C*} - \hat{\mu}_n$ and $\hat{\mu}_n^{R*} - \hat{\mu}_n$. Note that the robust bootstrap empirical distribution is less skewed than that of the classical bootstrap. We see that the lower tail of the classical bootstrap distribution estimate is heavier than that of the robust bootstrap. This happens because in the bootstrap samples the outlier is appearing in larger proportions than in the original sample.

The difference in the re-sampled distributions is reflected in the corresponding 99% basic percentile confidence intervals for μ_0. With the classical bootstrap we get $(79.537, 101.121)$ and with the robust bootstrap yields $(78.505, 94.869)$. The corresponding 95% confidence intervals are also different, but the disagreement is less severe. The classical bootstrap yields $(80.349, 93.79)$ and the robust bootstrap $(79.717, 91.994)$.

3.2.2 Two-sample location-scale: Seeded clouds

We consider data from an experiment conducted in the state of Florida (USA) between 1968 to 1972. The data we use is a subset corresponding to the period 1968 to

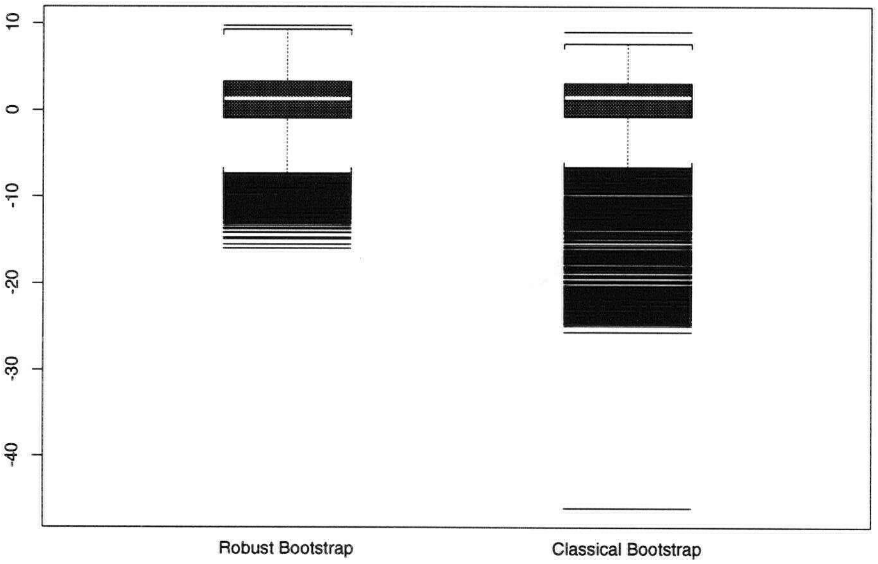


Figure 3.3: Comparison of the classical and robust bootstrap distribution estimates of $\hat{\mu}_n - \mu_0$ for the blood pressure data.

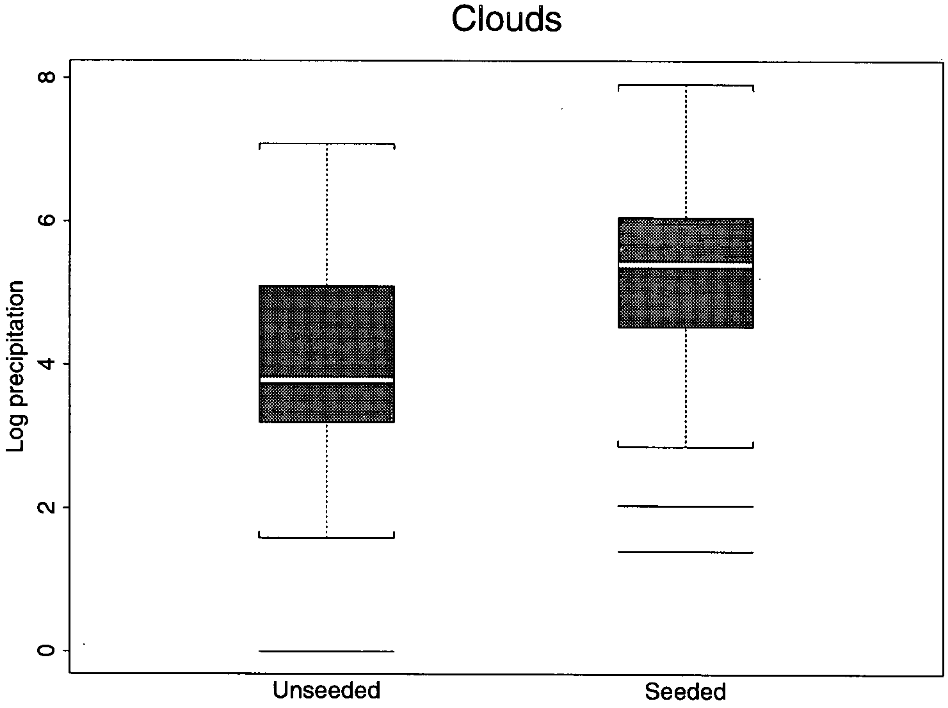


Figure 3.4: Precipitation data

1970. These data were originally analyzed by Simpson *et al.* (1975) using Bayesian techniques. The question of interest is whether "dynamic seeding" (massive silver iodide seeding of clouds) produces, under certain conditions, increased precipitation. The experimental unit is a single cumulus cloud and the outcome is the rain volume falling from the cloud. There are two groups of 26 clouds each. One group contains the "seeded clouds" and the other the "unseeded clouds".

We applied a log transformation to the data to get similar dispersions in both groups. Boxplots of the transformed data are provided in Figure 3.4.

Using a two-sample t-test for populations with equal variance, the p-value for the null hypothesis of equal means is $p = 0.0141$. The corresponding 99% confidence interval for the difference of the means yields $(-0.06, 2.35)$. Thus, at the 1% level, we would conclude that there is not enough evidence to support the alternative hypothesis of different mean rainfall. Figure 3.4 suggests that there is a difference in the location of the two boxplots, but that this shift is probably concealed by the two smallest observations in the seeded group. Indeed, after removing these two clouds the two-sample t-test assuming equal variances for the hypothesis of equal means now yields a p-value of 0.0014 (roughly ten times smaller than above). The corresponding confidence interval with nominal level of 99% becomes $(0.30, 2.56)$. Of course, the actual level of this confidence interval is unknown since its construction involved a subjective rejection rule.

If we bootstrap the two-sample t-test for populations with equal variances using the classical bootstrap, we obtain the following 95% and 99% basic percentile confidence intervals for the difference of the means: $(-4.561, -0.202)$ and $(-5.119, 0.599)$ respectively. These results also yield a p-value p that satisfies $0.01 < p < 0.05$. The inference based on bootstrapping the classical two-sample Welch test for populations with different variances leads to basic percentile confidence intervals that are equal to the ones shown above.

We now apply our robust bootstrap to construct a 99% confidence interval for the difference in the location parameters. We first describe a more general method to compare several location estimates. Given independent samples from k potentially different populations we want to build a confidence interval for $\mathbf{c}'\boldsymbol{\mu}$, where $\mathbf{c}' = (c_1, \ldots, c_k)$ are fixed constants with $c_1 \neq 0$ and $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_k)'$ is the vector of the population parameters.

Assume, for simplicity of the argument, that all the groups have the same number of observations, n. Equation (2.12) for the i-th population becomes

$$\frac{1}{n}\sum_{j=1}^{n} \psi\left(\left(y_{ij} - \hat{\mu}_n^i\right) / \hat{\sigma}_n^{(i)}\right) = 0, \tag{3.8}$$

where $y_{ij}, 1 \leq j \leq n$ is the data for the i-th population, and $\hat{\sigma}_n^{(i)}$ is the corresponding S-scale estimate. We base our inference on the distribution of

$$\sqrt{n}\,\mathbf{c}'\left(\hat{\boldsymbol{\mu}}_n - \boldsymbol{\mu}\right), \tag{3.9}$$

where $\hat{\boldsymbol{\mu}}_n' = \left(\hat{\mu}_n^1, \ldots \hat{\mu}_n^k\right)$ and $\hat{\mu}_n^i$, $1 \leq i \leq k$ is the robust location estimate for the i-th population given by (3.8). The distribution function of (3.9) can be obtained as follows.

$$F(u) = P\left(\sqrt{n}\,\mathbf{c}'\left(\hat{\boldsymbol{\mu}}_n - \boldsymbol{\mu}\right) \leq\ u\right) = \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} F_{v_1}\left(u/c_1 + \tilde{\mathbf{c}}'\mathbf{g}\right)\, dF_{v_2}\left(g_2\right)\cdots dF_{v_k}\left(g_k\right), \tag{3.10}$$

where $v_i = \sqrt{n}\ \left(\hat{\mu}_n^i - \mu_i\right)$, $F_{v_i}(u) = P\left(v_i \leq u\right)$ for $i = 1, \ldots, k$, $\mathbf{g} = (g_2, \ldots, g_k)$ and $\tilde{\mathbf{c}} = (-c_2/c_1, \ldots, -c_k/c_1)$. To simplify the notation we do not explicitly indicate that F and F_{v_i} above depend on the sample size n.

An estimate for (3.10) based on N bootstrap samples for each population ($k\,N$ independent bootstrap samples overall) can be constructed as follows. For $i = 1, \ldots, k$ and $j = 1, \ldots, N$ let $v_j^{*i} = \sqrt{n}\left(\hat{\mu}_{nj}^{R*i} - \mu_i\right)$ be the re-calculated v_j^*'s for the i-th population. Let $\hat{F}_{v_1}$ be the empirical distribution function of the v_j^{*1}, $j = 1, \ldots, N$. The estimate of (3.10) is given by

$$\hat{F}(u) = \frac{1}{N} \sum_{i_2=1}^{N} \cdots \frac{1}{N} \sum_{i_k=1}^{N} \hat{F}_{v_1} \left(u/c_1 + \tilde{c}_1 \, v_{i_2}^{*2} + \cdots \tilde{c}_{k-1} \, v_{i_k}^{*k} \right) ,$$

where $\tilde{c}_l$ denotes the l-th coordinate of the vector $\mathbf{c}$. With this method we can estimate the required quantiles of (3.9) by solving $\hat{F}(u) = \alpha$.

We apply this method with $k = 2$ and $\mathbf{c}' = (1, -1)$ using the same MM-location estimate as in the previous example (see page 95). The 99% confidence interval for $\mathbf{c}'\boldsymbol{\mu} = \mu_1 - \mu_2$ is $(0.22, 2.22)$. We thus conclude that the p-value satisfies $p < 0.01$ and reject the null hypothesis of equal means at the 1% level. Our conclusion is in agreement with the one reached by Simpson *et al.* (1975). Splus functions for one- and two-sample confidence intervals based on this method are available from the author.

3.3 Asymptotic properties

The following theorem shows that the robust bootstrap re-calculated quantities $\sqrt{n}\left(\hat{\mu}_n^{R*} - \hat{\mu}_n\right)$ have the same asymptotic distribution as the sequence $\sqrt{n}\left(\hat{\mu}_n - \mu\right)$ when $n \to \infty$. This result justifies the use of our method in order to perform inference on the parameter μ. Note that this result holds for any distribution F in the contamination neighbourhood $\mathcal{H}_\epsilon$ (see page 36).

Theorem 3.1 - Convergence of the robust bootstrap distribution - *Let* $\psi : \mathbb{R} \to \mathbb{R}$ *be odd, bounded, and non-decreasing in* $(0, \infty]$. *Let* $\rho : \mathbb{R} \to \mathbb{R}_+$ *be even,* *bounded, and non-decreasing in $(0,\infty]$. Let $b \in (0,1/2]$. Let $\hat{\mu}_n$ be the MM-location estimate based on ψ, let $\hat{\sigma}_n$ be the S-scale calculated with the function ρ and let $\tilde{\mu}_n$ be the associated S-location estimate (see Definitions 2.7 and 2.9). Let μ, σ and $\tilde{\mu}$ be the almost sure limits of the sequences $\hat{\mu}_n$, $\hat{\sigma}_n$ and $\tilde{\mu}_n$ respectively. Assume that the following conditions hold.*

1. *The following expected values exist and are finite for all $F \in \mathcal{H}_\epsilon$:*

$$E_F\left[\psi\left(X\right)\right], \quad E_F\left[\psi\left(X\right)/\,X\right], \quad E_F\left[\psi'\left(X\right)\right] \quad \text{and} \quad E_F\left[\psi'\left(X\right)X\right].$$

2. *For all $F \in \mathcal{H}_\epsilon$, $E_F\left[\rho'\left(X\right)X\right] \neq 0$ and finite.*

3. *The following functions are continuous: ψ, ρ', $\psi\left(u\right)/\,u$, $\rho'\left(u\right)/\,u$, ψ', ρ'', ψ'', ρ''', $\left(\psi\left(u\right)-\psi'\left(u\right)u\right)/\,u^2$, and $\left(\rho'\left(u\right)-\rho''\left(u\right)u\right)/\,u^2$.*

Let V^2 be the asymptotic variance of the sequence $\sqrt{n}\left(\hat{\mu}_n-\mu\right)$ (see Theorem 2.6), and let $\hat{\mu}_n^{R*}-\hat{\mu}_n$ be the robust bootstrap re-calculated estimates. Then along almost all sample sequences, conditional on the first n observations,

$$\sqrt{n}\left(\hat{\mu}_n^{R*}-\hat{\mu}_n\right) \xrightarrow[n\to\infty]{\mathcal{W}} \quad N\left(0,V^2\right).$$

Remark 3.3 Note that if ψ_c is a Huber-type function then

$$\left(\psi_c\left(u\right)-\psi_c'\left(u\right)\,u\right)/\,u^2=0 \qquad \text{for} \quad |u|\leq c\,.$$

When ρ_d is a re-descending function in Tukey's family we have

$$\left(\rho_d'\left(u\right)-\rho_d''\left(u\right)\,u\right)/\,u^2=\frac{4}{d^2}\left(\frac{u}{d}-\left(\frac{u}{d}\right)^3\right) \qquad \text{for} \quad |u|\leq d\,.$$

Hence assumption 3 is also satisfied if the distribution F generating the data does not have positive mass at the points where ψ is not differentiable. Apply Lemma 7.13 to verify that conditions 1 and 2 also hold for ψ_c and ρ_d in Huber's and Tukey's families respectively.

Remark 3.4 We can change condition 3 above so that it does not depend on the particular $F \in \mathcal{H}_\epsilon$ that generated the data. In this case we require that 3 hold everywhere. Clearly, now Huber's functions ψ_c do not satisfy the assumption. We have to use a "smoothed" version $\tilde{\psi}_c$ that coincides with ψ_c except on an arbitrary small neighbourhood around c and $-c$ where $\tilde{\psi}_c$ is continuously differentiable.

Proof of Theorem 3.1: For $i = 1, \ldots, n$ let $r_i = x_i - \hat{\mu}_n$ and $\tilde{r}_i = x_i - \tilde{\mu}_n$. Note that the estimates $\hat{\mu}_n$, $\hat{\sigma}_n$ and $\tilde{\mu}_n$ satisfy

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n} \psi\left(r_i / \hat{\sigma}_n\right) &= 0 \\ \frac{1}{n}\sum_{i=1}^{n} \rho\left(\tilde{r}_i / \hat{\sigma}_n\right) &= b \\ \frac{1}{n}\sum_{i=1}^{n} \rho'\left(\tilde{r}_i / \hat{\sigma}_n\right) &= 0\,. \end{aligned} \tag{3.11}$$

They can also be written as a weighted average of the observations as follows:

$$\begin{aligned} \hat{\mu}_n &= \sum_{i=1}^{n} \omega_i\, x_i \Bigg/ \sum_{i=1}^{n} \omega_i, \\ \hat{\sigma}_n &= \sum_{i=1}^{n} v_i\ \left(x_i - \tilde{\mu}_n\right), \\ \tilde{\mu}_n &= \sum_{i=1}^{n} \omega_i'\, x_i \Bigg/ \sum_{i=1}^{n} \omega_i', \end{aligned}$$

where

$$\omega_i\left(\hat{\mu}_n, \hat{\sigma}_n\right) = \psi\left(\left(x_i - \hat{\mu}_n\right) / \hat{\sigma}_n\right) / \left(x_i - \hat{\mu}_n\right),$$

$$v_i\left(\tilde{\mu}_n, \hat{\sigma}_n\right) = \rho\left(\left(x_i - \tilde{\mu}_n\right) / \hat{\sigma}_n\right) / \left(n\, b\, \left(x_i - \tilde{\mu}_n\right)\right),$$

and

$$\omega_i'\left(\tilde{\mu}_n, \hat{\sigma}_n\right) = \rho'\left(\left(x_i - \tilde{\mu}_n\right) / \hat{\sigma}_n\right) / \left(x_i - \tilde{\mu}_n\right).$$

The idea is to show that the vector $\left(\hat{\mu}_n, \hat{\sigma}_n, \tilde{\mu}_n\right) \in \mathbb{R}^3$ is the fixed point of a smooth function of means. Let $\mathbf{f} : \mathbb{R}^3 \rightarrow \mathbb{R}^3$ be defined by

$$\mathbf{f}\left(k, s, \tilde{k}\right) = \begin{pmatrix} d_n\left(k, s\right) / \, g_n\left(k, s\right) \\ s\, h_n\left(\tilde{k}, s\right) \\ u_n\left(\tilde{k}, s\right) \Big/ \, v_n\left(\tilde{k}, s\right) \end{pmatrix},$$

where

$$d_n\left(k, s\right) = \sum_{i=1}^{n} \omega_i\left(k, s\right) x_i\,,$$

$$g_n\left(k, s\right) = \sum_{i=1}^{n} \omega_i\left(k, s\right)\,,$$

$$h_n\left(\tilde{k}, s\right) = \sum_{i=1}^{n} v_i\left(\tilde{k}, s\right)\, \tilde{r}_i\,,$$

$$u_n\left(\tilde{k}, s\right) = \sum_{i=1}^{n} \omega_i'\left(\tilde{k}, s\right) x_i\,,$$

and

$$v_n\left(\tilde{k}, s\right) = \sum_{i=1}^{n} \omega_i'\left(\tilde{k}, s\right)\,.$$

The re-weighted representation of the estimates can be written as

$$\mathbf{f}\left(\hat{\mu}_n, \hat{\sigma}_n, \tilde{\mu}_n\right) = \left(\hat{\mu}_n, \hat{\sigma}_n, \tilde{\mu}_n\right)'.$$

Now, perform a Taylor expansion of $\mathbf{f}$ around the limiting values $(\mu, \sigma, \tilde{\mu})'$. We have

$$\begin{pmatrix} \hat{\mu}_n \\ \hat{\sigma}_n \\ \tilde{\mu}_n \end{pmatrix} = \mathbf{f}\left(\mu, \sigma, \tilde{\mu}\right) + \nabla \mathbf{f}\left(\mu, \sigma, \tilde{\mu}\right) \begin{pmatrix} \hat{\mu}_n - \mu \\ \hat{\sigma}_n - \sigma \\ \tilde{\mu}_n - \tilde{\mu} \end{pmatrix} + R_n, \tag{3.12}$$

where $R_n \in \mathbb{R}^3$ is the remainder term and $\nabla \mathbf{f}\left(\cdot\right) \in \mathbb{R}^{3\times 3}$ is the matrix of partial derivatives of $\mathbf{f}$. Each component of R_n is of the form $\mathbf{x}_n' H_n \mathbf{x}_n$, where $\|\mathbf{x}_n\| = O_P\left(1/\sqrt{n}\right)$ and H_n is a Hessian matrix. We have $\left|\mathbf{x}_n' H_n \mathbf{x}_n\right| \leq \|H_n \mathbf{x}_n\| \, \|\mathbf{x}_n\| \leq \|H_n\| \, \|\mathbf{x}_n\|^2$. Each entry in H_n is a second derivative of a component of $\mathbf{f}$. For example, if $\mathbf{f}_1$ denotes the first coordinate of $\mathbf{f}$, we have

$$H_{n\,(1,1)} = \frac{\partial^2 \mathbf{f}_1}{\partial^2 m} = \frac{\left[\frac{1}{s}\sum_{i=1}^n \psi''\left(r_i\right) \frac{1}{s}\right] \left[\sum_{i=1}^n \psi\left(r_i\right)/\, r_i\right] + \left[\frac{1}{s}\sum_{i=1}^n \psi'\left(r_i\right)\right]}{\left(\sum_{i=1}^n \psi\left(r_i\right)/\, r_i\right)^2}$$
$$\times \frac{\frac{1}{s}\left[\sum_{i=1}^n \psi\left(r_i\right)\right] \left[\sum_{i=1}^n \frac{\psi'(r_i)}{r_i}\right] - \left[\sum_{i=1}^n \psi\left(r_i\right)\right] \left[\sum_{i=1}^n \frac{\psi(r_i)}{r_i^2}\right]}{\left(\sum_{i=1}^n \psi\left(r_i\right)/\, r_i\right)^2}$$

By Lemma 7.1 and assumption 3 we have $\|H_n\| \, \|\mathbf{x}_n\| = o_P\left(1\right)$. Then, $\left|\mathbf{x}_n' H_n \mathbf{x}_n\right| = o_P\left(1/\sqrt{n}\right)$. Hence, $\|R_n\| = o_P\left(1/\sqrt{n}\right)$.

To simplify the notation let $\hat{\boldsymbol{\theta}}_n = \left(\hat{\mu}_n, \hat{\sigma}_n, \tilde{\mu}_n\right)'$, and $\boldsymbol{\theta} = \left(\mu, \sigma, \tilde{\mu}\right)'$. Equation (3.12) becomes

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}\right) = \left[\mathbf{I} - \nabla \mathbf{f}\left(\boldsymbol{\theta}\right)\right]^{-1} \sqrt{n}\left(\mathbf{f}\left(\boldsymbol{\theta}\right) - \boldsymbol{\theta}\right) + o_P\left(1\right). \tag{3.13}$$

The rest of the proof consists of showing that the bootstrap distribution of the right-hand side of (3.13) converges to the asymptotic distribution of $\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}\right)$. Note that the correction matrix in (3.13) has to be estimated. We will first show that a consistent estimate of $[\mathbf{I} - \nabla\mathbf{f}(\boldsymbol{\theta})]^{-1}$, namely $[\mathbf{I} - \nabla\mathbf{f}(\hat{\boldsymbol{\theta}}_n)]^{-1}$, yields the coefficients a_n, b_n and c_n in (3.7). First note that

$$\mathbf{I} - \nabla\mathbf{f}(\hat{\boldsymbol{\theta}}_n) = \begin{pmatrix} a_{11} & a_{12} & 0 \\ 0 & a_{22} & 0 \\ 0 & a_{32} & a_{33} \end{pmatrix},$$

The only entry that needs justification is $a_{23} = 0$. Remember that $\hat{\sigma}_n$ minimizes $s_n(t)$, $t \in \mathbb{R}$. It follows that for each $t \in \mathbb{R}$ we have

$$\frac{\partial}{\partial t}\left[\frac{1}{n}\sum_{i=1}^{n}\rho\left((x_i - t)/\, s_n(t)\right)\right] = 0\,,$$

hence

$$\frac{1}{n}\sum_{i=1}^{n}\rho'\left(\frac{x_i - t}{s_n(t)}\right)\left(\frac{-s_n(t) - (x_i - t)\,\dot{s_n}(t)}{s_n(t)^2}\right) = 0, \qquad (3.14)$$

where $\dot{s_n}(t)$ denotes the derivative of $s_n(t)$. Because $\tilde{\mu}_n$ minimizes $s_n(t)$, we have $\dot{s_n}(\tilde{\mu}_n) = 0$, which together with (3.14) implies

$$\frac{1}{n}\sum_{i=1}^{n}\rho'\left((x_i - \tilde{\mu}_n)/\,\hat{\sigma}_n\right) = 0.$$

It is easy to verify that

$$a_{23} = \frac{\partial}{\partial \tilde{k}}\left[\frac{1}{n}\sum_{i=1}^{n}\rho\left(\left(x_i - \tilde{k}\right)/\, s\right)\right]\Bigg|_{\left(k=\hat{\mu}_n, s=\hat{\sigma}_n, \tilde{k}=\tilde{\mu}_n\right)} = -\frac{1}{s}\frac{1}{n}\sum_{i=1}^{n}\rho'\left((x_i - \tilde{\mu}_n)/\,\hat{\sigma}_n\right)\,,$$

so that $a_{23} = 0$. Hence,

$$[\mathbf{I} - \nabla\mathbf{f}(\hat{\boldsymbol{\theta}}_n)]^{-1} = \begin{pmatrix} 1/\,a_{11} & -a_{12}/\,a_{11}a_{22} & 0 \\ 0 & 1/\,a_{22} & 0 \\ 0 & -a_{32}/\,a_{22}a_{33} & 1/\,a_{33} \end{pmatrix}.$$

Because we are interested in the asymptotic behaviour of the first coordinate of $\hat{\boldsymbol{\theta}}_n$ we only need the first row of $[\mathbf{I} - \nabla \mathbf{f}(\hat{\boldsymbol{\theta}}_n)]^{-1}$, and hence we only need to calculate a_{11}, a_{22} and a_{12}.

Simple calculations yield

$$\begin{aligned}
a_{11} &= \frac{\sum_{i=1}^n \psi'\left(\frac{x_i - \hat{\mu}_n}{\hat{\sigma}_n}\right)}{\sum_{i=1}^n \psi\left(\frac{x_i - \hat{\mu}_n}{\hat{\sigma}_n}\right) \Big/ \left(\frac{x_i - \hat{\mu}_n}{\hat{\sigma}_n}\right)} \\
a_{12} &= \frac{\sum_{i=1}^n \psi'\left(\frac{x_i - \hat{\mu}_n}{\hat{\sigma}_n}\right)\left(\frac{x_i - \hat{\mu}_n}{\hat{\sigma}_n}\right)}{\sum_{i=1}^n \psi\left(\frac{x_i - \hat{\mu}_n}{\hat{\sigma}_n}\right) \Big/ \left(\frac{x_i - \hat{\mu}_n}{\hat{\sigma}_n}\right)} \\
a_{22} &= \frac{1}{b} \sum_{i=1}^n \rho'\left(\frac{x_i - \tilde{\mu}_n}{\hat{\sigma}_n}\right)\left(\frac{x_i - \tilde{\mu}_n}{\hat{\sigma}_n}\right)
\end{aligned}$$

It is easy to verify that $(1/a_{11}, -a_{12}/a_{11}a_{22}) = (a_n, b_n)$ where a_n and b_n are the correction factors in (3.7).

We now show that the bootstrap distribution of the right-hand side of (3.13) converges to the same distribution as the sequence $\hat{\boldsymbol{\theta}}_n$. First we show that $\mathbf{f}(\boldsymbol{\theta}) - \boldsymbol{\theta}$ in (3.13) is a smooth function of means. It will then follow that we can bootstrap it to obtain an estimate of its distribution. Define the random vector $\mathbf{Y}(\boldsymbol{\theta}) \in \mathbb{R}^5$ by

$$\begin{aligned}
\mathbf{Y}(\boldsymbol{\theta}) = (\psi((X-\mu)/\sigma)\ ,\ \psi((X-\mu)/\sigma)/(X-\mu)\ , \\
\rho((X-\tilde{\mu})/\sigma)\ ,\ \rho'((X-\tilde{\mu})/\sigma)\ ,\ \rho'((X-\mu)/\sigma)/(X-\mu))'
\end{aligned}$$

We have

$$\boldsymbol{\mu}_{\mathbf{Y}(\boldsymbol{\theta})} = E_F \mathbf{Y}(\boldsymbol{\theta}) = (0, \#, b, 0, \#),$$

where # stands for something different from zero, but otherwise irrelevant in the analysis that follows. Let $\mathbf{g} : \mathbb{R}^5 \to \mathbb{R}^3$

$$\mathbf{g}\left(x_1, x_2, x_3, x_4, x_5\right) = \left(\frac{x_1}{x_2} + \mu \ , \ \frac{\sigma}{b} x_3 \ , \ \frac{x_4}{x_5} + \tilde{\mu}\right).$$

Then $\mathbf{g}\left(\boldsymbol{\mu}_{\mathbf{Y}(\boldsymbol{\theta})}\right) = (\mu, \sigma, \tilde{\mu}) = \boldsymbol{\theta}$. Let $\mathbf{Y}_i(\boldsymbol{\theta})$ be the corresponding vectors obtained with the observations x_i, $i = 1, \ldots, n$, and let $\bar{\mathbf{Y}}_n(\boldsymbol{\theta})$ be their sample mean. We have

$$\mathbf{g}\left(\bar{\mathbf{Y}}_n(\boldsymbol{\theta})\right) = \mathbf{f}(\boldsymbol{\theta}).$$

Hence

$$\sqrt{n}\left[\mathbf{f}(\boldsymbol{\theta}) - \boldsymbol{\theta}\right] = \sqrt{n}\left[\mathbf{g}\left(\bar{\mathbf{Y}}_n(\boldsymbol{\theta})\right) - \mathbf{g}\left(\boldsymbol{\mu}_{\mathbf{Y}(\boldsymbol{\theta})}\right)\right].$$

By Bickel and Freedman (1981), if $\mathbf{g}$ is smooth, we can bootstrap the last expression to obtain a consistent estimate of its asymptotic distribution.

For any vector $\boldsymbol{\theta}$ let $\bar{\mathbf{Y}}_n^*(\boldsymbol{\theta})$ be the sample mean of the vectors $\bar{\mathbf{Y}}_n(\boldsymbol{\theta})$ obtained with a bootstrap sample $x_1^*, \ldots, x_n^*$. Because $\boldsymbol{\theta}$ is unknown and we want to estimate it with $\hat{\boldsymbol{\theta}}_n$, we still have to show that $\sqrt{n}\left[\bar{\mathbf{Y}}_n^*(\hat{\boldsymbol{\theta}}_n) - \bar{\mathbf{Y}}_n(\hat{\boldsymbol{\theta}}_n)\right]$ is asymptotically equivalent to $\sqrt{n}\left[\bar{\mathbf{Y}}_n(\boldsymbol{\theta}) - \boldsymbol{\mu}_{\mathbf{Y}(\boldsymbol{\theta})}\right]$.

Consider the metric $d_2(F_1, F_2)$ for distribution functions defined by

$$d_2^2(F_1, F_2) = \inf E\left[(X - Y)^2\right] \tag{3.15}$$

where the infimum is taken over all the possible joint distributions for the random vector (X, Y) such that its marginal laws are F_1 and F_2 respectively. This metric was introduced in Mallows (1972) and Tanaka (1973). For a detailed discussion see Bickel and Freedman (Section 8, 1981). d_2 metrizes weak convergence in the following sense:

$$d_2(F_\alpha, F) \to 0 \quad \text{iff} \quad F_\alpha \xrightarrow{\mathcal{W}} F \ \text{ and } \ \lim_\alpha E_{F_\alpha} X^2 = E_F X^2$$

where $\xrightarrow{\mathcal{W}}$ denotes weak convergence. Let $Z_1, \ldots, Z_n$ be i.i.d. random variables with common distribution function F. Let $F^{(n)}$ denote the distribution function of

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(Z_i - E\left[Z_i\right]\right) .$$

For any pair of distribution functions F_1 and F_2, d_2 satisfies $d_2\left(F_1^{(n)}, F_2^{(n)}\right) \leq d_2\left(F_1, F_2\right)$ (Bickel and Freedman, 1981).

Let G_n be the empirical distribution function of the vectors $\mathbf{y}_i\left(\hat{\boldsymbol{\theta}}_n\right)$, $i = 1, \ldots, n$. Conditional on the first n observations, the distribution of

$$\sqrt{n}\left[\bar{\mathbf{y}}_n^*\left(\hat{\boldsymbol{\theta}}_n\right) - \bar{\mathbf{y}}_n\left(\hat{\boldsymbol{\theta}}_n\right)\right]$$

is $G_n^{(n)}$. Let Z represent the asymptotic distribution of $\sqrt{n}\left[\bar{\mathbf{y}}_n\left(\boldsymbol{\theta}\right) - \boldsymbol{\mu}\left(\boldsymbol{\theta}\right)\right]$. In what follows we will show that

$$d_2\left(G_n^{(n)}, Z\right) \xrightarrow[n\to\infty]{} 0 \text{ almost surely.}$$

Let F_n denote the distribution function of $\sqrt{n}\left[\bar{\mathbf{y}}_n\left(\boldsymbol{\theta}\right) - \boldsymbol{\mu}\left(\boldsymbol{\theta}\right)\right]$ and $F_n^{(n)}$ the empirical distribution of $\sqrt{n}\left[\bar{\mathbf{y}}_n^*\left(\boldsymbol{\theta}\right) - \bar{\mathbf{y}}_n\left(\boldsymbol{\theta}\right)\right]$. Following the notation in Bickel and Freedman (1981) we have that

$$d_2\left(G_n^{(n)}, Z\right) \leq d_2\left(G_n^{(n)}, F_n^{(n)}\right) + d_2\left(F_n^{(n)}, Z\right) \leq d_2\left(G_n, F_n\right) + d_2\left(F_n^{(n)}, Z\right) .$$

Bickel and Freedman show in their Theorem 2.1 that $d_2\left(F_n^{(n)}, Z\right) \to 0$ almost surely as $n \to \infty$. Lemma 7.18 shows that $d_2\left(G_n, F_n\right) \to 0$ almost surely. We have shown that conditionally on the first n observations, as n goes to infinity,

$$\sqrt{n}\left[\bar{\mathbf{y}}_n^*\left(\hat{\boldsymbol{\theta}}_n\right) - \bar{\mathbf{y}}_n\left(\hat{\boldsymbol{\theta}}_n\right)\right]$$

converges weakly to the same limit as $\sqrt{n}\left[\bar{\mathbf{y}}_n\left(\boldsymbol{\theta}_0\right)-\boldsymbol{\mu}\left(\boldsymbol{\theta}_0\right)\right]$. To complete the proof note that the function $\mathbf{g}$ satisfies the regularity conditions required in Lemma 8.1 of Bickel and Freedman (1981), hence the above conclusion applies to $\mathbf{g}\left(\bar{\mathbf{y}}_n^*\left(\hat{\boldsymbol{\theta}}_n\right)\right)$. ■

Remark 3.5 - If assumption 3 in the previous theorem does not hold, then the remainder term R_n in (3.12) does not necessarily satisfy $\|R_n\| = o_P(1/\sqrt{n})$. The theorem is still valid, but the proof follows a different approach. Let $\mathbf{f}^*(\boldsymbol{\theta})$ be the evaluation of $\mathbf{f}(\boldsymbol{\theta})$ with a bootstrap sample of the $x_1, \ldots, x_n$. We can show that $\sqrt{n}\left[\mathbf{f}^*\left(\hat{\boldsymbol{\theta}}_n\right)-\hat{\boldsymbol{\theta}}_n\right]$ in (3.13) is asymptotically normal. Let $\Sigma_{\mathbf{f}}$ be the asymptotic covariance matrix of the robust bootstrapped estimates. Let

$$A_n = \left[\mathbf{I}-\nabla\mathbf{f}\left(\hat{\boldsymbol{\theta}}_n\right)\right]^{-1} \Sigma_{\mathbf{f}} \left[\mathbf{I}-\nabla\mathbf{f}\left(\hat{\boldsymbol{\theta}}_n\right)\right]^{\prime -1},$$

then it is easy to show that the element $[A_n]_{(1,1)}$ converges to the asymptotic variance of $\sqrt{n}\left(\hat{\mu}_n-\mu\right)$.

The following theorem shows that the bootstrap variance of the robust bootstrap estimates converges to the asymptotic variance. Note that this is not a consequence of the previous theorem. See Ghosh *et al.* (1984) for a counterexample.

Theorem 3.2 - Convergence of the robust bootstrap variances - *Assume the same regularity conditions as in Theorem 3.1. Then along almost all sample sequences,*

$$n \ \operatorname{var}^*\left(\hat{\mu}_n^{R*}\right) \rightarrow \sigma^2$$

where σ^2 is the asymptotic variance of the sequence $\sqrt{n}\left(\hat{\mu}_n-\mu\right)$, and var* *denotes the bootstrap variance.*

Proof: To fix ideas we first consider the simple case of known scale. In this case (3.6) and (3.7) become (see also (1.6))

$$\sqrt{n}\,(\hat{\mu}_n^* - \hat{\mu}_n) = a_n \frac{\sum_{i=1}^n \psi\,(x_i^* - \hat{\mu}_n)}{\sum_{i=1}^n \psi\,(x_i^* - \hat{\mu}_n)/\,(x_i^* - \hat{\mu}_n)},$$

where

$$a_n = \frac{\sum_{i=1}^n \psi\,(x_i - \hat{\mu}_n)/\,(x_i - \hat{\mu}_n)}{\sum_{i=1}^n \psi'\,(x_i - \hat{\mu}_n)}. \tag{3.16}$$

Then we can write

$$\sqrt{n}\,(\hat{\mu}_n^* - \hat{\mu}_n) = a_n \frac{\bar{Y}_n^*}{\bar{Z}_n^*} = a_n\, g\left(\bar{Y}_n^*, \bar{Z}_n^*\right),$$

where $g\,(x,y) = x/\,y$, $y_i^* = \psi\,(x_i^* - \hat{\mu}_n)$ and $z_i^* = \psi\,(x_i^* - \hat{\mu}_n)/\,(x_i^* - \hat{\mu}_n)$. By Bickel and Doksum (1977, page 52), conditionally on the first n observations,

$$n \text{ var}_*\,(\hat{\mu}_n^*) = a_n^2\, \frac{\frac{1}{n}\sum_{i=1}^n y_i^2}{\left(\frac{1}{n}\sum_{i=1}^n z_i\right)^2} + o\,(1/\,n)\,.$$

The result now follows by replacing a_n with its value in (3.16) and taking the limit as $n \to \infty$. The general proof follows the same lines, and it is based on the matrix representation used in the proof of Theorem 3.1 ■

3.4 Robustness properties

In this section we study the robustness properties of the quantile estimates of the robust bootstrap. Let $t \in (0,1)$, and let q_t be the t-th upper quantile of a statistic $\hat{\mu}_n$, that is, q_t satisfies

$$P\,[\,\hat{\mu}_n > q_t\,] = t\,.$$

Following Singh (1998) we define the upper breakdown point of a quantile estimate $\hat{q}_t$ as the minimum proportion of asymmetric contamination that can drive it over any finite bound. Equivalently, it is the smallest proportion of arbitrarily large outliers in the original data set such that we expect the re-calculated estimate to be unbounded in at least $t \times 100$ % of the bootstrap samples.

It is easy to see that if the breakdown point of $\hat{\mu}_n$ is ϵ^*, the corresponding upper breakdown point of $\hat{q}_t$ is the smallest $\delta \in [0,1]$ such that

$$P\left(\text{ Binomial}\,(n,\delta) \geq [\epsilon^* n] \right) \geq t, \tag{3.17}$$

where $[x]$ denotes the smallest integer larger or equal to x. Lemma 7.17 shows that the function $f\,(\delta) = P\left(\text{ Binomial}\,(n,\delta) \geq [\epsilon^* n] \right)$ is continuous and non-decreasing for $\delta \in [0,1]$, and hence we can always find the upper breakdown point of $\hat{q}_t$ as defined above.

In the same paper Singh proves that if we re-sample from the Winsorized data points, the resulting quantile estimates are asymptotically equivalent to those obtained by the classical bootstrap but have the highest possible breakdown point, namely, the minimum between 50% and the breakdown point of the robust estimate.

There are two closely related scenarios in which the quantile estimates based on the robust bootstrap can break down. The first unfavourable situation is when the proportion of outliers in the original data is larger than the breakdown point of the estimate. In this case the estimate is already unreliable, and so are the inferences we derive from it. The second case is related to the number of outliers appearing in the bootstrap samples. Let τ^* be the expected proportion of bootstrap samples that contain more outliers than the breakdown point of the estimate. In other words, we expect $\tau^* \times 100\%$ of the re-calculated $\hat{\mu}_n^*$'s to be unreliable. The estimate $\hat{q}_t$ may be severely affected by the outliers when $\tau^* > t$. The following theorem summarizes this discussion.

Theorem 3.3 - Breakdown point of the robust bootstrap quantiles for the location-scale model - *Let $x_1, \dots x_n$ be i.i.d. observations following model (2.1). Let $0 < t < 1/2$ and let $\hat{\mu}_n$ be an MM-location estimate with breakdown point ϵ^*. The breakdown point of the t-th robust bootstrap quantile estimate $\hat{q}_t$ is* $\min\left(t^{1/n}, \epsilon^*\right)$.

Proof: Noting that $\hat{\sigma}_n^*$ in (3.5) satisfies

$$\hat{\sigma}_n^* = \hat{\sigma}_n \, \frac{1}{n\,b} \sum_{i=1}^{n} \rho\left(\frac{y_i^* - \tilde{\mu}_n}{\hat{\sigma}_n}\right)$$

we see that $\hat{\sigma}_n^*$ remains bounded for any bootstrap sample $y_1^*, \dots y_n^*$.

Let $x_1^*, \dots, x_n^*$ be a bootstrap sample. To simplify the notation, and without loss of generality, assume that $x_1^*, \dots, x_{g^*}^*$, with $0 \leq g \leq n$ are points in the bootstrap sample that are not outliers. The robust bootstrap evaluation of $\hat{\mu}_n$ satisfies

$$\hat{\mu}_n^{R*} - \hat{\mu}_n = \frac{\sum_1^{g^*} \left[\psi\left(\frac{x_i^* - \hat{\mu}_n}{\hat{\sigma}_n}\right) \Big/ (x_i^* - \hat{\mu}_n)\right] x_i^* + \sum_{g^*+1}^{n} \left[\psi\left(\frac{x_i^* - \hat{\mu}_n}{\hat{\sigma}_n}\right) \Big/ (x_i^* - \hat{\mu}_n)\right] x_i^*}{\sum_1^{g^*} \psi\left(\frac{x_i^* - \hat{\mu}_n}{\hat{\sigma}_n}\right) \Big/ (x_i^* - \hat{\mu}_n) + \sum_{g^*+1}^{n} \psi\left(\frac{x_i^* - \hat{\mu}_n}{\hat{\sigma}_n}\right) \Big/ (x_i^* - \hat{\mu}_n)},$$

If we now take the limit when the outliers $x_{g^*+1}^*, \dots, x_n^*$ approach infinity, we have

$$\frac{\psi\left(\frac{x_i^* - \hat{\mu}_n}{\hat{\sigma}_n}\right)}{(x_i^* - \hat{\mu}_n)} \xrightarrow[x_i^* \to \infty]{} 0\,, \qquad g^* + 1 \leq i \leq n\,,$$

and

$$\frac{\psi\left(\frac{x_i^* - \hat{\mu}_n}{\hat{\sigma}_n}\right)}{(x_i^* - \hat{\mu}_n)} x_i^* \xrightarrow[x_i^* \to \infty]{} 1\,, \qquad g^* + 1 \leq i \leq n.$$

As a result $\hat{\mu}_n^{R*} - \hat{\mu}_n$ remains bounded as long as $g^* \geq 1$. It is easy to see that the probability of obtaining a bootstrap sample with $g^* = 0$, that is, where all the points are outliers, is ϵ^n, where ϵ is the proportion of outliers in the original sample. Hence, to drive the t-th quantile out of bounds we should have $\epsilon^n \geq t$, that is $\epsilon \geq t^{1/n}$. The proof is complete. ∎

The previous theorem shows that the breakdown point of the t-th robust bootstrap quantile increases with the sample size and with the value of t. For example, if the MM-estimate has a breakdown point of 50%, for any sample size $n \geq 10$ and any $t \geq 0.001$ the breakdown point of the t-th robust bootstrap quantile is 50%.

Table 3.1 shows some classical and robust bootstrap quantile breakdown points for an M-location estimate with breakdown point of 50%. The breakdown points of the robust bootstrap quantile estimates are calculated with the formula proved in Theorem 3.3 above. The corresponding breakdown points for the classical bootstrap quantile estimates are calculated with formula (3.17). For example, the entry 0.22 for $n = 10$ and $t = 0.01$ means that if there are at least 22% outliers (more than 2 outliers) in a sample of size 10, then the classical bootstrap estimate of $q_{0.01}$ might be severely affected by the value of those outliers. Note that the breakdown point decreases as we move further out into the tails of the distribution of $\hat{\mu}_n$, and it increases with the sample size. As an example, for the same quantile $q_{0.01}$ as before, if the sample size is $n = 20$ the classical bootstrap quantile will breakdown only when the proportion of outliers reaches 28%. It is intuitively clear that the lower breakdown points will

	Classical bootstrap				Robust bootstrap			
n	$q_{0.05}$	$q_{0.01}$	$q_{0.005}$	$q_{0.001}$	$q_{0.05}$	$q_{0.01}$	$q_{0.005}$	$q_{0.001}$
5	0.19	0.11	0.08	0.05	0.50	0.40	0.35	0.25
10	0.30	0.22	0.19	0.14	0.50	0.50	0.50	0.50
20	0.35	0.28	0.26	0.21	0.50	0.50	0.50	0.50
50	0.40	0.35	0.33	0.30	0.50	0.50	0.50	0.50

Table 3.1: Comparison of breakdown points of classical and robust bootstrap quantile estimates for MM-location estimators

be found further out into the tails of the distribution, as these quantiles are typically more difficult to estimate. Note that for $n \geq 10$ the breakdown of the robust bootstrap quantile estimates considered here have the highest attainable value, namely 50%. In this sense our method compares favourably with Singh's Winsorized bootstrap, which yields quantile estimates $\hat{q}_t$ with breakdown point 50% for any n and t.

3.5 Studentizing the robust bootstrap

The basic idea behind the studentized bootstrap (both classical and robust) is as follows. Let T_n be a statistic such that $\sqrt{n}\,(T_n - \mu) \to N\,(0, U^2)$ and let $\hat{U}_n^2$ be a consistent estimate of U^2. Under certain regularity conditions (see, for example, Hall, 1992) we have

$$P\left(\sqrt{n}\,(T_n - \mu)/\,\hat{U}_n \leq x\right) - P\left(\sqrt{n}\,(T_n^* - \hat{\mu}_n)/\,\hat{U}_n^* \leq x \middle| \mathcal{X}\right) = O_p\left(n^{-1}\right)\,, \quad (3.18)$$

where $\hat{U}_n^*$ denotes the re-calculated $\hat{U}_n$ with the bootstrap samples, and $P\,(\cdot\,|\,\mathcal{X})$ denotes the bootstrap distribution conditional on the sample $\mathcal{X}$. Note that $\sqrt{n}\,(T_n - \mu)/\,\hat{U}_n$

is an asymptotically pivotal statistic. Under certain regularity conditions, the property of the bootstrap distribution stated in (3.18) holds when the statistic being bootstrapped is asymptotically pivotal (see Hall, 1992).

On the other hand, if $\Phi\left(\cdot\right)$ denotes the standard normal cumulative distribution function we have

$$P\left(\sqrt{n}\,(T_n - \mu)/\,\hat{U}_n \leq x\right) - \Phi\left(x\right) = O\left(n^{-1/2}\right) \, . \tag{3.19}$$

When (3.18) holds we say that the bootstrap estimate of the distribution function of $\hat{\mu}_n$ has a higher order of convergence than the one given by the normal asymptotic distribution.

Because the robust bootstrap re-computes a non-pivotal statistic the resulting estimate of the distribution function of $\hat{\mu}_n$ may not be more accurate than the one derived from its asymptotic normal distribution. To improve the accuracy of our distribution estimates we consider a studentized version of the robust bootstrap, in the same spirit as the bootstrap-t confidence intervals (see Efron, 1979; Hall, 1992; DiCiccio and Efron, 1996). We will refer to them as robust bootstrap-t confidence intervals.

Let $\hat{\mu}_n$, $\hat{\sigma}_n$ and $\tilde{\mu}_n$ be the MM-location, S-scale and S-location estimates, respectively. Under certain regularity conditions the vector $(\hat{\mu}_n, \hat{\sigma}_n, \tilde{\mu}_n)'$ has an asymptotic normal distribution. Let $\Sigma = \Sigma\left(\mu, \sigma, \tilde{\mu}, F\right)$ denote the corresponding asymptotic covariance matrix in $\mathbb{R}^{3\times 3}$. Let Σ_n be the empirical estimate of this matrix, that is $\Sigma_n = \Sigma\left(\hat{\mu}_n, \hat{\sigma}_n, \tilde{\mu}_n, F_n\right)$ where F_n is the empirical distribution function of the sample. Following the same approach as in the proof of Theorem 3.1 it is easy but tedious to see that along almost all sample sequences we have

$$\sqrt{n}\left[\begin{pmatrix} \hat{\mu}_n^* \\ \hat{\sigma}_n^* \\ \tilde{\mu}_n^* \end{pmatrix} - \begin{pmatrix} \hat{\mu}_n \\ \hat{\sigma}_n \\ \tilde{\mu}_n \end{pmatrix}\right] \to \mathbf{N}\left(\mathbf{0}, \tilde{\Sigma}\right),$$

for a certain matrix $\tilde{\Sigma}(\mu, \sigma, \tilde{\mu}, F) \in \mathbb{R}^{3\times 3}$, where $(\hat{\mu}_n^*, \hat{\sigma}_n^*, \tilde{\mu}_n^*)$ are the bootstrap re-calculations obtained from the re-weighted expressions in (3.11). As before, let $\tilde{\Sigma}_n = \tilde{\Sigma}(\hat{\mu}_n, \hat{\sigma}_n, \tilde{\mu}_n, F_n)$ be the empirical estimate of $\tilde{\Sigma}$. Let $A_n = \mathbf{I} - \nabla\mathbf{f}(\hat{\mu}_n, \hat{\sigma}_n, \tilde{\mu}_n)$ be the estimated correction matrices used in the proof of Theorem 3.1 (see page 105 for the definitions). We can show that under certain regularity conditions, $A_n\,\tilde{\Sigma}_n\,A_n' \to \Sigma$ almost surely.

For a matrix C that is symmetric and definite positive, let $C^{1/2}$ be its unique square root. That is, we have $C = C^{1/2}\,[C^{1/2}]'$. Let $C^{-1/2}$ be the inverse of $C^{1/2}$.

If we use classical studentized bootstrap on the vector $(\hat{\mu}_n, \hat{\sigma}_n, \tilde{\mu}_n)$ we use $\Sigma_n^{-1/2}$, the inverse of the square root of the estimate Σ_n, as follows:

$$\sqrt{n}\;\Sigma_n^{-1/2\,c*}\left[\begin{pmatrix} \hat{\mu}_n^{c*} \\ \hat{\sigma}_n^{c*} \\ \tilde{\mu}_n^{c*} \end{pmatrix} - \begin{pmatrix} \hat{\mu}_n \\ \hat{\sigma}_n \\ \tilde{\mu}_n \end{pmatrix}\right],$$

where $\Sigma_n^{-1/2\,c*}$ denotes the evaluation of Σ_n with the bootstrap samples, and $(\hat{\mu}_n^{c*}, \hat{\sigma}_n^{*c}, \tilde{\mu}_n^{c*})'$ are the classical bootstrap re-calculated statistics. From Theorem 3.1 we know that

$$\sqrt{n}\left[\begin{pmatrix} \hat{\mu}_n^{c*} \\ \hat{\sigma}_n^{c*} \\ \tilde{\mu}_n^{c*} \end{pmatrix} - \begin{pmatrix} \hat{\mu}_n \\ \hat{\sigma}_n \\ \tilde{\mu}_n \end{pmatrix}\right] \sim \sqrt{n}\,A_n\left[\begin{pmatrix} \hat{\mu}_n^* \\ \hat{\sigma}_n^* \\ \tilde{\mu}_n^* \end{pmatrix} - \begin{pmatrix} \hat{\mu}_n \\ \hat{\sigma}_n \\ \tilde{\mu}_n \end{pmatrix}\right],$$

where $(\hat{\mu}_n^*, \hat{\sigma}_n^*, \tilde{\mu}_n^*)'$ are the robust bootstrap re-calculated statistics, and $X_n \sim Y_n$ means that both sequences have the same asymptotic distribution. Simple algebra yields $\Sigma_n = A_n \, \tilde{\Sigma}_n \, A_n'$. Hence, $\Sigma_n^{-1/2} = \tilde{\Sigma}_n^{-1/2} \, A_n^{-1}$. Based on these observations we propose to studentize our bootstrap as follows. Let $\tilde{\Sigma}_n^{-1/2\,*}$ and $A_n^{-1\,*}$ be bootstrap evaluations of $\tilde{\Sigma}_n^{-1/2}$ and A_n^{-1} respectively. Let

$$\mathbf{e}^* = \sqrt{n} \; \tilde{\Sigma}_n^{-1/2\,*} \; A_n^{-1\,*} \; A_n \left[\begin{pmatrix} \hat{\mu}_n^* \\ \hat{\sigma}_n^* \\ \tilde{\mu}_n^* \end{pmatrix} - \begin{pmatrix} \hat{\mu}_n \\ \hat{\sigma}_n \\ \tilde{\mu}_n \end{pmatrix} \right] , \tag{3.20}$$

be the studentized robust bootstrap re-calculations of $(\hat{\mu}_n, \hat{\sigma}_n, \tilde{\mu}_n)'$. If we transform $\mathbf{e}^*$ in (3.20) with the covariance matrix estimate $\Sigma_n^{1/2}$ calculated with the original data, we get a bootstrap sample of the joint distribution of $(\hat{\mu}_n, \hat{\sigma}_n, \tilde{\mu}_n)'$. That is, if $\mathbf{h}^* = \Sigma_n^{1/2} \, \mathbf{e}^*$, we can use the first coordinate of these $\mathbf{h}^*$ vectors to get our estimate for the asymptotic distribution of $\hat{\mu}_n$.

Unfortunately, we were unable to prove that the proposed studentized robust bootstrap method achieves (3.18). The reason seems to be that the correction factor we apply to $\hat{\mu}_n^{R*} - \hat{\mu}_n$ is of order $O_P\left(1/\sqrt{n}\right)$.

Numerical experiments show that the studentized robust bootstrap performs slightly better that the non-studentized robust bootstrap, but the difference does not seem to be of the order of magnitude suggested by (3.18) and (3.19). See Chapter 6 for a proposal on how this could be improved.

3.6 Inference

We consider two related approaches to performing statistical inference about the location parameter μ.

The first approach is to approximate the distribution of $\sqrt{n}\,(\hat{\mu}_n - \mu)$ with its normal asymptotic distribution (see Theorem 2.6). If we proceed in this fashion, it is important to have a good estimate of the asymptotic variance V^2 given in that Theorem. We will then build an asymptotic $1-\alpha$ confidence interval for μ of the form $\left(\hat{\mu}_n - z_{\alpha/2}\,\hat{V}_n,\, \hat{\mu}_n + z_{\alpha/2}\,\hat{V}_n\right)$, where $\hat{V}_n$ is an estimate of V, and z_α is the quantile that leaves area equal to α to its right under a standard normal curve. We can also wish to estimate V in order to assess the precision of the point estimate $\hat{\mu}_n$. In either case, it is of interest to have a reliable estimate of V. In Section 3.6.1 we compare the accuracy of four estimates $\hat{V}_n$: the robust bootstrap (RB), the classical bootstrap (CB), Singh's Winsorized bootstrap (WB) (Singh, 1998) and the empirical estimate (AV) based on the formula given in Theorem 2.6. The first three methods use the empirical variance of the re-calculated $\hat{\mu}_n^*$'s to estimate V^2 (see for example Davison and Hinkley, 1997, page 16). The difference among them lies in the way in which the re-computed $\hat{\mu}_n^*$'s are obtained. See Section 3.6.1 for a description of the WB method.

The second approach is based on estimating the distribution of $\sqrt{n}\,(\hat{\mu}_n - \mu)$ without using Theorem 2.6. In this case we focus in constructing confidence intervals for μ of the form $\left(\mathcal{Z}_{1-\alpha/2},\, \mathcal{Z}_{\alpha/2}\right)$, where $\mathcal{Z}_\eta$ satisfies $\lim_{n\to\infty} P\left[\ \sqrt{n}\ (\hat{\mu}_n - \mu) > \mathcal{Z}_\eta\ \right] = \eta$, for $\eta \in [0,1]$. We can use bootstrap methods to obtain such estimates. The basic idea (see for example Davison and Hinkley, 1997, page 18) is to use the empirical distribution of the re-calculated $\hat{\mu}_n^*$'s to obtain estimated quantiles $\hat{\mathcal{Z}}_\eta$. Note that with this method we do not use the symmetry assumption that underlies the asymptotic normal approximation in the previous approach. To estimate these quantiles we considered the studentized classical bootstrap (see Davison and Hinkley, 1997, page 29), the studentized robust bootstrap and the Winsorized bootstrap (Singh, 1998). We have also studied the non-studentized classical and robust bootstrap, but the studentized methods performed better. Details and the results of our experiment are discussed in Section 3.6.2.

In all the simulation experiments we used an MM-location estimate with $\psi_{1.345}$ in Huber's family. The S-scale was calculated with the function $\rho_{1.04086}$ in (2.14). This election yields an estimate $\hat{\mu}_n$ that has 50% breakdown point and that is 95% efficient when the data is normally distributed.

Remark 3.6 - Note that the correction coefficient b_n in (3.7)

$$b_n = n\, b \frac{\left[\sum_{i=1}^n \psi'\left(r_i / \hat{\sigma}_n\right) r_i\right]}{\left[\sum_{i=1}^n \psi'\left(r_i / \hat{\sigma}_n\right)\right] \left[\sum_{i=1}^n \rho'\left(\tilde{r}_i / \hat{\sigma}_n\right) \tilde{r}_i / \hat{\sigma}_n\right]},$$

could potentially be unstable due to small values in the denominator. It is clear that, almost surely,

$$\lim_{n\to\infty} b_n \;=\; b_\infty\left(F\right) \;=\; b \frac{E_F\left[\psi'\left(u\right)\, u\right]}{E_F\left[\psi'\left(u\right)\right]\; E_F\left[\rho'\left(u\right)\, u\right]}.$$

Lemma 7.20 gives a bound for $b_\infty\left(F\right)$ for any distribution F_ϵ in an ϵ neighbourhood of the standard normal distribution, when $\psi = \psi_c$ belongs to Huber's family and $\rho = \rho_k$ given by (2.14).

In our simulation studies we used $c = 1.345$, $b = 1/2$ and $k = 1.041$, so that the previous lemma yields

$$|b_\infty(F_\epsilon)| \leq 1$$

for $\epsilon \in (0, 0.3)$. Hence we restricted $b_n \in (-1.5, 1.5)$.

3.6.1 Asymptotic variance estimation

In this section we compare the following estimates of the asymptotic variance V^2 of the sequence $\hat{\mu}_n$: the classical bootstrap (CB), the robust bootstrap (RB), Singh's Winsorized bootstrap (WB) (Singh, 1998) and the empirical estimate based on the asymptotic formula (AV).

As mentioned before, the three bootstrap-based estimates of V^2 are the empirical variance of the re-calculated $\hat{\mu}_n^*$'s. They differ in the way in which the re-computed statistics are obtained. The last estimate AV of V^2 is given by $\hat{V}_n^2 = V(\hat{\mu}_n, \hat{\sigma}_n, F_n)^2$ where $V(\mu, \sigma, F)$ is given in (2.71).

We now briefly describe Singh's Winsorized bootstrap (Singh, 1998). Let $x_1, \ldots, x_n$ be the original sample. Let $\hat{\sigma}_n$ be a robust scale estimate and $\hat{\mu}_n$ a robust location estimate calculated as in (2.12) with a function ψ_c from Huber's family (2.5). For a fixed constant h and a bootstrap sample $x_1^*, \ldots, x_n^*$ define the Winsorized bootstrap sample as follows

$$y_i^* = \begin{cases} x_i^* & \text{if } |(x_i^* - \hat{\mu}_n)/\,\hat{\sigma}_n| \leq h \\ \hat{\mu}_n - h\,\hat{\sigma}_n & \text{if } (x_i^* - \hat{\mu}_n)/\,\hat{\sigma}_n < -h \\ \hat{\mu}_n + h\,\hat{\sigma}_n & \text{if } (x_i^* - \hat{\mu}_n)/\,\hat{\sigma}_n > h. \end{cases}$$

for $i = 1, \ldots, n$. Singh proposes to use $h = 1.5\,c$ where c is the constant used in ψ_c. The Winsorized bootstrap evaluation of $\hat{\mu}_n^{W*}$ is the solution of

$$\sum_{i=1}^{n} \psi_c\left((y_i^* - \hat{\mu}_n^{W*})/\,\hat{\sigma}_n\right) = 0$$

We then estimate the asymptotic standard error V of $\hat{\mu}_n$ by calculating the empirical standard deviation of the bootstrapped $\hat{\mu}_n^{W*}$'s.

In this study we considered distributions in the family

$$F_\epsilon(x) = (1-\epsilon)\,\Phi(x) + \epsilon\,\Phi\left((x-7)/\,0.1\right), \tag{3.21}$$

where Φ denotes the standard normal cumulative distribution function. That is, $100 \times \epsilon$ per cent of the observations are outliers centered around 7. Other values for the center of the contamination yielded similar results.

We needed to simulate observations of a random variable X with distribution function F_ϵ given by (3.21). A realization of such a random variable can be easily generated in the following way. For each $i = 1, \ldots, n$ let $B_i \sim \text{Binomial}(1, \epsilon)$, independent from each other. Then

$$X_i = \begin{cases} Z_i^P & \text{if } B_i = 0 \\ Z_i^C & \text{if } B_i = 1 \end{cases} \tag{3.22}$$

where $Z_i^P \sim N(0,1)$ and $Z_i^C \sim N(7, 0.1)$, both independent of B_i. When drawn in this fashion, every sample $x_1, \ldots, x_n$ will contain a different number of outliers. This random proportion of outliers has expected value equal to ϵ.

We used samples of size 20, 30 and 50. The proportions ϵ of contamination considered were 0.0, 0.1, 0.2 and 0.3. For each F_ϵ we computed the correct asymptotic variance $V^2(\mu, \sigma, F_\epsilon)$ of the sequence $\hat{\mu}_n$ (see Theorem 2.6). We then simulated 3,000 samples from F_ϵ and obtained the asymptotic variance estimates $\hat{V}_i$, $i = 1, \ldots, 3000$ with each of the four methods. We report the averages of the following two "loss functions", where V is the actual asymptotic standard deviation of $\hat{\mu}_n$ when $X \sim F_\epsilon$:

$$d_q(\hat{V}_i, V) = \left(\frac{\hat{V}_i}{V} - 1 \right)^2, \tag{3.23}$$

and

$$d_l(\hat{V}_i, V) = \left(\log \left(\frac{\hat{V}_i}{V} \right) \right)^2. \tag{3.24}$$

The quadratic measure (3.23) is more sensitive to over-estimation of V, but does not penalize under-estimation with the same intensity (intuitively: the worst under-estimation for V is $\hat{V}_i = 0$ and hence there is an upper bound for $d_q(\hat{V}_i, V)$ when $\hat{V}_i < V$, while d_q is unbounded for $\hat{V}_i > V$). The logarithmic loss d_l (3.24) penalizes under-estimation with more intensity because now $d_l(\hat{V}_i, V) \to +\infty$ when $\hat{V}_i \to 0$. With this loss function over-estimation receives less weight than with d_q because $d_l(\hat{V}_i, V) / d_q(\hat{V}_i, V) \to 0$ when $\hat{V}_i \to +\infty$.

Tables 3.2 and 3.3 provide detailed results of this Monte Carlo study. In Figures 3.5 and 3.6 we summarize these results. We see that in all cases there is a value ϵ^*

such that for $\epsilon \leq \epsilon^*$ all methods behave similarly, but when the proportion of outliers exceeds ϵ^* the robust bootstrap shows a clear advantage in performance.

For example, with d_q and $n = 30$, for $\epsilon \leq 0.10$ there is not much difference among the different methods. But when $\epsilon \geq 0.20$ there is a clear change in the pattern: the robust bootstrap remains relatively stable, the Winsorized bootstrap and the empirical variance grow notably faster and the classical bootstrap is completely unreliable.

For $n = 50$ and d_q the classical bootstrap breaks-down for $\epsilon > 0.1$ while the other three methods remain close to each other. When $\epsilon > 0.20$ we see that the robust bootstrap remains stable while the other methods break-down (the Winsorized bootstrap resulted the second best method in this study).

With d_l the differences in performance are smaller but the pattern is the same as with d_q. Note that the robust bootstrap has a comparable average loss for values of ϵ where most of the methods considered here remain close to each other. When the proportion of contamination is large and there is a clear differentiation in performances, the robust bootstrap is consistently the best method.

This study illustrates the numerical stability of our method when the interest lies in estimating V^2, the asymptotic variance of the sequence $\hat{\mu}_n$. These results show that inference based on our method is more stable than that based on the other three proposals for high proportions of contamination, and at the same time, remains comparable for small proportions of outliers. The robust bootstrap is also much faster

$$E\left(\hat{V}_n \big/ V - 1\right)^2$$

		Robust	Classical	Winsorized	Empirical
n	ϵ	Bootstrap	Bootstrap	Bootstrap	AV
20	0.00	0.268 (0.993)	0.219 (0.475)	0.218 (0.464)	0.271 (1.120)
	0.10	3.258 (41.46)	33.27 (411.9)	2.908 (51.35)	3.256 (50.89)
	0.20	17.99 (88.02)	189.7 (632.3)	19.95 (115.9)	29.14 (189.6)
	0.30	10.17 (26.84)	89.81 (134.0)	11.83 (31.14)	18.92 (66.11)
30	0.00	0.135 (0.268)	0.123 (0.230)	0.122 (0.221)	0.132 (0.261)
	0.10	1.304 (29.35)	9.798 (329.8)	1.505 (54.10)	1.418 (40.43)
	0.20	13.60 (114.1)	116.0 (595.7)	20.36 (191.2)	34.46 (456.0)
	0.30	17.94 (53.36)	124.1 (241.3)	38.63 (116.4)	76.22 (294.3)
50	0.00	0.067 (0.112)	0.061 (0.094)	0.060 (0.093)	0.065 (0.114)
	0.10	0.233 (0.759)	0.361 (1.786)	0.203 (0.609)	0.215 (0.685)
	0.20	3.417 (38.15)	31.16 (366.3)	6.124 (161.3)	6.307 (161.7)
	0.30	12.34 (63.11)	95.15 (301.1)	40.62 (224.2)	122.3 (1012)

Table 3.2: Comparison of asymptotic variance estimates - quadratic measure

to compute than these alternatives.

3.6.2 Coverage and lengths of confidence intervals

In this study we compare the coverage and mean lengths of confidence intervals based on the following methods: studentized classical bootstrap (B-t), studentized robust bootstrap (RB-t), and Singh's Winsorized bootstrap (WB).

Let $\hat{\mu}_n$ be an MM-location estimate and let $\hat{\sigma}_n$ be the associated S-scale. The classical bootstrap-t (B-t) method was implemented as described above with $T_n = \hat{\mu}_n$ and $\hat{U}_n = V\left(\hat{\mu}_n, \hat{\sigma}_n, F_n\right)$, where $V\left(\mu, \sigma, F\right)$ is given in Theorem 2.6. The robust

$$E\left(\log\left(\hat{V}_n\big/V\right)\right)^2$$

n	ϵ	Robust Bootstrap	Classical Bootstrap	Winsorized Bootstrap	Empirical AV
20	0.00	0.198 (0.322)	0.177 (0.264)	0.195 (0.301)	0.198 (0.327)
	0.10	0.436 (0.863)	0.875 (1.982)	0.390 (0.791)	0.415 (0.841)
	0.20	1.130 (1.871)	2.896 (4.091)	1.081 (1.902)	1.166 (2.075)
	0.30	1.642 (1.730)	3.696 (3.211)	1.809 (1.863)	1.907 (2.075)
30	0.00	0.120 (0.187)	0.110 (0.171)	0.114 (0.182)	0.119 (0.186)
	0.10	0.279 (0.552)	0.418 (1.048)	0.245 (0.503)	0.269 (0.540)
	0.20	0.813 (1.570)	1.903 (3.314)	0.764 (1.693)	0.844 (1.858)
	0.30	1.559 (1.974)	3.556 (3.667)	1.922 (2.579)	2.096 (3.127)
50	0.00	0.065 (0.094)	0.059 (0.085)	0.060 (0.087)	0.064 (0.093)
	0.10	0.143 (0.226)	0.166 (0.299)	0.134 (0.206)	0.139 (0.216)
	0.20	0.480 (0.900)	0.913 (1.921)	0.405 (0.897)	0.476 (0.967)
	0.30	1.049 (1.596)	2.570 (3.332)	1.257 (2.350)	1.364 (2.797)

Table 3.3: Comparison of asymptotic variance estimates - logarithmic measure

bootstrap-t was performed as discussed in Section 3.5. For a description of the Winsorized bootstrap (WB) see Section 3.6.1.

To construct $1-\alpha$ confidence intervals for μ we estimate the quantiles q_t that satisfy

$$\lim_{n\to\infty} P\left(q_{1-\alpha/2} \leq \sqrt{n}\,(\hat{\mu}_n - \mu)/\,\hat{U}_n \leq q_{\alpha/2}\right) = 1-\alpha\,.$$

Let $\hat{F}^*$ be the empirical distribution function of the re-computed $\sqrt{n}\,(\hat{\mu}_n^* - \hat{\mu}_n)/\,\hat{U}_n^*$ with each bootstrap-t method. An estimate of q_α is given by the solution $\hat{q}_\alpha^*$ of $\hat{F}^*\,(\hat{q}_\alpha^*) = \alpha$. The confidence interval is

$$\left(\hat{\mu}_n - \hat{q}_{\alpha/2}^*\,\hat{U}_n\,,\, \hat{\mu}_n - \hat{q}_{1-\alpha/2}^*\,\hat{U}_n\right).$$

Confidence intervals based on Singh's Winsorized bootstrap used the empirical distribution function of the recomputed estimates as an estimate of the cumulative distribution function of $\sqrt{n}\ (\hat{\mu}_n - \mu)$. If $\hat{F}_W^*$ denotes that estimate, then, for $\alpha \in [0,1]$ the α-th quantile estimate based on the Winsorized bootstrap $\hat{q}_{W\,\alpha}^*$ solves $\hat{F}_W^*\ (\hat{q}_{W\,\alpha}^*) = \alpha$. The interval is

$$\left(\hat{\mu}_n - \hat{q}_{W\,\alpha/2}^*\,,\, \hat{\mu}_n - \hat{q}_{W\,1-\alpha/2}^*\right).$$

The tails of the distribution estimates obtained by the classical bootstrap are potentially unstable (see the discussion on page 86). We expect the robust bootstrap to show more clearly its advantage over the other methods when we estimate extreme quantiles (for example the quantile $q_{0.005}$ needed to build a 99% confidence interval).

We considered data generated by distribution functions F_ϵ in (3.21) with ϵ equal to 0.00, 0.10, 0.20 and 0.30 and n = 20, 30 and 50. For each combination of n and ϵ we generated 3,000 samples and constructed the confidence intervals as described above. Tables 3.4 and 3.5 show the results. We display the same results in Figures 3.7 to 3.12. For each sample size (20, 30 and 50) we plot the empirical coverage level on the x-axis and the the mean length on the y-axis. The labels 0, 1, 2 and 3 correspond to 0%, 10%, 20% and 30% of outliers. The results corresponding to each method are joined by a line. The ideal trajectory will stay near the $1-\alpha$ line without moving upward.

As expected we see that for larger proportions of contamination we obtain larger mean lengths. For $n = 20$ or 30 the studentized bootstrap confidence intervals are noticeably longer that the ones based on the other methods. We also note that the studentized bootstrap confidence intervals have smaller coverage levels than nominal. There is no important difference in the performance of the robust bootstrap compared with the Winsorized bootstrap, but the computational demands are significantly less for the robust bootstrap. We conclude that the robust bootstrap performs equivalently to other robust proposals and is faster to compute, so we recommend its use.

n	ϵ	Studentized robust bootstrap	Studentized bootstrap	Winsorized bootstrap
20	0.00	0.942 (1.01)	0.953 (1.09)	0.938 (0.91)
	0.10	0.950 (1.25)	0.931 (1.26)	0.937 (1.11)
	0.20	0.962 (1.74)	0.912 (1.87)	0.958 (1.49)
	0.30	0.970 (2.56)	0.940 (4.01)	0.972 (2.11)
30	0.00	0.949 (0.79)	0.954 (0.81)	0.948 (0.74)
	0.10	0.958 (0.97)	0.933 (0.95)	0.950 (0.90)
	0.20	0.964 (1.28)	0.910 (1.20)	0.963 (1.18)
	0.30	0.971 (2.01)	0.919 (2.10)	0.971 (1.73)
50	0.00	0.946 (0.59)	0.948 (0.59)	0.942 (0.57)
	0.10	0.958 (0.72)	0.948 (0.71)	0.958 (0.70)
	0.20	0.967 (0.94)	0.924 (0.88)	0.966 (0.91)
	0.30	0.980 (1.43)	0.908 (1.29)	0.982 (1.32)

Table 3.4: Coverage and length of 95% confidence intervals for the location-scale model

n	ϵ	Studentized robust bootstrap	Studentized bootstrap	Winsorized bootstrap
20	0.00	0.988 (1.43)	0.988 (1.60)	0.968 (1.12)
	0.10	0.991 (1.81)	0.986 (2.04)	0.983 (1.42)
	0.20	0.994 (2.52)	0.982 (4.55)	0.985 (1.90)
	0.30	0.998 (3.80)	0.990 (11.5)	0.995 (2.63)
30	0.00	0.988 (1.08)	0.989 (1.12)	0.979 (0.93)
	0.10	0.993 (1.33)	0.976 (1.31)	0.981 (1.14)
	0.20	0.997 (1.77)	0.967 (1.67)	0.988 (1.50)
	0.30	0.999 (2.68)	0.978 (3.22)	0.995 (2.14)
50	0.00	0.989 (0.80)	0.988 (0.80)	0.979 (0.73)
	0.10	0.992 (0.97)	0.988 (0.94)	0.991 (0.89)
	0.20	0.999 (1.27)	0.974 (1.16)	0.993 (1.16)
	0.30	0.999 (1.97)	0.978 (1.86)	0.997 (1.70)

Table 3.5: Coverage and length of 99% confidence intervals for the location-scale model

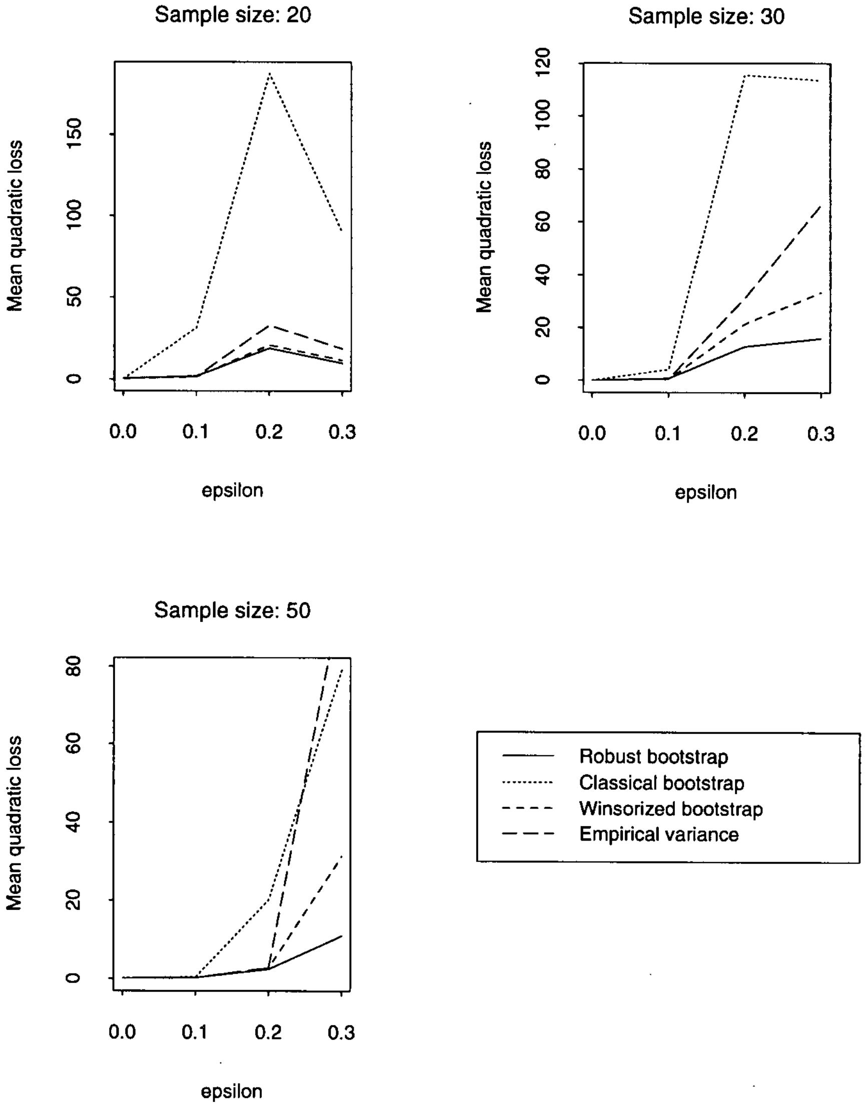


Figure 3.5: Comparison of asymptotic variance estimates - quadratic measure

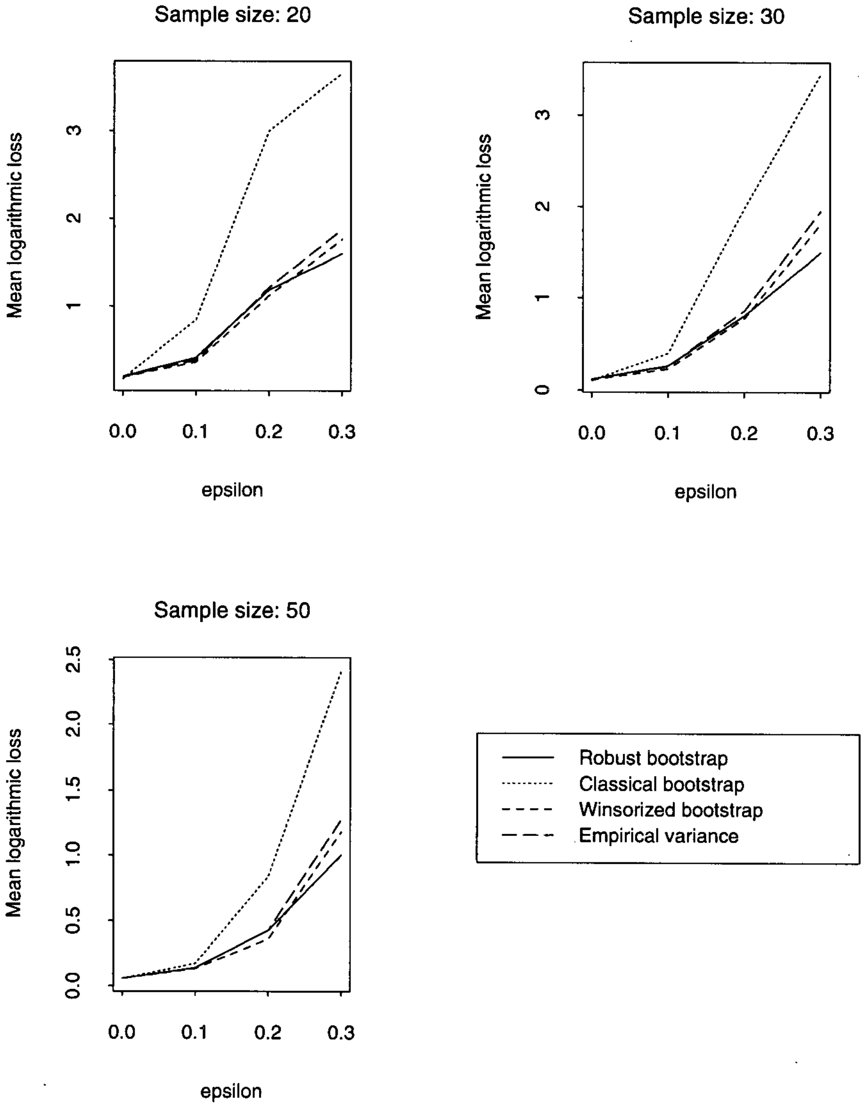


Figure 3.6: Comparison of asymptotic variance estimates - logarithmic measure

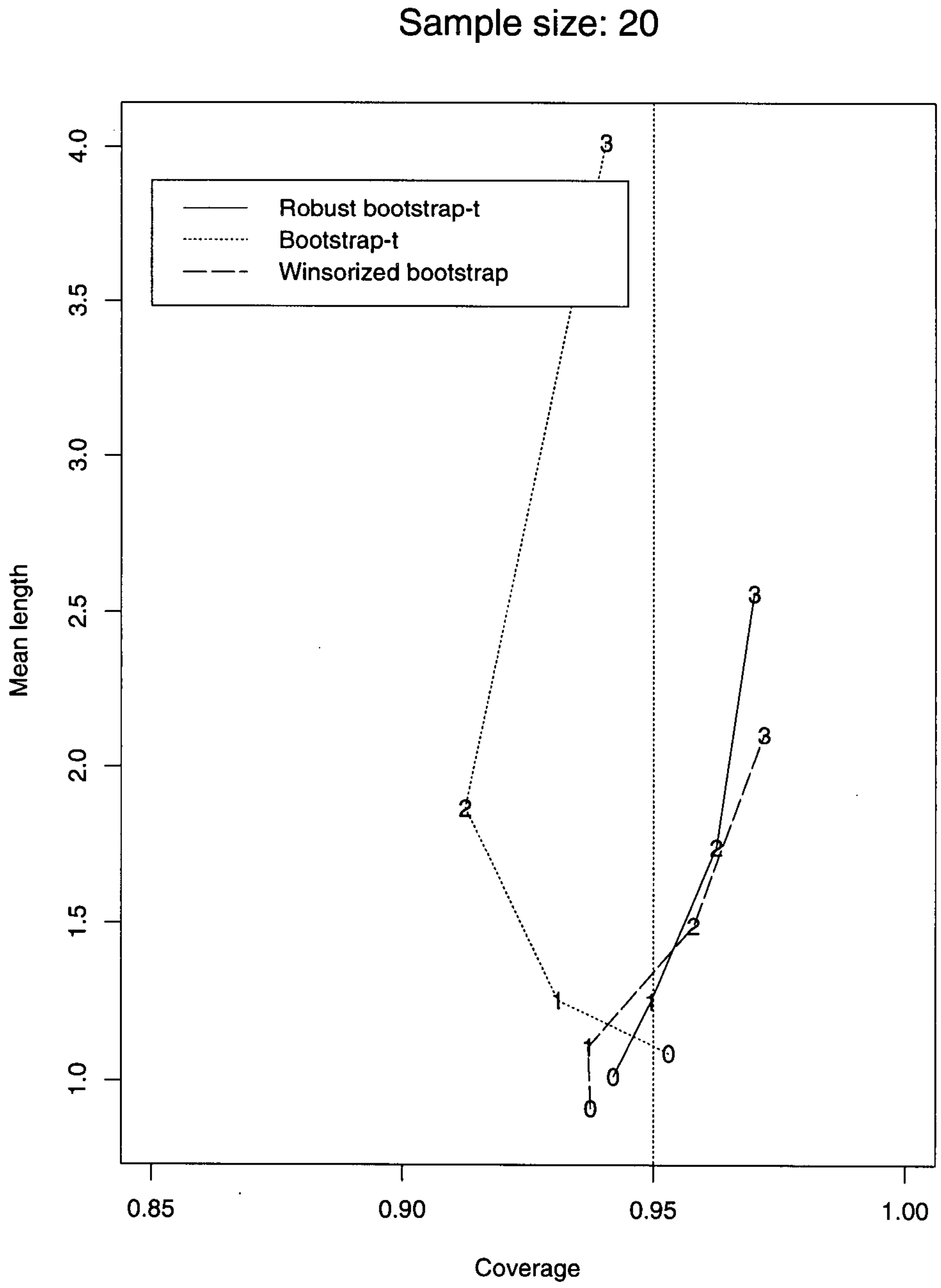


Figure 3.7: Location-scale model 95% confidence intervals for $n = 20$ - Labels 0, 1, 2 and 3 correspond to $\epsilon = 0.0$, 0.1, 0.2 and 0.3 respectively.

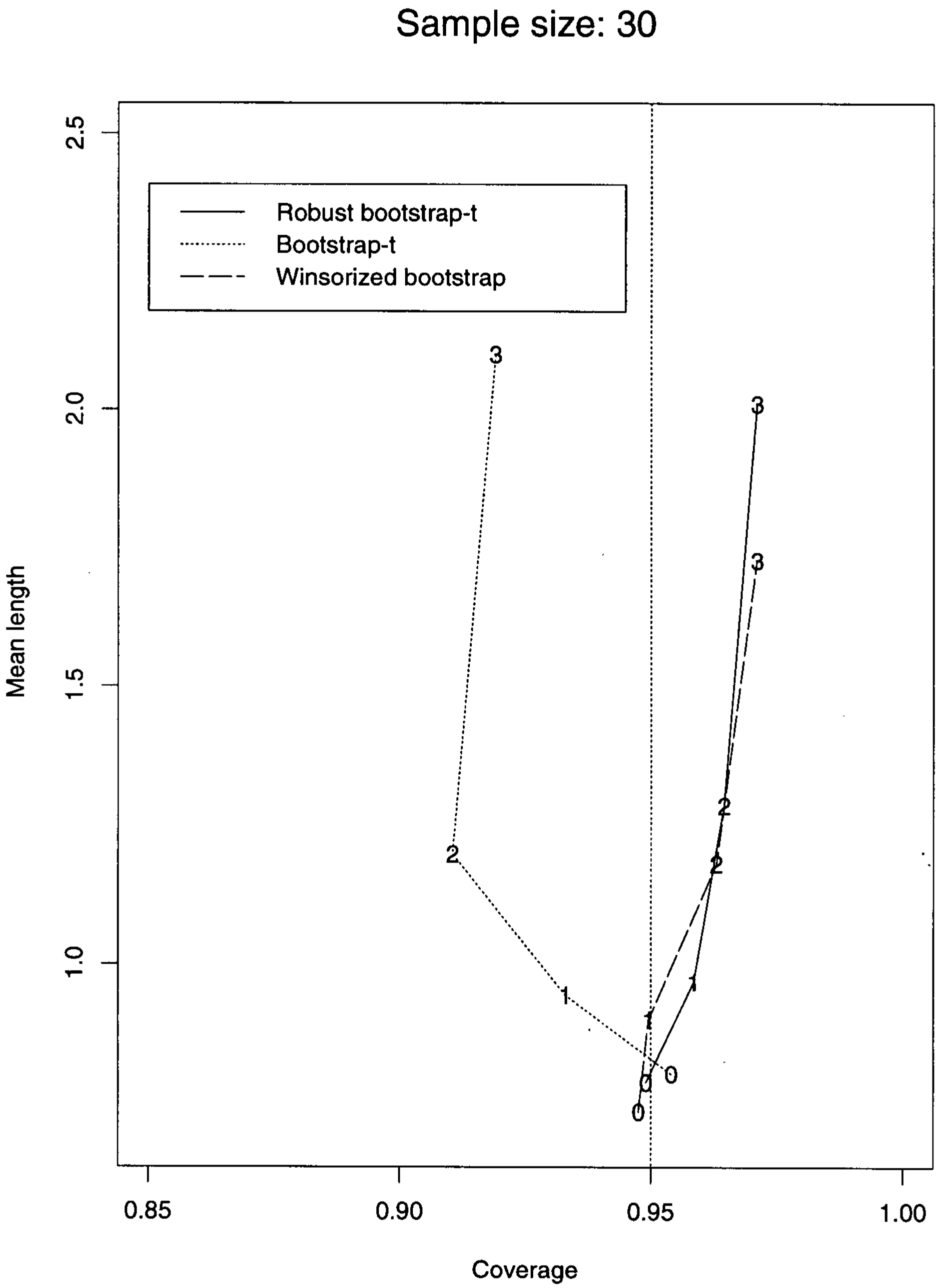


Figure 3.8: Location-scale model 95% confidence intervals for $n = 30$ - Labels 0, 1, 2 and 3 correspond to $\epsilon = 0.0$, 0.1, 0.2 and 0.3 respectively.

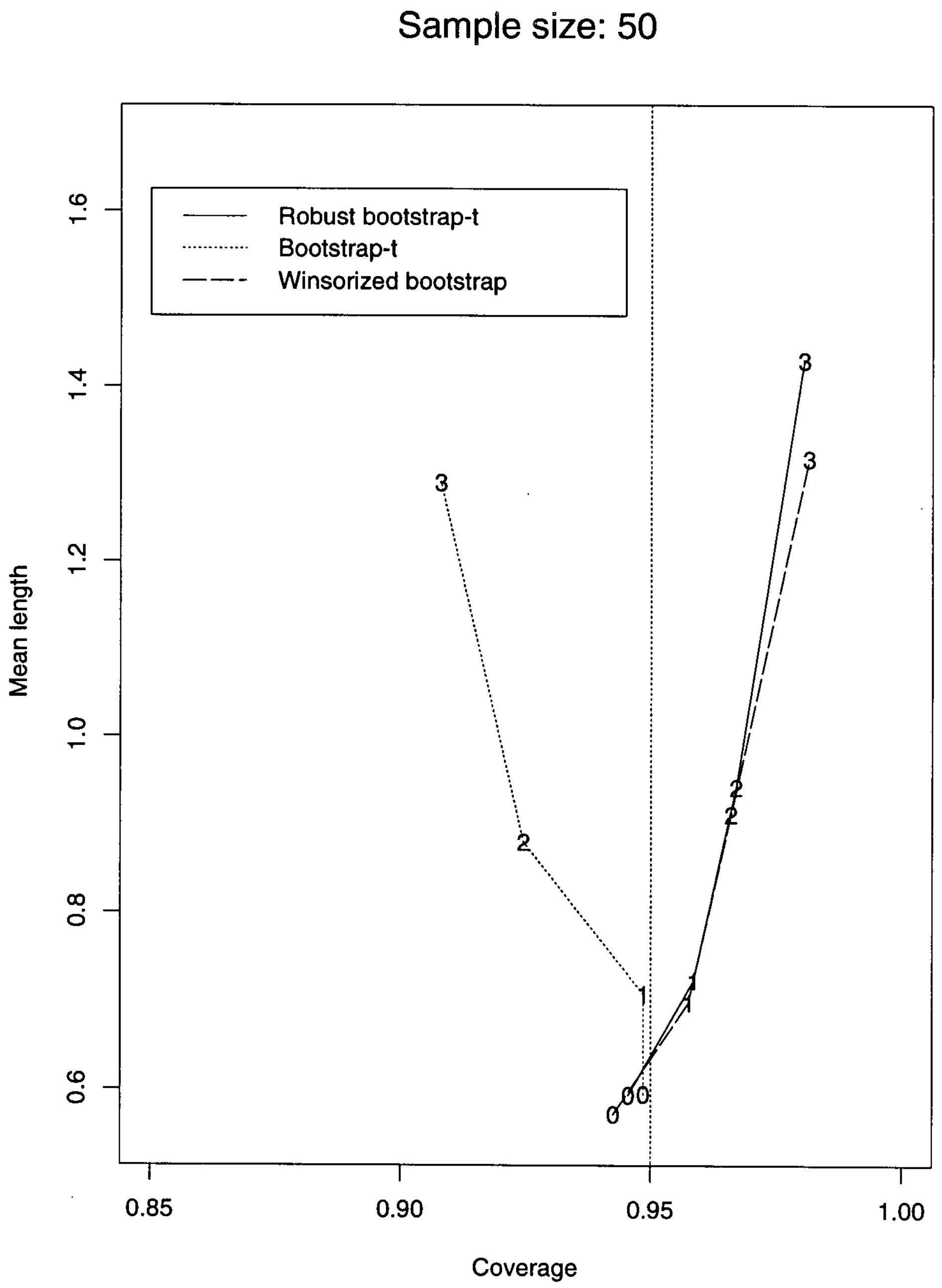


Figure 3.9: Location-scale model 95% confidence intervals for $n = 50$ - Labels 0, 1, 2 and 3 correspond to $\epsilon = 0.0$, 0.1, 0.2 and 0.3 respectively.

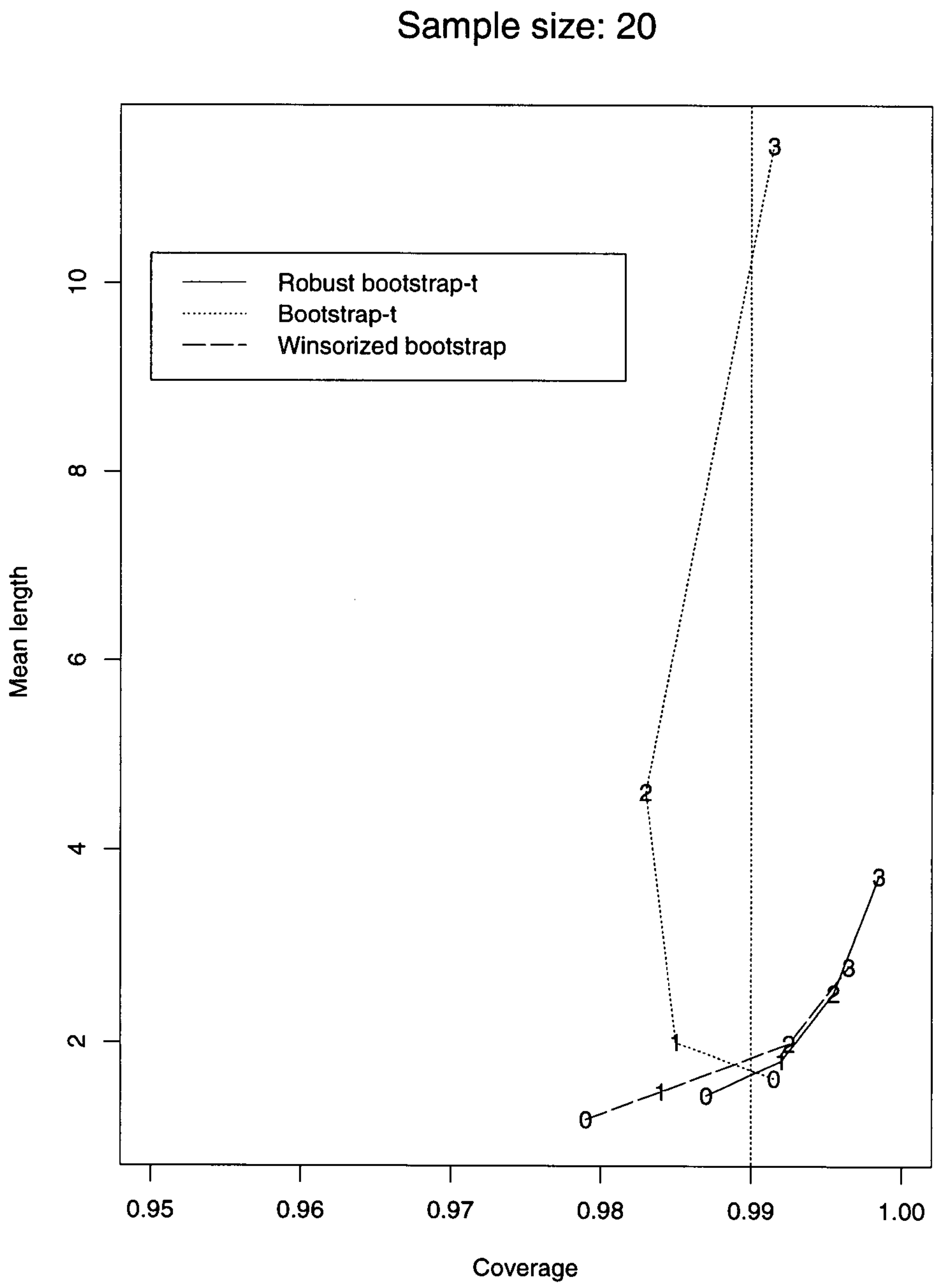


Figure 3.10: Location-scale model 99% confidence intervals for $n = 20$ - Labels 0, 1, 2 and 3 correspond to $\epsilon = 0.0, 0.1, 0.2$ and 0.3 respectively.

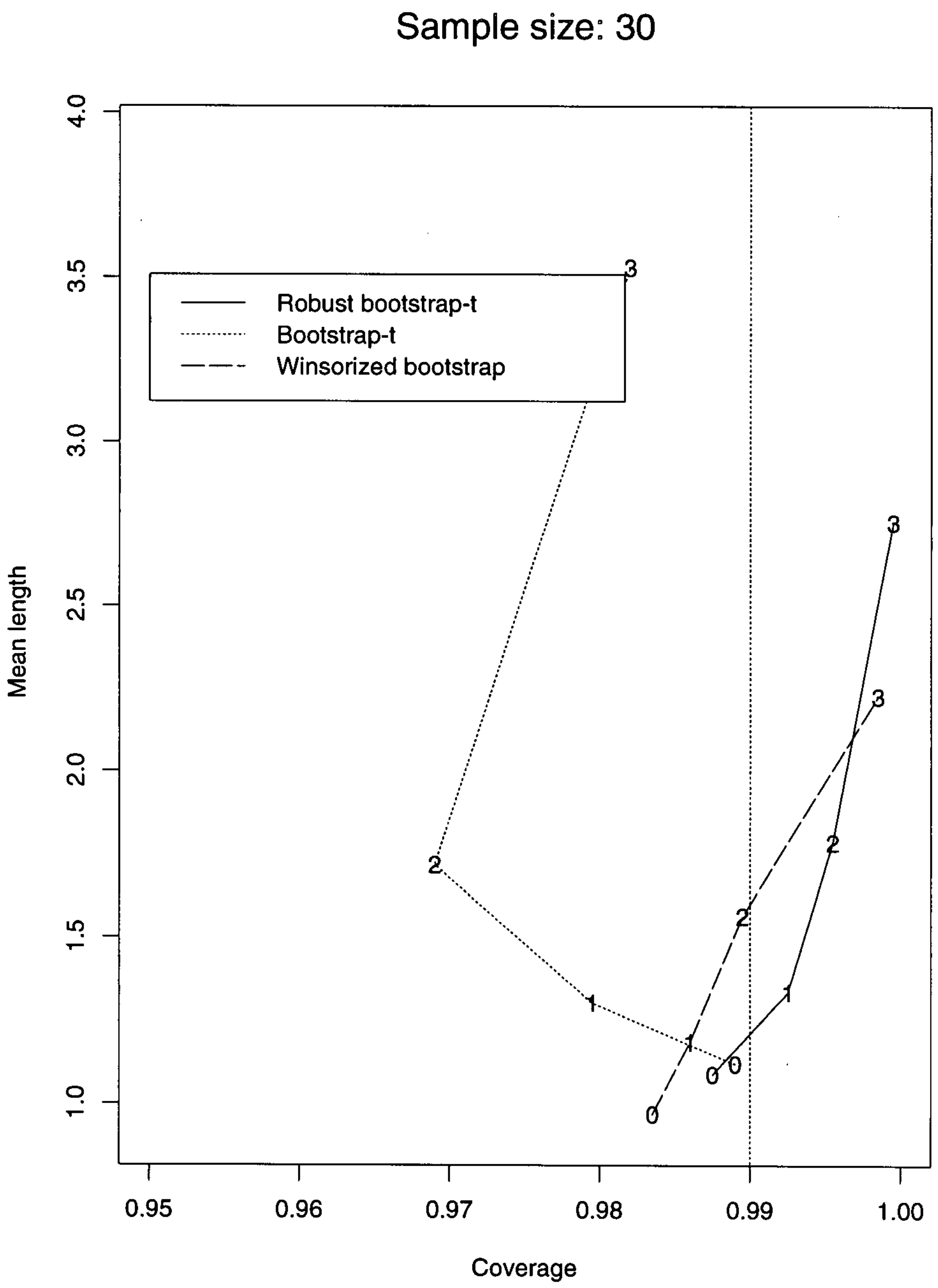


Figure 3.11: Location-scale model 99% confidence intervals for $n = 30$ - Labels 0, 1, 2 and 3 correspond to $\epsilon = 0.0$, 0.1, 0.2 and 0.3 respectively.

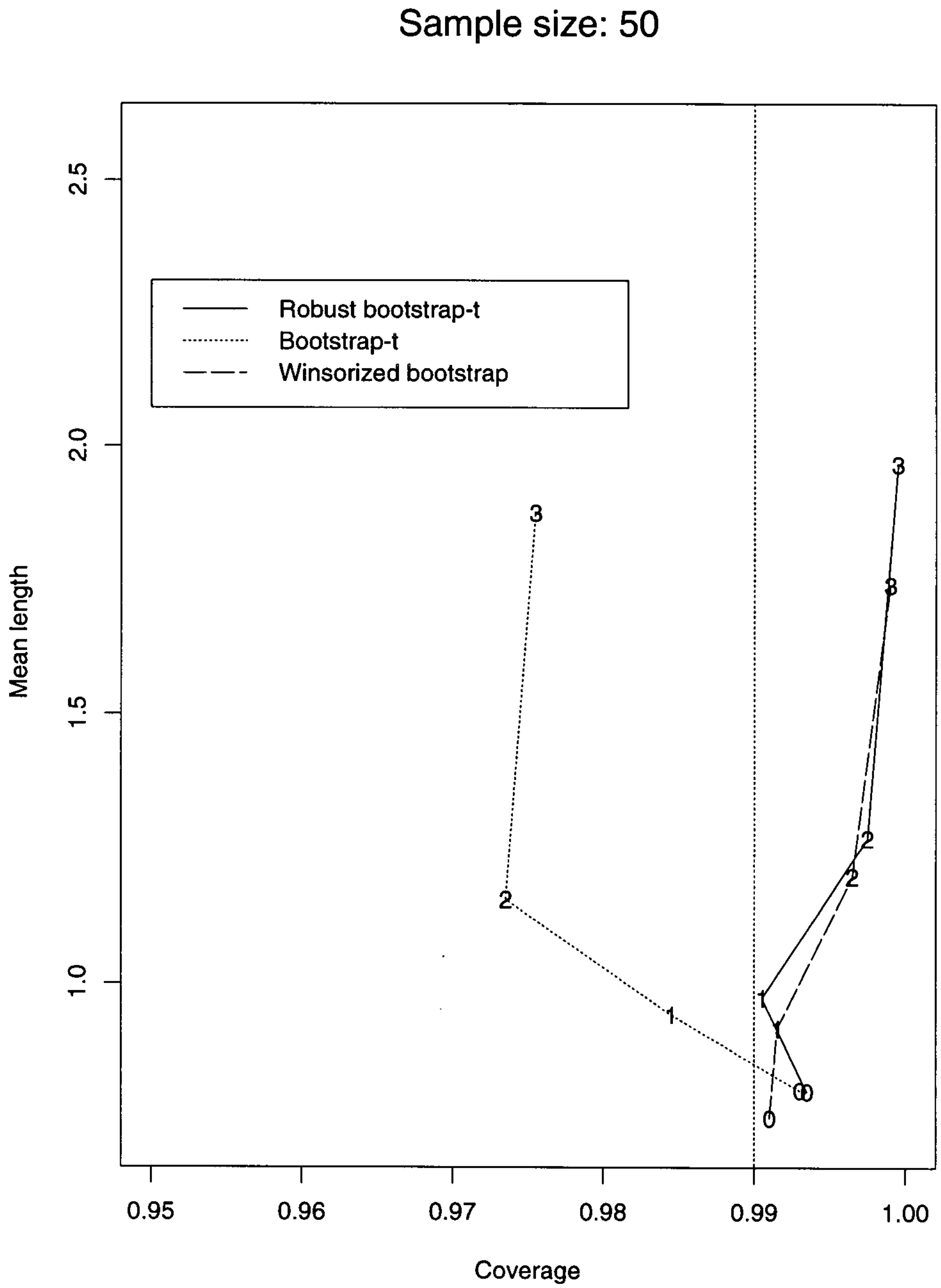


Figure 3.12: Location-scale model 99% confidence intervals for $n = 50$ - Labels 0, 1, 2 and 3 correspond to $\epsilon = 0.0$, 0.1, 0.2 and 0.3 respectively.

Chapter 4

Global asymptotic properties of robust estimates for the linear regression model

In this chapter we extend the results of Chapter 2 to the linear regression model. We first describe the class of MM-regression estimates (Yohai, 1987) and discuss its robustness properties when the data are generated by a distribution belonging to the ϵ-contamination neighbourhood of a central distribution H_0. This gross-error neighbourhood allows for contamination both in the errors and in the predictor variables (see (4.10)). We show that under certain regularity conditions the S-scale, S-regression and the MM-regression estimates are consistent for any distribution in this neighbourhood. We also show that with some additional regularity conditions, the S-

and MM-regression estimates are asymptotically normal for arbitrary distributions in the neighbourhood.

4.1 Definitions

Consider the following linear regression model. Let $y_1, \ldots, y_n$ be n independent observations satisfying

$$y_i = \boldsymbol{\beta}_0' \mathbf{x}_i + \epsilon_i, \qquad i = 1, \ldots, n, \tag{4.1}$$

where $\boldsymbol{\beta}_0 \in \mathbb{R}^p$ is the parameter of interest, $\mathbf{x}_i$ are n p-dimensional covariates, and the errors ϵ_i are i.i.d. with mean zero and constant variance σ_0^2. We will consider random covariates in (4.1). Asymptotic theory for robust regression estimates with fixed explanatory variables has yet to be studied in detail. To our knowledge, only results for M-regression estimates (Yohai and Maronna, 1979), S-regression estimates (Davies, 1993) and GM-estimates (Wiens, 1996) have been published. The errors ϵ_i are assumed to be independent of the explanatory variables. We consider σ_0, the dispersion parameter of the errors, a nuisance parameter. If the model (4.1) includes an intercept write $\mathbf{x}_i = (1, \mathbf{z}_i')'$ where $\mathbf{z}_i$ are the explanatory variables. Otherwise we have $\mathbf{x}_i = \mathbf{z}_i$.

Robust regression estimates were first introduced by Huber (1973, 1981). Let $\psi : \mathbb{R} \to \mathbb{R}$ be odd, non-decreasing and bounded. The M-regression estimate is the solution $\hat{\boldsymbol{\beta}}_n$ of the equation

$$\sum_{i=1}^{n} \psi\left((y_i - \mathbf{x}_i' \hat{\boldsymbol{\beta}}_n) / \hat{\sigma}_n\right) \mathbf{x}_i = \mathbf{0}, \tag{4.2}$$

where $\hat{\sigma}_n$ is a robust estimate of the scale of the residuals. If ρ is such that $\psi = \rho'$ then $\hat{\boldsymbol{\beta}}_n$ can be defined as

$$\hat{\boldsymbol{\beta}}_n = \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^p} \sum_{i=1}^{n} \rho\left((y_i - \mathbf{x}_i' \boldsymbol{\theta}) / \hat{\sigma}_n \right) . \tag{4.3}$$

Note that because ψ is non-decreasing the corresponding ρ is unbounded. These estimates have breakdown point 0 (see Definition 4.4) because the above equation does not take into account the leverage of the observations (see Weisberg, 1985, page 111) to down-weight them.

A generalization of the above class of estimates is given by the generalized M-estimates (GM-estimates). Let $\eta : \mathbb{R}^p \times \mathbb{R} \rightarrow \mathbb{R}$ satisfy:

- for all $\mathbf{x} \in \mathbb{R}^p$, $\eta(\mathbf{x}, \cdot)$ is continuous except on a finite set $\mathcal{C}(\mathbf{x})$;
- for each $\mathbf{x} \in \mathbb{R}^p$, $\eta(\mathbf{x}, \cdot)$ is odd;
- $\eta(\mathbf{x}, r) \geq 0$ for $\mathbf{x} \in \mathbb{R}^p$ and $r \geq 0$.

The GM-regression estimate $\hat{\boldsymbol{\beta}}_n$ is defined by

$$\sum_{i=1}^{n} \eta\left(\mathbf{x}_i, (y_i - \mathbf{x}_i' \hat{\boldsymbol{\beta}}_n) / \hat{\sigma}_n\right) \mathbf{x}_i = \mathbf{0} .$$

(see for instance Hill, 1977; Krasker, 1980; Krasker and Welsch, 1982 and Hampel *et al.*, 1986). All proposals for η can be written as $\eta(\mathbf{x}, r) = \omega(\mathbf{x})\ \psi(r\, v(\mathbf{x}))$, for different choices of the weight functions $\omega : \mathbb{R}^p \rightarrow \mathbb{R}_+$, $\psi : \mathbb{R} \rightarrow \mathbb{R}$ and $v : \mathbb{R}^p \rightarrow \mathbb{R}_+$. For example, setting $\omega(\mathbf{x}) \equiv v(\mathbf{x}) \equiv 1$ we obtain Huber's estimates (4.2). If $v(\mathbf{x}) \equiv 1$ we obtain Mallow's family $\eta(\mathbf{x}, r) = \omega(\mathbf{x})\ \psi(r)$. Maronna, Bustos and Yohai (1979) showed that these estimates have breakdown point at most $1/(p+1)$ where p is the number of covariates in the model, including the intercept if present.

Rousseeuw (1984) introduced the Least Median of Squares (LMS) estimate and the Least Trimmed Squares (LTS) estimates. These estimates minimize the median and the trimmed mean of the squared residuals respectively. The LMS has the highest achievable breakdown point, namely 50%, independently of the number of explanatory variables. Unfortunately, the LMS does not have a $\sqrt{n}$-asymptotic distribution (Davies, 1990) and the LTS is computationally very demanding.

Rousseeuw and Yohai (1984) introduced the class of S-regression estimates. They are defined as the set of coefficients $\hat{\boldsymbol{\beta}}_n$ that minimizes an M-scale of the corresponding residuals (compare with Definition 2.6).

Consider a loss function $\rho : \mathbb{R} \to \mathbb{R}_+$ that satisfies the following set of regularity conditions:

R.1 $\rho(-u) = \rho(u)$ for all $u \in \mathbb{R}$, and $\rho(0) = 0$;

R.2 ρ is continuously differentiable;

R.3 $\sup_x \rho(x) = 1$;

R.4 if $\rho(u) < 1$ and $0 \leq v < u$ then $\rho(v) < \rho(u)$.

Definition 4.1 - **S-regression estimates** *Let $\rho : \mathbb{R} \to \mathbb{R}_+$ satisfy conditions R.1 to R.4 above. Let $b \in (0,1]$. The S-regression estimate $\tilde{\boldsymbol{\beta}}_n$ solves*

$$\tilde{\boldsymbol{\beta}}_n = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \hat{\sigma}_n(\boldsymbol{\beta}) \ ,$$

where $\hat{\sigma}_n(\boldsymbol{\beta})$ satisfies

$$\frac{1}{n}\sum_{i=1}^{n}\rho\left((y_i-\mathbf{x}_i'\boldsymbol{\beta})/\hat{\sigma}_n(\boldsymbol{\beta})\right)=b\,. \tag{4.4}$$

The corresponding S-scale estimate $\hat{\sigma}_n$ is

$$\hat{\sigma}_n=\inf_{\boldsymbol{\beta}\in\mathbb{R}^p}\hat{\sigma}_n(\boldsymbol{\beta})=\hat{\sigma}_n(\tilde{\boldsymbol{\beta}}_n)\,. \tag{4.5}$$

S-regression estimates have been shown to be asymptotically normal when the errors distribution is symmetric (Rousseeuw and Yohai, 1984). Unfortunately, for these estimates there is a trade-off between high breakdown point and high efficiency when the errors follow a standard normal distribution. The function ρ in the above definition can be chosen so that the resulting S-regression estimate is highly efficient, but this choice of ρ yields a poor breakdown point. If, on the other hand, we choose ρ to obtain a high breakdown point, the asymptotic efficiency of $\tilde{\boldsymbol{\beta}}_n$ decreases notably.

Note that S-regression estimates are a special type of M-regression estimates with a bounded loss function ρ and a special scale estimate. Because of the monotonicity of ρ and the inequality $\hat{\sigma}_n\leq\hat{\sigma}_n(\boldsymbol{\beta})$ for any $\boldsymbol{\beta}\in\mathbb{R}^p$ it is easy to see that

$$\frac{1}{n}\sum_{i=1}^{n}\rho\left(\frac{y_i-\tilde{\boldsymbol{\beta}}_n{}'\mathbf{x}_i}{\hat{\sigma}_n}\right)\leq\frac{1}{n}\sum_{i=1}^{n}\rho\left(\frac{y_i-\boldsymbol{\beta}'\mathbf{x}_i}{\hat{\sigma}_n}\right)\qquad\forall\,\boldsymbol{\beta}\in\mathbb{R}^p\,, \tag{4.6}$$

and hence $\tilde{\boldsymbol{\beta}}_n$ minimizes

$$\frac{1}{n}\sum_{i=1}^{n}\rho\left(\frac{y_i-\boldsymbol{\beta}'\mathbf{x}_i}{\hat{\sigma}_n}\right)\,,$$

where ρ is bounded (compare with (4.3)). In order to obtain simultaneously high breakdown point and high efficiency at the standard normal model, Yohai (1987) introduced the class of MM-regression estimates.

Definition 4.2 - MM-regression estimates *Let* $\rho_0 : \mathbb{R} \to \mathbb{R}^+$ *and* $\rho_1 : \mathbb{R} \to \mathbb{R}_+$ *be two functions satisfying conditions R.1 to R.4 above and such that* $\rho_1(u) \leq \rho_0(u)$ *for all* $u \in \mathbb{R}$ *and* $\sup_{u\in\mathbb{R}} \rho_1(u) = \sup_{u\in\mathbb{R}} \rho_0(u)$. *The MM-regression estimate is defined in the following three steps:*

- *let* $\tilde{\boldsymbol{\beta}}_n$ *be a high-breakdown point estimate for* $\boldsymbol{\beta}$*;*

- *let* $\hat{\sigma}_n$ *be the M-scale estimate of the residuals based on* $\tilde{\boldsymbol{\beta}}_n$. *That is,* $\hat{\sigma}_n$ *satisfies*

$$\frac{1}{n}\sum_{i=1}^{n} \rho_0\left((y_i - \mathbf{x}_i'\tilde{\boldsymbol{\beta}}_n)/\,\hat{\sigma}_n\right) = b\,;$$

- *the MM-estimate* $\hat{\boldsymbol{\beta}}_n$ *is defined as any solution of*

$$\sum_{i=1}^{n} \rho_1'\left((y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}_n)/\,\hat{\sigma}_n\right)\mathbf{x}_i = \mathbf{0}$$

with $S(\hat{\boldsymbol{\beta}}_n) \leq S(\tilde{\boldsymbol{\beta}}_n)$*, where*

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{n} \rho_1\left((y_i - \mathbf{x}_i'\boldsymbol{\beta})/\,\hat{\sigma}_n\right).$$

In particular, we will consider MM-estimates obtained with the steps described above when $\tilde{\boldsymbol{\beta}}_n$ is a S-regression estimate, and $\hat{\sigma}_n$ is the corresponding S-scale. Note that if ρ_0 and ρ_1 are continuously differentiable then the estimates $\hat{\boldsymbol{\beta}}_n$, $\tilde{\boldsymbol{\beta}}_n$ and $\hat{\sigma}_n$ satisfy the following equations:

$$\sum_{i=1}^{n} \rho_1'\left((y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}_n)/\,\hat{\sigma}_n\right)\mathbf{x}_i = \mathbf{0}, \tag{4.7}$$

$$\frac{1}{n}\sum_{i=1}^{n} \rho_0\left((y_i - \mathbf{x}_i'\tilde{\boldsymbol{\beta}}_n)/\,\hat{\sigma}_n\right) = b, \tag{4.8}$$

and

$$\sum_{i=1}^{n} \rho_0' \left((y_i - \mathbf{x}_i' \tilde{\boldsymbol{\beta}}_n) / \hat{\sigma}_n \right) \mathbf{x}_i = \mathbf{0}. \tag{4.9}$$

This class of regression estimates simultaneously achieves high breakdown point and high asymptotic efficiency (see Yohai, 1987). Yohai (1987) also proves that, if the data follow model (4.1) and the sequence $\tilde{\boldsymbol{\beta}}_n$ is consistent to the true parameter $\boldsymbol{\beta}_0$, then $\hat{\sigma}_n \to \sigma_0$ and $\hat{\boldsymbol{\beta}}_n$ is also strongly consistent to $\boldsymbol{\beta}_0$.

4.2 Robustness properties

Let F_0 be the distribution function of the errors ϵ and let G_0 be the distribution function of the explanatory variables $\mathbf{z}$ in $\mathbf{x}$ (see model (4.1)). Let $\mathcal{D}$ be a set of distribution functions in $\mathbb{R}^p$ where p is the number of random components in the vector of covariates $\mathbf{x}$. Let H_0 the distribution of the pair $(y, \mathbf{z})$.

We model the presence of outliers in the data in such a way that both the response variable and the covariates can be affected. In other words, the contamination might upset both F_0 and G_0. Consider the following ϵ-contamination neighbourhood of H_0

$$\mathcal{H}_\epsilon = \left\{ H \in \mathcal{D} \, : \, H = (1-\epsilon) \; H_0 + \epsilon \, H^* \right\}, \tag{4.10}$$

where H^* is arbitrary. Let ρ_0 and ρ_1 be real functions satisfying the regularity conditions of Definition 4.2. Let $b \in (0, 1/2]$. For each $\boldsymbol{\theta} \in \mathbb{R}^p$ define the functional

$\boldsymbol{\sigma}(H,\boldsymbol{\theta}) : \mathcal{D} \rightarrow \mathbb{R}_{+}$ by the following equation

$$E_H\left[\rho_0\left(\frac{Y-\boldsymbol{\theta}'\mathbf{X}}{\boldsymbol{\sigma}(H,\boldsymbol{\theta})}\right)\right] = b\,. \tag{4.11}$$

Let $\boldsymbol{\sigma} : \mathcal{D} \rightarrow \mathbb{R}_{+}$ be

$$\boldsymbol{\sigma}(H) = \inf_{\boldsymbol{\theta}\in\mathbb{R}^p} \boldsymbol{\sigma}(H,\boldsymbol{\theta})\ . \tag{4.12}$$

The associated S-regression functional $\tilde{\boldsymbol{\beta}} : \mathcal{D} \rightarrow \mathbb{R}^p$ is

$$\tilde{\boldsymbol{\beta}}(H) = \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^p} \boldsymbol{\sigma}(H,\boldsymbol{\theta})\ . \tag{4.13}$$

Finally, let the functional of the MM-regression estimate $\boldsymbol{\beta} : \mathcal{D} \rightarrow \mathbb{R}^p$ be defined by

$$\boldsymbol{\beta}(H) = \arg\inf_{\boldsymbol{\theta}\in\mathbb{R}^p} E_H\left[\rho_1\left(\frac{Y-\mathbf{X}'\boldsymbol{\theta}}{\boldsymbol{\sigma}(H)}\right)\right]\ . \tag{4.14}$$

Most asymptotic bias results for robust regression estimates have been established for the linear model without intercept, that is when $\mathbf{x}_i = \mathbf{z}_i$ in (4.1). When we have a linear regression model through the origin, the definition of asymptotic bias for the parameters $\boldsymbol{\beta}$ is as follows.

Definition 4.3 - **Maximum asymptotic bias for models without intercept** - *The maximum asymptotic bias of $\boldsymbol{\mu}$ over $\mathcal{H}_\epsilon$ is given by*

$$\mathbf{B}(\epsilon) = \sup_{H\in\mathcal{H}_\epsilon} \left\|(\boldsymbol{\beta}(H)-\boldsymbol{\beta}(H_0))'\ \mathbf{A}(G_0)\ (\boldsymbol{\beta}(H)-\boldsymbol{\beta}(H_0))\right\|, \tag{4.15}$$

where G_0 is the distribution of $\mathbf{x}$ in (4.1), $\mathbf{A}(G_0)$ is an equivariant dispersion estimate, and H_0 is the central distribution of the contamination neighbourhood $\mathcal{H}_\epsilon$.

For MM-regression estimates we have the following lower bound for $\mathbf{B}(\epsilon)$ (Berrendero Díaz, 1996; and Berrendero Díaz *et al.*, 1998):

$$\mathbf{B}(\epsilon) \geq \left(\frac{B^{+}_{\rho_0}(\epsilon)}{h^{-1}_{\rho_1}\left(h_{\rho_1}\left(B^{+}_{\rho_0}(\epsilon)\right)\right) + \epsilon/(1+\epsilon)} \right)^{2} - 1\,, \tag{4.16}$$

where

$$h_{\rho}(s) = \int_{0}^{\infty} \rho(y/s)\; dF_0(y)\,,$$

and

$$B^{+}_{\rho}(\epsilon) = h^{-1}_{\rho}\left(\frac{b-\epsilon}{1-\epsilon}\right).$$

The lower bound in (4.16) is an equality for small values of ϵ (see Berrendero Díaz, 1996).

Definition 4.4 - Asymptotic breakdown point *Let* $\mathbf{B}(\epsilon)$ *be as in (4.15). The asymptotic breakdown point of* $\boldsymbol{\beta}$ *is*

$$\epsilon^{*} = \inf\left\{\epsilon\,:\, \mathbf{B}(\epsilon) = \infty\right\}. \tag{4.17}$$

Yohai (1987) shows that if $\tilde{\boldsymbol{\beta}}_n$ has breakdown point equal to 1/2, then the MM-regression estimate $\hat{\boldsymbol{\beta}}_n$ also has $\epsilon^{*} = 1/2$. Rousseeuw and Yohai (1984) show that the S-regression estimates have $\epsilon^{*} = 1/2$ and hence the MM-regression estimates obtained with an initial S-regression estimate inherit this property.

4.3 Asymptotic properties

Under the central model H_0 the sequence $\hat{\boldsymbol{\beta}}_n$ is consistent to the true $\boldsymbol{\beta}_0$ in (4.1) (see Yohai, 1987). We also have that $\sqrt{n}\,(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})$ is asymptotically normal with covariance matrix

$$\boldsymbol{\Sigma}\left(F_0, G_0, \boldsymbol{\beta}_0, \sigma_0\right) =$$
$$= \sigma_0^2 \left[E_{F_0}\, {\rho_1'}^2 \left(U/\sigma_0\right) \Big/ \left(E_{F_0} \rho_1'' \left(U/\sigma_0\right)\right)^2 \right] \left[E_{G_0} \mathbf{X}\, \mathbf{X}' \right]^{-1}, \quad (4.18)$$

where $U = Y - \boldsymbol{\beta}_0' \mathbf{X}$ and $\sigma_0 = \boldsymbol{\sigma}\left(H_0\right)$ (see Yohai, 1987).

The asymptotic properties of these estimates when $H \in \mathcal{H}_\epsilon$ and $H \neq H_0$ are very difficult to study. In the next sections we obtain asymptotic results that hold for arbitrary distributions H in $\mathcal{H}_\epsilon$. In particular, we show that with some additional regularity conditions the sequences $\tilde{\boldsymbol{\beta}}_n$, $\hat{\sigma}_n$ and $\hat{\boldsymbol{\beta}}_n$ are consistent, and that $\hat{\boldsymbol{\beta}}_n$ is asymptotically normal.

4.3.1 Consistency of the S-scale estimate

In this section we will show that the S-scale estimates (4.5) for the linear regression model are strongly consistent to their asymptotic value (4.12). We will need the following regularity conditions on the function ρ and the explanatory variables

R.5 $\sup_u \left|\rho'\left(u\right)\right| < \infty$;

X.1 $P\left(\boldsymbol{\theta}' \mathbf{X} = 0 \right) = 0$ for all $\boldsymbol{\theta} \in \mathbb{R}^p$.

Remark 4.1 It is not difficult to see that if ρ belongs to Tukey's family (2.8) then it satisfies R.1-R.5.

To simplify the notation, for each $\boldsymbol{\theta} \in \mathbb{R}^p$ and $s > 0$ let

$$g(\boldsymbol{\theta}, s) = E\rho\left(\frac{Y - \boldsymbol{\theta}'\mathbf{X}}{s}\right). \tag{4.19}$$

Recall that the S-scale estimate $\hat{\sigma}_n$ satisfies $\hat{\sigma}_n = \inf_{\boldsymbol{\theta}} \hat{\sigma}_n(\boldsymbol{\theta})$, where $\hat{\sigma}_n(\boldsymbol{\theta})$ solves (4.4). Denote by Ω the underlying probability space, and let $\omega \in \Omega$ be an arbitrary event. In the statement of Theorem 4.1 below we add the argument ω to $\hat{\sigma}_n(\boldsymbol{\theta})$ to explicitly indicate that it is a random variable.

The following theorem is the main result in this section. Let $H \in \mathcal{H}_\epsilon$ be an arbitrary but fixed distribution. To simplify the notation, drop the argument H from $\boldsymbol{\sigma}(\boldsymbol{\theta}, H)$ (see (4.11)).

Theorem 4.1 *Let ρ be a real function satisfying conditions R.1-R.5 and let $\mathbf{X}$ be a random vector in $\mathbb{R}^p$ that satisfies X.1 above. Then*

i) for any $\epsilon > 0$ there exists $K_2 > 0$ such that $\boldsymbol{\sigma}(\boldsymbol{\theta}) - \epsilon \leq \hat{\sigma}_n(\boldsymbol{\theta}, \omega) \leq \boldsymbol{\sigma}(\boldsymbol{\theta}) + \epsilon$, a.s. uniformly in $\{\boldsymbol{\theta} : \|\boldsymbol{\theta}\| \leq K_2\}$. That is, there exists a null set $\mathcal{M}$ such that for any $\epsilon > 0$ and $\omega \notin \mathcal{M}$, there exists $n_0 = n_0(\epsilon, \omega)$ such that for any $n \geq n_0$

$$\boldsymbol{\sigma}(\boldsymbol{\theta}) - \epsilon \leq \hat{\sigma}_n(\boldsymbol{\theta}, \omega) \leq \boldsymbol{\sigma}(\boldsymbol{\theta}) + \epsilon \qquad \forall\, \boldsymbol{\theta} \in \{\boldsymbol{\theta} : \|\boldsymbol{\theta}\| \leq K_2\},$$

where n_0 does not depend on $\boldsymbol{\theta}$;

ii) $\hat{\sigma}_n \to \boldsymbol{\sigma}$ almost surely.

The following lemmas are needed for the proof of Theorem 4.1.

Unless explicitly stated otherwise, we will assume that the function ρ and the vector $\mathbf{X}$ satisfy conditions R.1-R.5 and X.1 above. For a function $f : \mathbb{R}^k \to \mathbb{R}$ let

$$\int_A f(\mathbf{u})\, dP(\mathbf{u}) \,,$$

denote the integral of f over $A \subseteq \mathbb{R}^k$ with respect to the measure P.

Lemma 4.1 *The function $g(\boldsymbol{\theta}, s)$ defined in (4.19) is continuous in $\boldsymbol{\theta}$ uniformly on $s \in (\eta, \infty)$ for any $\eta > 0$.*

Proof: Fix $\tilde{\epsilon} > 0$. Choose a bounded set $\mathcal{K} \subset \mathbb{R}^{p+1}$ such that $P_H\left[(Y, \mathbf{X}) \in \mathcal{K}\right] < \tilde{\epsilon}/4$. Let $a_1 = \sup_{\mathbf{x} \in \mathcal{K}} \|\mathbf{x}\| < \infty$ and let $a_2 = \sup_{u \in \mathbb{R}} |\rho'(u)| < \infty$. Choose $\delta = \delta(\tilde{\epsilon})$ such that $0 < \delta < (\tilde{\epsilon}\,\eta)/(2\,a_1\,a_2)$, and $\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\| < \delta$. Then we have

$$\begin{aligned}
\left| g(\boldsymbol{\theta}_1, s) - g(\boldsymbol{\theta}_2, s) \right| &\leq \left| \int_{\mathcal{K}} \rho\left(\frac{Y - \boldsymbol{\theta}_1'\mathbf{X}}{s}\right) - \rho\left(\frac{Y - \boldsymbol{\theta}_2'\mathbf{X}}{s}\right) dH(Y, \mathbf{X}) \right| + \tilde{\epsilon}/2 \\
&= \left| \int_{\mathcal{K}} \rho'\left(\frac{Y - \tilde{\boldsymbol{\theta}}'\mathbf{X}}{s}\right) \frac{1}{s} (\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)'\,\mathbf{X}\, dH(Y, \mathbf{X}) \right| + \tilde{\epsilon}/2 \\
&\leq \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\| \frac{1}{\eta}\, a_1\, a_2 + \tilde{\epsilon}/2 \\
&< \tilde{\epsilon} \,,
\end{aligned}$$

for any $s \geq \eta$. ■

The next result shows that the scale functional $\boldsymbol{\sigma}(H, \boldsymbol{\theta})$ defined in (4.11) with $H \in \mathcal{H}_\epsilon$ and $\boldsymbol{\theta} \in \mathbb{R}^p$ remains bounded away from zero and infinity if $\boldsymbol{\theta}$ belongs to an arbitrary neighbourhood of $\mathbf{0} \in \mathbb{R}^p$.

Lemma 4.2 *Let $\boldsymbol{\sigma}(H, \boldsymbol{\theta})$ be as in (4.11). For each arbitrary $K > 0$ and $H \in \mathcal{H}_\epsilon$, there exist two constants $S_1 = S_1(H, K)$ and $S_2 = S_2(H, K)$ such that*

$$0 < S_1 \leq \boldsymbol{\sigma}(H, \boldsymbol{\theta}) \leq S_2 < \infty \qquad \forall \; \|\boldsymbol{\theta}\| \leq K \,, \tag{4.20}$$

where $\|\cdot\|$ is the Euclidean norm in $\mathbb{R}^p$.

Proof: Fix the distribution function H and the constant K. Consider the function $f(s) : \mathbb{R}_+ \to \mathbb{R}_+$ defined by

$$f(s) = \max_{\|\boldsymbol{\theta}\| \leq K} E_H \rho \left(\frac{Y - \boldsymbol{\theta}'\mathbf{X}}{s} \right) = \max_{\|\boldsymbol{\theta}\| \leq K} g(\boldsymbol{\theta}, s) \,.$$

where $g(\boldsymbol{\theta}, s)$ is defined in (4.19). Note that for any $\boldsymbol{\theta}$ and $s_1 \leq s_2$ we have

(i) $E_H \rho \left(\frac{Y - \boldsymbol{\theta}'\mathbf{X}}{s_2} \right) \leq E_H \rho \left(\frac{Y - \boldsymbol{\theta}'\mathbf{X}}{s_1} \right)$;

(ii) $\lim_{s \to \infty} E_H \rho \left(\frac{Y - \boldsymbol{\theta}'\mathbf{X}}{s} \right) = 0$.

We will now show that the above properties hold for $f(s)$ as well. That f is non-increasing follows immediately from the first inequality above. To simplify the notation let $\mathcal{K} = \{ \boldsymbol{\theta} \in \mathbb{R}^p : \|\boldsymbol{\theta}\| \leq K \}$. Let $\epsilon > 0$ be arbitrary and fix a sufficiently small $\eta > 0$. For each $\boldsymbol{\theta} \in \mathcal{K}$, let $s(\boldsymbol{\theta}) < \infty$ be such that

$$\left| g(\boldsymbol{\theta}, s) \right| < \epsilon/4, \qquad \forall \, s > \max(\eta, s(\boldsymbol{\theta})) \,.$$

Also for each $\boldsymbol{\theta} \in \mathcal{K}$ define the set

$$\mathcal{A}_{\boldsymbol{\theta}} = \left\{ \, \tilde{\boldsymbol{\theta}} \in \mathcal{K} : \left| g(\tilde{\boldsymbol{\theta}}, s) \right| < \epsilon/4, \quad \forall \, s > \max(\eta, s(\boldsymbol{\theta})) \, \right\} .$$

We will show that $\mathcal{A}_{\boldsymbol{\theta}}$ is open. Note that $\boldsymbol{\theta} \in \mathcal{A}_{\boldsymbol{\theta}}$ and hence is not empty. Let $\tilde{\boldsymbol{\theta}} \in \mathcal{A}_{\boldsymbol{\theta}}$. By the continuity of g, there exist a $\delta = \delta\left(\epsilon, \tilde{\boldsymbol{\theta}}, \eta\right)$ such that

$$\|\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}\| < \delta \;\Rightarrow\; \left|g\left(\hat{\boldsymbol{\theta}}, s\right) - g\left(\tilde{\boldsymbol{\theta}}, s\right)\right| < \epsilon/8 \qquad \forall\, s \geq \eta\,.$$

Hence, $\hat{\boldsymbol{\theta}} \in \mathcal{A}_{\boldsymbol{\theta}}$ whenever $\|\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}\| < \delta\left(\epsilon, \tilde{\boldsymbol{\theta}}, \eta\right)$, and ϵ is fixed throughout the argument. Hence $\mathcal{A}_{\boldsymbol{\theta}}$ is open. By a standard compactness argument it follows that there exists a finite collection $\boldsymbol{\theta}_1$, ..., $\boldsymbol{\theta}_k$ such that

$$\mathcal{K} \subseteq \bigcup_{j=1}^{k} \mathcal{A}_{\boldsymbol{\theta}_j}\,.$$

Take $s_0 > \max\left(\eta, s\left(\boldsymbol{\theta}_1\right), \ldots, s\left(\boldsymbol{\theta}_k\right)\right)$. Let $s > s_0$. It is easy to see that for any $\boldsymbol{\theta} \in \mathcal{K}$ we have $\left|g\left(\boldsymbol{\theta}, s\right)\right| < \epsilon/4$. Then,

$$f\left(s\right) = \max_{\boldsymbol{\theta} \in \mathcal{K}} g\left(\boldsymbol{\theta}, s\right) \leq \epsilon/4 < \epsilon \qquad \text{if} \quad s > s_0\,.$$

We can find $S_2 = S_2\left(H, K\right)$ such that $f\left(s\right) < b$ for $s > S_2$. Hence, $g\left(\boldsymbol{\theta}, s\right) < b$ for all $\boldsymbol{\theta} \in \mathcal{K}$. Hence $\boldsymbol{\sigma}\left(H, \boldsymbol{\theta}\right) \leq S_2$ for all $\boldsymbol{\theta} \in \mathcal{K}$.

The argument for the other inequality is simple. Note that for any fixed $\boldsymbol{\theta} \in \mathbb{R}^p$ we have $\lim_{s \to 0} g\left(\boldsymbol{\theta}, s\right) = 1$. Hence, $\underline{\lim}_{s \to 0} \max_{\boldsymbol{\theta} \in \mathcal{K}} g\left(\boldsymbol{\theta}, s\right) \geq 1$. The result follows by noting that $\overline{\lim}_{s \to 0} \max_{\boldsymbol{\theta} \in \mathcal{K}} g\left(\boldsymbol{\theta}, s\right) \leq 1$ because $\rho\left(u\right) \leq 1$. ■

The following lemma shows that the infimum in the definition (4.12) for $\boldsymbol{\sigma}$ can be taken inside a certain neighbourhood around $\mathbf{0} \in \mathbb{R}^p$.

Lemma 4.3 *There exists $K_1 > 0$ such that*

$$\boldsymbol{\sigma}\left(H, 0\right) < \boldsymbol{\sigma}\left(H, \boldsymbol{\theta}\right) \qquad \forall\ \|\boldsymbol{\theta}\| > K_1\,,$$

and hence

$$\boldsymbol{\sigma}(H) = \inf_{\boldsymbol{\theta}\in\mathbb{R}^p} \boldsymbol{\sigma}(H,\boldsymbol{\theta}) = \inf_{\|\boldsymbol{\theta}\|\le K_1} \boldsymbol{\sigma}(H,\boldsymbol{\theta}) \, .$$

Proof: It is enough to show that for large $\|\boldsymbol{\theta}\|$ we have $\boldsymbol{\sigma}(H,\mathbf{0}) < \boldsymbol{\sigma}(H,\boldsymbol{\theta})$. Note that by the Dominated Convergence Theorem $g(\boldsymbol{\theta},\boldsymbol{\sigma}(\mathbf{0})) \to 1$ when $\|\boldsymbol{\theta}\| \to \infty$. Hence, given $0 < \eta < 1 - b$ there exists K_1 such that $g(\boldsymbol{\theta},\boldsymbol{\sigma}(\mathbf{0})) > b + \eta$ for all $\boldsymbol{\theta}$ with $\|\boldsymbol{\theta}\| > K_1$. Hence, $\boldsymbol{\sigma}(\boldsymbol{\theta}) > \boldsymbol{\sigma}(\mathbf{0})$ for $\|\boldsymbol{\theta}\| > K_1$. ■

Remark 4.2 Note that we can consider any other neighbourhood around $\mathbf{0} \in \mathbb{R}^p$ larger than the one given by the previous lemma. Specifically, let K_1 be as in Lemma 4.3 and let $K_2 \ge K_1$. Then

$$\boldsymbol{\sigma} \le \inf_{\|\boldsymbol{\theta}\|\le K_2} \boldsymbol{\sigma}(\boldsymbol{\theta}) \le \inf_{\|\boldsymbol{\theta}\|\le K_1} \boldsymbol{\sigma}(\boldsymbol{\theta}) = \boldsymbol{\sigma} \, ,$$

and hence we have

$$\boldsymbol{\sigma} = \inf_{\|\boldsymbol{\theta}\|\le K_2} \boldsymbol{\sigma}(\boldsymbol{\theta}) \qquad \text{for any } K_2 \ge K_1 \, .$$

We now show that there exists a compact set where both $\boldsymbol{\sigma}$ and $\hat{\sigma}_n$ are attained almost surely. Recall that for each $\boldsymbol{\theta} \in \mathbb{R}^p$, $\hat{\sigma}_n(\boldsymbol{\theta})$ is defined in (4.4) by

$$\frac{1}{n}\sum_{i=1}^{n} \rho\left(\frac{y_i - \boldsymbol{\theta}'\mathbf{x}_i'}{\hat{\sigma}_n(\boldsymbol{\theta})}\right) = b \, ,$$

and that $\hat{\sigma}_n = \inf_{\boldsymbol{\theta}\in\mathbb{R}^p} \hat{\sigma}_n(\boldsymbol{\theta})$.

Lemma 4.4 *Let K_1 be as in Lemma 4.3. Then there exists $K_2 \ge K_1$ such that if*

$$A_n = \left\{ \inf_{\|\boldsymbol{\theta}\|>K_2} \hat{\sigma}_n(\boldsymbol{\theta}) \le \hat{\sigma}_n(\mathbf{0}) \right\} ,$$

then $P(A_n \ i.o.) = 0$.

Proof: We will first show that

$$\lim_{K\to\infty} E\left[\inf_{\|\boldsymbol{\theta}\|>K} \rho\left(\frac{Y-\boldsymbol{\theta}'\mathbf{X}}{s}\right) I\left(|Y|\leq K\right)\right] = 1\,. \tag{4.21}$$

It is easy to see that for any $y \in \mathbb{R}$ and $s \in \mathbb{R}_+$ we have

$$\lim_{M\to\infty} \rho\left(\frac{|y|-M}{s}\right) I\left(|y|\leq M\right) = 1\,. \tag{4.22}$$

By hypothesis and Lemma 7.4, for any $\epsilon > 0$ there exist $\alpha > 0$ and $\mathcal{C}_1, \ldots, \mathcal{C}_s$ such that $\bigcup_{i=1}^{s} \mathcal{C}_i \supset \{\, \boldsymbol{\theta} : \|\boldsymbol{\theta}\| = 1\,\}$, and $P\left[\inf_{\boldsymbol{\theta}\in\mathcal{C}_i} |\boldsymbol{\theta}'\mathbf{X}| \geq \alpha\right] \geq 1-\epsilon$, for $1 \leq i \leq s$. We have

$$\begin{aligned}
\inf_{\|\boldsymbol{\theta}\|>L} \rho\left(\frac{y-\boldsymbol{\theta}'\mathbf{x}}{s}\right) &\geq \inf_{\|\boldsymbol{\theta}\|>L} \rho\left(\frac{y-\boldsymbol{\theta}'\mathbf{x}}{s}\right) I\left[|\boldsymbol{\theta}'\mathbf{x}| \geq M\right] \\
&\geq \inf_{\|\boldsymbol{\theta}\|>L} \rho\left(\frac{|y|-M}{s}\right) I\left[|\tilde{\boldsymbol{\theta}}'\mathbf{x}| \geq M/\left\|\boldsymbol{\theta}\right\|\right] I\left[|Y| \leq K\right]\,,
\end{aligned}$$

where $\tilde{\boldsymbol{\theta}} = \boldsymbol{\theta}/\left\|\boldsymbol{\theta}\right\|$. Set $L = M/\alpha$ to obtain

$$\begin{aligned}
\inf_{\|\boldsymbol{\theta}\|>L} \rho\left(\frac{y-\boldsymbol{\theta}'\mathbf{x}}{s}\right) &\geq \inf_{\|\boldsymbol{\theta}\|=1} \rho\left(\frac{|y|-M}{s}\right) I\left[|\boldsymbol{\theta}'\mathbf{x}| \geq \alpha\right] I\left[|Y| \leq K\right] \\
&\geq \min_{1\leq j\leq s} \inf_{\boldsymbol{\theta}\in\mathcal{C}_j} \rho\left(\frac{|y|-M}{s}\right) I\left[|\boldsymbol{\theta}'\mathbf{x}| \geq \alpha\right] I\left[|Y| \leq K\right] \\
&\geq (1-\delta) \min_{1\leq j\leq s} \inf_{\boldsymbol{\theta}\in\mathcal{C}_j} I\left[|\boldsymbol{\theta}'\mathbf{x}| \geq \alpha\right]\,,
\end{aligned}$$

by (4.22) for M large enough. Hence

$$\begin{aligned}
E\left[\inf_{\|\boldsymbol{\theta}\|>L} \rho\left(\frac{Y-\boldsymbol{\theta}'\mathbf{X}}{s}\right) I\left(|Y|\leq K\right)\right] &\geq (1-\delta)\, E\left[\min_{1\leq j\leq s} \inf_{\boldsymbol{\theta}\in\mathcal{C}_j} I\left[|\boldsymbol{\theta}'\mathbf{x}| \geq \alpha\right] I\left[|Y| \leq K\right]\right] \\
&= (1-\delta)\, P\left[\inf_{\boldsymbol{\theta}\in\mathcal{C}_j} |\boldsymbol{\theta}'\mathbf{x}| \geq \alpha\,,\ \forall\, 1 \leq j \leq s\right] \\
&\geq (1-\delta)\ (1-s\,\epsilon)\ ,
\end{aligned}$$

which can be made as close to 1 as desired. Hence (4.21) holds.

We now show that for an arbitrary $\boldsymbol{\theta} \in \mathbb{R}^p$, $\delta > 0$ we have

$$P_H\Big(\big|\hat{\sigma}_n(\boldsymbol{\theta}) - \boldsymbol{\sigma}(\boldsymbol{\theta})\big| > \delta\Big) \leq 2\exp(-n\gamma) \quad \text{for some } \gamma = \gamma(\delta) > 0\,. \tag{4.23}$$

Fix $\delta > 0$ and $\boldsymbol{\theta} \in \mathbb{R}^p$. We know that there exists $\delta_0 = \delta_0(\boldsymbol{\theta}, \delta) > 0$ such that

$$E\rho\left(\frac{Y - \boldsymbol{\theta}'\mathbf{X}}{\boldsymbol{\sigma}(\boldsymbol{\theta}) + \delta}\right) \geq b + \delta_0\,,$$

and

$$E\rho\left(\frac{Y - \boldsymbol{\theta}'\mathbf{X}}{\boldsymbol{\sigma}(\boldsymbol{\theta}) - \delta}\right) \leq b - \delta_0\,.$$

By Bernstein's Inequality (Lemma 7.6) we have

$$\begin{aligned}
P\Big(\hat{\sigma}_n(\boldsymbol{\theta}) - \boldsymbol{\sigma}(\boldsymbol{\theta}) < \delta\Big) = P\Big(\hat{\sigma}_n(\boldsymbol{\theta}) < \boldsymbol{\sigma}(\boldsymbol{\theta}) + \delta\Big) &\geq P\left(\frac{1}{n}\sum_{i=1}^n \rho\left(\frac{y_i - \boldsymbol{\theta}'\mathbf{x}_i}{\boldsymbol{\sigma}(\boldsymbol{\theta}) + \delta}\right) < b\right) \\
&\geq P\left(\left|\frac{1}{n}\sum_{i=1}^n \rho\left(\frac{y_i - \boldsymbol{\theta}'\mathbf{x}_i}{\boldsymbol{\sigma}(\boldsymbol{\theta}) + \delta}\right) - E\rho\left(\frac{Y - \boldsymbol{\theta}'\mathbf{X}}{\boldsymbol{\sigma}(\boldsymbol{\theta}) + \delta}\right)\right| < \delta_0/2\right) \\
&\geq 1 - \exp(-n\gamma_1)\,,
\end{aligned}$$

for some $\gamma_1 > 0$. Similarly we obtain

$$\begin{aligned}
P\Big(\hat{\sigma}_n(\boldsymbol{\theta}) - \boldsymbol{\sigma}(\boldsymbol{\theta}) > -\delta\Big) &= P\Big(\hat{\sigma}_n(\boldsymbol{\theta}) > \boldsymbol{\sigma}(\boldsymbol{\theta}) - \delta\Big) \\
&\geq P\left(\left|\frac{1}{n}\sum_{i=1}^n \rho\left(\frac{y_i - \boldsymbol{\theta}'\mathbf{x}_i}{\boldsymbol{\sigma}(\boldsymbol{\theta}) + \delta}\right) - E\rho\left(\frac{Y - \boldsymbol{\theta}'\mathbf{X}}{\boldsymbol{\sigma}(\boldsymbol{\theta}) + \delta}\right)\right| < \delta_0/2\right) \\
&\geq 1 - \exp(-n\gamma_2)\,,
\end{aligned}$$

where $\gamma_2 > 0$. Finally

$$\begin{aligned}
P\Big(\big|\hat{\sigma}_n(\boldsymbol{\theta}) - \boldsymbol{\sigma}(\boldsymbol{\theta})\big| > \delta\Big) &\leq P\Big(\hat{\sigma}_n(\boldsymbol{\theta}) - \boldsymbol{\sigma}(\boldsymbol{\theta}) > \delta\Big) + P\Big(\hat{\sigma}_n(\boldsymbol{\theta}) - \boldsymbol{\sigma}(\boldsymbol{\theta}) < -\delta\Big) \\
&\leq 2\exp(-n\min(\gamma_1, \gamma_2))\,.
\end{aligned}$$

Hence (4.23) holds with $\gamma = \min(\gamma_1, \gamma_2)$.

We will show that if

$$A_n = \left\{ \inf_{\|\boldsymbol{\theta}\|>K_2} \hat{\sigma}_n(\boldsymbol{\theta}) \leq \hat{\sigma}_n(\mathbf{0}) \right\},$$

then $P(A_n \text{ i.o.}) = 0$. Let $\delta_1 > 0$ be arbitrary. Choose $K_2 \geq K_1$ such that

$$E\left(\inf_{\|\boldsymbol{\theta}\|>K_2} \rho\left(\frac{Y - \boldsymbol{\theta}'\mathbf{X}}{\boldsymbol{\sigma}(\mathbf{0}) + \delta} \right) \right) \geq b + \delta_1 . \tag{4.24}$$

We have

$$\begin{aligned}
&P\left(\inf_{\|\boldsymbol{\theta}\|>K_2} \hat{\sigma}_n(\boldsymbol{\theta}) > \hat{\sigma}_n(\mathbf{0}) \right) \\
&\quad \geq P\left(\inf_{\|\boldsymbol{\theta}\|>K_2} \frac{1}{n}\sum_{i=1}^{n} \rho\left(\frac{y_i - \boldsymbol{\theta}'\mathbf{x}_i}{\boldsymbol{\sigma}(\mathbf{0}) + \delta} \right) > b, \left| \hat{\sigma}_n(\mathbf{0}) - \boldsymbol{\sigma}(\mathbf{0}) \right| < \delta \right) \\
&\quad \geq 1 - P\left(\frac{1}{n}\sum_{i=1}^{n} \inf_{\|\boldsymbol{\theta}\|>K_2} \rho\left(\frac{y_i - \boldsymbol{\theta}'\mathbf{x}_i}{\boldsymbol{\sigma}(\mathbf{0}) + \delta} \right) \leq b \right) - P\left(\left| \hat{\sigma}_n(\mathbf{0}) - \boldsymbol{\sigma}(\mathbf{0}) \right| \geq \delta \right) \\
&\quad \geq 1 - 2\, \exp(-n\,\gamma) - P\left(\frac{1}{n}\sum_{i=1}^{n} \inf_{\|\boldsymbol{\theta}\|>K_2} \rho\left(\frac{y_i - \boldsymbol{\theta}'\mathbf{x}_i}{\boldsymbol{\sigma}(\mathbf{0}) + \delta} \right) \leq b \right),
\end{aligned}$$

where $\gamma > 0$ is given by (4.23) (set $\boldsymbol{\theta} = \mathbf{0}$). Now note that by (4.24) and Bernstein's Inequality (Lemma 7.6) we have

$$\begin{aligned}
&P\left(\frac{1}{n}\sum_{i=1}^{n} \inf_{\|\boldsymbol{\theta}\|>K_2} \rho\left(\frac{y_i - \boldsymbol{\theta}'\mathbf{x}_i}{\boldsymbol{\sigma}(\mathbf{0}) + \delta} \right) \leq b \right) \\
&\qquad \leq P\left(\left| \frac{1}{n}\sum_{i=1}^{n} \inf_{\|\boldsymbol{\theta}\|>K_2} \rho\left(\frac{y_i - \boldsymbol{\theta}'\mathbf{x}_i}{\boldsymbol{\sigma}(\mathbf{0}) + \delta} \right) - E\left(\rho\left(\frac{Y - \boldsymbol{\theta}'\mathbf{X}}{\boldsymbol{\sigma}(\mathbf{0}) + \delta} \right) \right) \right| > \delta_1 \right) \\
&\qquad \leq \exp(-n\,\gamma'),
\end{aligned}$$

where $\gamma' > 0$. Hence $P(A_n) \leq 3\, \exp(-n\, \min(\gamma, \gamma'))$ and the result follows from the Borel-Cantelli Lemma (Lemma 7.3). ■

The following Lemma shows that $\boldsymbol{\sigma}(H, \boldsymbol{\theta})$ in (4.11) is continuous as a function of $\boldsymbol{\theta}$.

Lemma 4.5 *Assume that the conditions of Lemma 4.2 hold and that*

$$\frac{\partial}{\partial s} g(\boldsymbol{\theta}, s) < 0 \qquad \forall\ \boldsymbol{\theta} \in \mathbb{R}^p \ \ \forall\ s > 0\,.$$

Then $\boldsymbol{\sigma}(\boldsymbol{\theta})$ is continuous in $\boldsymbol{\theta}$ and hence uniformly continuous if $\boldsymbol{\theta} \in \mathcal{K}_\Theta$, where $\mathcal{K}_\Theta$ is an arbitrary compact set in $\mathbb{R}^p$.

Proof: We adapt an argument in Martin and Zamar (1993). Fix $\delta > 0$. Let S_1 and S_2 as in Lemma 4.2 and let K_1 as Lemma 4.3. Let $\mathcal{K}_s = [S_1, S_2]$ and $\mathcal{K}_\Theta = \{\|\boldsymbol{\theta}\| \leq K_2\}$. Let $\mathcal{B} = \mathcal{K}_s \times \mathcal{K}_\Theta$. Let δ_0 be given by

$$\delta_0 = \delta \min_{(\boldsymbol{\theta},s)\in\mathcal{B}} \left| \frac{\partial}{\partial s} g(\boldsymbol{\theta}, s) \right| > 0\,.$$

By Lemma 4.1 there exists $\gamma > 0$ such that

$$\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\| < \gamma \quad \Rightarrow \quad |g(\boldsymbol{\theta}_1, \boldsymbol{\sigma}(\boldsymbol{\theta}_2) - \delta) - g(\boldsymbol{\theta}_2, \boldsymbol{\sigma}(\boldsymbol{\theta}_2) - \delta)| < \delta_0/\,4\,.$$

Using the Mean Value Theorem we have

$$\begin{aligned} g(\boldsymbol{\theta}_1, \boldsymbol{\sigma}(\boldsymbol{\theta}_2) - \delta) - b &\geq g(\boldsymbol{\theta}_2, \boldsymbol{\sigma}(\boldsymbol{\theta}_2) - \delta) - b - \frac{\delta_0}{4} \\ &\geq \delta \min_{(\boldsymbol{\theta},s)\in\mathcal{B}} \left| \frac{\partial}{\partial s} g(\boldsymbol{\theta}, s) \right| - \frac{\delta_0}{4} > 0\,. \end{aligned}$$

Hence $\boldsymbol{\sigma}(\boldsymbol{\theta}_1) \geq \boldsymbol{\sigma}(\boldsymbol{\theta}_2) - \delta$. Similarly we have

$$\begin{aligned} g(\boldsymbol{\theta}_1, \boldsymbol{\sigma}(\boldsymbol{\theta}_2) + \delta) - b &\leq g(\boldsymbol{\theta}_2, \boldsymbol{\sigma}(\boldsymbol{\theta}_2) + \delta) - b + \frac{\delta_0}{4} \\ &\leq -\delta \min_{(\boldsymbol{\theta},s)\in\mathcal{B}} \left| \frac{\partial}{\partial s} g(\boldsymbol{\theta}, s) \right| + \frac{\delta_0}{4} < 0\,, \end{aligned}$$

so that $\boldsymbol{\sigma}(\boldsymbol{\theta}_1) \leq \boldsymbol{\sigma}(\boldsymbol{\theta}_2) + \delta$. ■

The following Lemma states that if $s \neq \boldsymbol{\sigma}(\boldsymbol{\theta})$ then $E\rho\left((Y - \boldsymbol{\theta}'\mathbf{X})/s\right)$ remains uniformly away from b when $\boldsymbol{\theta}$ belongs to an arbitrary compact set.

Lemma 4.6 *Let $\mathcal{K} \subset \mathbb{R}^p$ be an arbitrary compact set, let $\delta > 0$ be arbitrary and let b be as in the definition of the S-scale estimate. Then, there exists a positive constant $\epsilon = \epsilon(\mathcal{K}, \delta)$ such that*

$$E\rho\left(\frac{Y - \boldsymbol{\theta}'\mathbf{X}}{\boldsymbol{\sigma}(\boldsymbol{\theta}) + \delta}\right) \leq b - \epsilon \qquad \forall \boldsymbol{\theta} \in \mathcal{K},$$

and

$$E\rho\left(\frac{Y - \boldsymbol{\theta}'\mathbf{X}}{\boldsymbol{\sigma}(\boldsymbol{\theta}) - \delta}\right) \geq b + \epsilon \qquad \forall \boldsymbol{\theta} \in \mathcal{K}.$$

Proof: Follows easily from Lemma 4.2 and a Taylor expansion of first order, after noting that

$$\lambda_1 = \inf_{\boldsymbol{\theta} \in \mathcal{K}} E\left[\rho'\left(\frac{Y - \boldsymbol{\theta}'\mathbf{X}}{\boldsymbol{\sigma}(\boldsymbol{\theta}) + \delta}\right)\left(\frac{Y - \boldsymbol{\theta}'\mathbf{X}}{\boldsymbol{\sigma}(\boldsymbol{\theta}) + \delta}\right)\right] > 0,$$

and

$$\lambda_2 = \inf_{\boldsymbol{\theta} \in \mathcal{K}} E\left[\rho'\left(\frac{Y - \boldsymbol{\theta}'\mathbf{X}}{\boldsymbol{\sigma}(\boldsymbol{\theta}) - \delta}\right)\left(\frac{Y - \boldsymbol{\theta}'\mathbf{X}}{\boldsymbol{\sigma}(\boldsymbol{\theta}) - \delta}\right)\right] > 0.$$

■

Lemma 4.7 shows that the estimating equation is uniformly continuous in $\boldsymbol{\theta}$ and s bounded away from zero, for n sufficiently large, almost surely. When needed in the proof, we will explicitly indicate that $\hat{\sigma}_n$ or $\hat{\sigma}_n(\boldsymbol{\theta})$ are random variables by including the argument $\omega \in \Omega$, where Ω denotes the underlying probability space.

Lemma 4.7 *For each $\boldsymbol{\theta} \in \mathbb{R}^p$, $s > 0$ and $\omega \in \Omega$, let*

$$\bar{\rho}_n (\boldsymbol{\theta}, \omega, s) = \frac{1}{n} \sum_{i=1}^{n} \rho \left(\frac{y_i(\omega) - \boldsymbol{\theta}' \mathbf{x}_i(\omega)}{s} \right) . \tag{4.25}$$

and $\eta > 0$ be arbitrary. Then there exists a null set $\mathcal{N}$ such that for any $\epsilon > 0$ and $\omega \notin \mathcal{N}$ there exist $\delta = \delta(M, \epsilon, \eta) > 0$ and $n_0 = n_0(M, \omega)$ such that if $\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\| < \delta$ then

$$\left| \bar{\rho}_n(\boldsymbol{\theta}_1, \omega, s) - \bar{\rho}_n(\boldsymbol{\theta}_2, \omega, s) \right| < \epsilon \qquad \forall\, n > n_0(M, \omega) \qquad \forall\, s \geq \eta .$$

Proof: Fix $\tilde{\epsilon} > 0$. Let $\mathcal{K} \subset \mathbb{R}^{p+1}$ be a compact set such that $P_H \left[(Y, \mathbf{X}) \notin \mathcal{K} \right] < \tilde{\epsilon}/8$. Take away a null set such that for every remaining ω the strong law of large numbers assures the existence of $n_0(\omega)$ such that

$$\frac{1}{n} \sum_{i=1}^{n} I(\mathbf{x}_i(\omega) \notin \mathcal{K}) < \tilde{\epsilon}/4 \qquad \forall n > n_0(\omega) .$$

Let $a_1 = \sup_t \rho'(t) < \infty$ and let $a_2 = \sup_{\mathbf{x} \in \mathcal{K}} \|\mathbf{x}\| < \infty$. Using the Mean Value Theorem and R.3 we have

$$\begin{aligned}
&\left| \bar{\rho}_n(\boldsymbol{\theta}_1, \omega) - \bar{\rho}_n(\boldsymbol{\theta}_2, \omega) \right| \\
&\leq \frac{1}{n} \sum_{i=1}^{n} \left| \rho\left(\frac{y_i(\omega) - \boldsymbol{\theta}_1 \mathbf{x}_i(\omega)}{s} \right) - \rho\left(\frac{y_i(\omega) - \boldsymbol{\theta}_2 \mathbf{x}_i(\omega)}{s} \right) \right| I(\mathbf{x}_i \in \mathcal{K}) + \frac{2}{n} \sum_{i=1}^{n} I(\mathbf{x}_i \notin \mathcal{K}) \\
&\leq \frac{1}{n} \sum_{i=1}^{n} \left| \rho'\left(\frac{y_i(\omega) - \tilde{\boldsymbol{\theta}}' \mathbf{x}_i(\omega)}{s} \right) \frac{1}{s} \mathbf{x}_i(\omega)' (\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2) \right| I(\mathbf{x}_i(\omega) \in \mathcal{K}) + \tilde{\epsilon}/4 \\
&\leq \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\| a_2 \frac{1}{\eta} a_1 + \tilde{\epsilon}/4 .
\end{aligned}$$

Hence, for almost all ω

$$\left| \bar{\rho}_n(\boldsymbol{\theta}_1, \omega) - \bar{\rho}_n(\boldsymbol{\theta}_2, \omega) \right| < \tilde{\epsilon} ,$$

if $\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|$ is sufficiently small, and $n \geq n_0(\tilde{\epsilon}, \omega)$. ■

We now show that $\bar{\rho}_n(\boldsymbol{\theta}, \omega, s)$ defined in (4.25) and $E\rho((Y - \boldsymbol{\theta}'\mathbf{X})/s)$ are uniformly close if $\boldsymbol{\theta}$ belongs to an arbitrary compact set, for large n, almost surely.

Lemma 4.8 *Let $\mathcal{K} \subset \mathbb{R}^p$ be an arbitrary compact set, let ρ, $\bar{\rho}_n$ be as in Lemma 4.7, and let $\eta > 0$ be arbitrary. There exists a null set $\mathcal{N}$ such that for any $\omega \notin \mathcal{N}$ and $\epsilon > 0$ there exists $n_0 = n_0(\omega, \epsilon)$ such that*

$$\sup_{\boldsymbol{\theta} \in \mathcal{K}} \left| \bar{\rho}_n(\boldsymbol{\theta}, \omega, s) - E\rho(\boldsymbol{\theta}, s) \right| < \epsilon \qquad \forall\, n > n_0(\omega, \epsilon) \qquad \forall\, s \geq \eta\,.$$

Proof: Consider the same null $\mathcal{N}$ as in the proof of Lemma 4.7. By Lemma 4.1 the function

$$g(\boldsymbol{\theta}, s) = E\rho\left(\frac{Y - \boldsymbol{\theta}'\mathbf{X}}{s}\right)$$

is continuous in $\boldsymbol{\theta}$, uniformly on $s \geq \eta$. By Lemma 4.7 there exists $\delta > 0$ such that if $\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\| < \delta$ then

$$\left|\bar{\rho}_n(\boldsymbol{\theta}_1, \omega, s) - \bar{\rho}_n(\boldsymbol{\theta}_2, \omega, s)\right| < \epsilon \qquad \forall\, n > n_0(\omega, \epsilon) \qquad \forall\, s \geq \eta\,.$$

Construct a finite collection of open balls $B(\boldsymbol{\theta}_j, \delta)$ of radius δ such that they cover $\mathcal{K}$ and such that

$$\left| E\rho\left(\frac{Y - \boldsymbol{\theta}_j'\mathbf{X}}{s}\right) - E\rho\left(\frac{Y - \boldsymbol{\theta}'\mathbf{X}}{s}\right) \right| < \epsilon/3 \qquad \forall\, s \geq \eta\,,$$

for $\boldsymbol{\theta} \in B(\boldsymbol{\theta}_j, \delta)$. Such a collection exists by Lemma 4.1 and a standard compactness argument. Let $\boldsymbol{\theta} \in \mathcal{K}$ be arbitrary. Then $\boldsymbol{\theta} \in B(\boldsymbol{\theta}_j, \delta)$ for a particular ball. We have

$$\begin{aligned}
&\left|\bar{\rho}_n(\boldsymbol{\theta}, \omega, s) - E\rho\left(\frac{Y - \boldsymbol{\theta}'\mathbf{X}}{s}\right)\right| \\
&\qquad \leq \left|\bar{\rho}_n(\boldsymbol{\theta}, \omega, s) - \bar{\rho}_n(\boldsymbol{\theta}_j, \omega, s)\right| + \left|\bar{\rho}_n(\boldsymbol{\theta}_j, \omega, s) - E\rho\left(\frac{Y - \boldsymbol{\theta}_j'\mathbf{X}}{s}\right)\right| \\
&\qquad\qquad + \left|E\rho\left(\frac{Y - \boldsymbol{\theta}_j'\mathbf{X}}{s}\right) - E\rho\left(\frac{Y - \boldsymbol{\theta}'\mathbf{X}}{s}\right)\right| < \epsilon, \qquad (4.26)
\end{aligned}$$

if n is large enough, not depending on the particular $\boldsymbol{\theta} \in \mathbb{R}^p$. ■

We now prove the main result of this section.

Proof of Theorem 4.1. **(i)**: Let S_1 as in Lemma 4.2. Fix $\delta > 0$ such that $S_1 - \delta > 0$. By the previous lemmas there exist $\epsilon > 0$ such that $E\rho((Y - \boldsymbol{\theta}'\mathbf{X})/(\boldsymbol{\sigma}(\boldsymbol{\theta}) + \delta)) \leq b - \epsilon$ for $\boldsymbol{\theta} \in \mathcal{K}$. Also $|\bar{\rho}_n(\boldsymbol{\theta}, s) - E\rho((Y - \boldsymbol{\theta}'\mathbf{X})/s)| < \epsilon/2$ for all $s \geq S_1 - \delta$ and $\boldsymbol{\theta} \in \mathcal{K}$, for n sufficiently large, almost surely. Note that for any $\boldsymbol{\theta} \in \mathcal{K}$ we have $\boldsymbol{\sigma}(\boldsymbol{\theta}) + \delta \geq S_1 - \delta$ and then

$$\begin{aligned}
\bar{\rho}_n(\boldsymbol{\theta}, \boldsymbol{\sigma}(\boldsymbol{\theta}) + \delta) &< E\rho((Y - \boldsymbol{\theta}'\mathbf{X})/(\boldsymbol{\sigma}(\boldsymbol{\theta}) + \delta)) + \epsilon/2 \\
&< b - \epsilon/2,
\end{aligned}$$

so that $\hat{\sigma}_n(\boldsymbol{\theta}) \leq \boldsymbol{\sigma}(\boldsymbol{\theta}) + \delta$ for n large enough, almost surely. Similarly we obtain $\hat{\sigma}_n(\boldsymbol{\theta}) \geq \boldsymbol{\sigma}(\boldsymbol{\theta}) - \delta$ for n large enough, almost surely.

(ii): This follows from (i) and Lemma 7.16. ■

4.3.2 Consistency of the S- and MM-regression estimates

In this section we show that under certain regularity conditions the S- and MM-regression estimates are consistent. Note that because both estimators minimize a loss measure, the following Theorem applies to both.

Theorem 4.2 *Assume that $\rho : \mathbb{R} \to \mathbb{R}_+$ and $\mathbf{X} \in \mathbb{R}^p$ satisfy conditions R.1 to R.5 and X.1 above. Let $g(\boldsymbol{\theta}, s)$ be as in (4.19). Assume that $\hat{\sigma}_n \to \boldsymbol{\sigma}$ almost surely, and let $\tilde{\boldsymbol{\beta}}_n$ be defined by*

$$\tilde{\boldsymbol{\beta}}_n = \arg \inf_{\boldsymbol{\theta} \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \rho\left(\frac{y_i - \boldsymbol{\theta}' \mathbf{x}_i}{\hat{\sigma}_n}\right) .$$

Let $\tilde{\boldsymbol{\beta}}$ be

$$\tilde{\boldsymbol{\beta}} = \arg \inf_{\boldsymbol{\theta} \in \mathbb{R}^p} E\rho\left(\frac{Y - \boldsymbol{\theta}'\mathbf{X}}{\boldsymbol{\sigma}}\right) = \arg \inf_{\boldsymbol{\theta} \in \mathbb{R}^p} g(\boldsymbol{\theta}, \boldsymbol{\sigma}) .$$

If $g(\boldsymbol{\theta}, \boldsymbol{\sigma})$ has a unique minimum as a function of $\boldsymbol{\theta} \in \mathbb{R}^p$, then $\tilde{\boldsymbol{\beta}}_n \to \tilde{\boldsymbol{\beta}}$ a.s.

Proof: We will first show that there exists $L > 0$ such that

$$\overline{\lim_{n\to\infty}} \, \|\tilde{\boldsymbol{\beta}}_n\| \leq L \qquad \text{a.s.}$$

Because $\hat{\sigma}_n \to \boldsymbol{\sigma}$ almost surely, it is enough to show that for any $\sigma > 0$ there exists L and $\eta = \eta(L) > 0$ such that

$$\lim_{n\to\infty} \inf_{\|\boldsymbol{\theta}\| > L} \frac{1}{n} \sum_{i=1}^{n} \rho\left(\frac{y_i - \boldsymbol{\theta}'\mathbf{x}_i}{\sigma}\right) \geq b + \eta \quad \text{a.s.} \tag{4.27}$$

The Dominated Convergence Theorem shows that for any $\sigma > 0$,

$$\lim_{M\to\infty} E\rho\left(\frac{|Y| - M}{\sigma}\right) = 1 . \tag{4.28}$$

By Lemma 7.4 and because $P\left(\boldsymbol{\theta}'\mathbf{X} = 0\right) = 0$ for all $\boldsymbol{\theta} \in \mathbb{R}^p$ by hypothesis, there exist $\alpha > 0$, $\gamma > 0$ and a finite collection of compact sets $\mathcal{C}_1, \ldots, \mathcal{C}_s$ in R^p such that

$$\bigcup_{j=1}^{s} \mathcal{C}_j \supset \left\{\boldsymbol{\theta} \in \mathbb{R}^p : \|\boldsymbol{\theta}\| = 1\right\} ,$$

and

$$P\left(\inf_{\boldsymbol{\theta} \in \mathcal{C}_j} |\boldsymbol{\theta}'\mathbf{X}| \geq \alpha\right) \geq b + \gamma . \tag{4.29}$$

By (4.28) we can find M and $\eta > 0$ such that

$$(b+\gamma)\; E\rho\left(\frac{|Y| - M}{\sigma}\right) P\Big(|Y| \leq M\Big) \geq b + \eta . \tag{4.30}$$

Now we have

$$\begin{aligned}
\inf_{\|\boldsymbol{\theta}\|>L} \frac{1}{n}\sum_{i=1}^{n} \rho\left(\frac{y_i - \boldsymbol{\theta}'\mathbf{x}_i}{\sigma}\right) &\geq \inf_{\|\boldsymbol{\theta}\|>L} \frac{1}{n}\sum_{i=1}^{n} \rho\left(\frac{y_i - \boldsymbol{\theta}'\mathbf{x}_i}{\sigma}\right) I\left(|\boldsymbol{\theta}'\mathbf{x}_i| \geq M\right) I\left(|y_i| \leq M\right) \\
&\geq \inf_{\|\boldsymbol{\theta}\|>L} \frac{1}{n}\sum_{i=1}^{n} \rho\left(\frac{|y_i| - |\boldsymbol{\theta}'\mathbf{x}_i|}{\sigma}\right) I\left(|\boldsymbol{\theta}'\mathbf{x}_i| \geq M\right) I\left(|y_i| \leq M\right) ,
\end{aligned}$$

because $|y_i - \boldsymbol{\theta}'\mathbf{x}_i| \geq |y_i| - |\boldsymbol{\theta}'\mathbf{x}_i|$. Hence

$$\begin{aligned}
\inf_{\|\boldsymbol{\theta}\|>L} \frac{1}{n}\sum_{i=1}^{n} \rho\left(\frac{y_i - \boldsymbol{\theta}'\mathbf{x}_i}{\sigma}\right) &\geq \inf_{\|\boldsymbol{\theta}\|>L} \frac{1}{n}\sum_{i=1}^{n} \rho\left(\frac{|y_i| - M}{\sigma}\right) I\left(|\boldsymbol{\theta}'\mathbf{x}_i| \geq M\right) I\left(|y_i| \leq M\right) \\
&\geq \inf_{\|\boldsymbol{\theta}\|>L} \frac{1}{n}\sum_{i=1}^{n} \rho\left(\frac{|y_i| - M}{\sigma}\right) I\left(|\tilde{\boldsymbol{\theta}}'\mathbf{x}_i| \geq M/\left\|\boldsymbol{\theta}\right\|\right) I\left(|y_i| \leq M\right) ,
\end{aligned}$$

where $\tilde{\boldsymbol{\theta}} = \boldsymbol{\theta}/\left\|\boldsymbol{\theta}\right\|$. Also note that if $L > M/\alpha$ then $\{|\tilde{\boldsymbol{\theta}}'\mathbf{x}_i| \geq M/\left\|\boldsymbol{\theta}\right\|\} \supseteq \{|\tilde{\boldsymbol{\theta}}'\mathbf{x}_i| \geq \alpha\}$, so that

$$\begin{aligned}
\inf_{\|\boldsymbol{\theta}\|>L} \frac{1}{n}\sum_{i=1}^{n} \rho\left(\frac{y_i - \boldsymbol{\theta}'\mathbf{x}_i}{\sigma}\right) &\geq \inf_{\|\boldsymbol{\theta}\|=1} \frac{1}{n}\sum_{i=1}^{n} \rho\left(\frac{|y_i| - M}{\sigma}\right) I\left(|\boldsymbol{\theta}'\mathbf{x}_i| \geq \alpha\right) \\
&\geq \min_{1\leq j\leq s} \inf_{\boldsymbol{\theta}\in\mathcal{C}_j} \frac{1}{n}\sum_{i=1}^{n} \rho\left(\frac{|y_i| - M}{\sigma}\right) I\left(|\boldsymbol{\theta}'\mathbf{x}_i| \geq \alpha\right) \\
&\geq \min_{1\leq j\leq s} \frac{1}{n}\sum_{i=1}^{n} \inf_{\boldsymbol{\theta}\in\mathcal{C}_j} \rho\left(\frac{|y_i| - M}{\sigma}\right) I\left(|\boldsymbol{\theta}'\mathbf{x}_i| \geq \alpha\right) .
\end{aligned}$$

Equation (4.27) now follows from (4.29), (4.30) and the Strong Law of Large Numbers.

We now show that $\tilde{\boldsymbol{\beta}}_n \to \tilde{\boldsymbol{\beta}}$ almost surely. Consider an arbitrary open neighbourhood $B\left(\tilde{\boldsymbol{\beta}}\right)$ of $\tilde{\boldsymbol{\beta}}$ and the compact set

$$\mathcal{K} = \left\{\boldsymbol{\theta} : \|\boldsymbol{\theta}\| \leq L\right\} \cap B\left(\tilde{\boldsymbol{\beta}}\right)^c .$$

By Lemma 7.5

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \rho\left(\frac{y_i - \tilde{\boldsymbol{\beta}}'\mathbf{x}_i}{\hat{\sigma}_n}\right) = E\left(\rho\left(\frac{Y - \tilde{\boldsymbol{\beta}}'\mathbf{X}}{\boldsymbol{\sigma}}\right)\right) \quad \text{a.s.}.$$

So, it is enough to show that there exists $\gamma > 0$ and $\boldsymbol{\sigma}_1 > \boldsymbol{\sigma}$ such that

$$\varliminf_{n\to\infty} \inf_{\boldsymbol{\theta}\in\mathcal{K}} \frac{1}{n} \sum_{i=1}^{n} \rho\left(\frac{y_i - \boldsymbol{\theta}'\mathbf{x}_i}{\boldsymbol{\sigma}_1}\right) \geq E\left(\rho\left(\frac{Y - \tilde{\boldsymbol{\beta}}'\mathbf{X}}{\boldsymbol{\sigma}}\right)\right) + \gamma \quad \text{a.s.} \tag{4.31}$$

The assumption on uniqueness of the minimum of the function $g\left(\boldsymbol{\theta}, \boldsymbol{\sigma}\right)$, the Dominated Convergence Theorem and a standard compactness argument yield $\boldsymbol{\sigma}_1 > \boldsymbol{\sigma}$, $\gamma > 0$ and a finite family of sets $\mathcal{C}_1, \ldots, \mathcal{C}_s$ such that

$$\bigcap_{i=1}^{s} \mathcal{C}_i \supset \mathcal{K} = \left\{\boldsymbol{\theta} : \|\boldsymbol{\theta}\| \leq L\right\} \cap B\left(\tilde{\boldsymbol{\beta}}\right)^c ,$$

and

$$E\left[\inf_{\boldsymbol{\theta}\in\mathcal{C}_i} \rho\left(\frac{Y - \boldsymbol{\theta}'\mathbf{X}}{\boldsymbol{\sigma}_1}\right)\right] \geq E\left(\rho\left(\frac{Y - \tilde{\boldsymbol{\beta}}'\mathbf{X}}{\boldsymbol{\sigma}}\right)\right) + \gamma \qquad 1 \leq i \leq s .$$

Hence

$$\varliminf_{n\to\infty} \inf_{\boldsymbol{\theta}\in\mathcal{K}} \frac{1}{n} \sum_{i=1}^{n} \rho\left(\frac{y_i - \boldsymbol{\theta}'\mathbf{x}_i}{\boldsymbol{\sigma}_1}\right) \geq \min_{1\leq j\leq s} \varliminf_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \inf_{\boldsymbol{\theta}\in\mathcal{C}_j} \rho\left(\frac{y_i - \boldsymbol{\theta}'\mathbf{x}_i}{\boldsymbol{\sigma}_1}\right) ,$$

and the result follows from the Strong Law of Large Numbers. ■

4.3.3 Asymptotic distribution of the MM-regression estimate

The asymptotic distribution of these estimates for distributions belonging to (4.10) can be derived along the same lines we used in Section 2.3.6. In general the scale estimate $\hat{\sigma}_n$ is not asymptotically independent of $\hat{\boldsymbol{\beta}}_n$. Nevertheless, provided the sequences $\tilde{\boldsymbol{\beta}}_n$, $\hat{\boldsymbol{\beta}}_n$ and $\hat{\sigma}_n$ are consistent, if $\hat{\sigma}_n$ is a S-scale, we can find the asymptotic distribution of $\sqrt{n}\,(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})$. The proof of the following theorem is based on the same techniques used in the proof of Theorem 2.6 and is omitted here.

Theorem 4.3 - **Asymptotic normality** - *Assume that* $\hat{\boldsymbol{\beta}}_n \xrightarrow{P} \boldsymbol{\beta}\,(H)$, $\tilde{\boldsymbol{\beta}}_n \xrightarrow{P} \tilde{\boldsymbol{\beta}}\,(H)$, *and* $\hat{\sigma}_n \xrightarrow{P} \boldsymbol{\sigma}\,(H)$. *Let* $F \in \mathcal{H}_\epsilon$ *and let* ρ_0 *and* ρ_1 *be as in Definition 4.2. Assume that* ρ_0 *and* ρ_1 *have third derivatives that are continuous and the following expected values exist:*

$$E_H\left[\rho_0'\left(\frac{y-\boldsymbol{\beta}'\mathbf{X}}{\boldsymbol{\sigma}}\right)\left(\frac{y-\boldsymbol{\beta}'\mathbf{X}}{\boldsymbol{\sigma}}\right)\right],\quad E_H\left[\rho_0''\left(\frac{y-\boldsymbol{\beta}'\mathbf{X}}{\boldsymbol{\sigma}}\right)\mathbf{X}\,\mathbf{X}'\right],$$
$$E_H\left[\rho_0''\left(\frac{y-\boldsymbol{\beta}'\mathbf{X}}{\boldsymbol{\sigma}}\right)(y-\boldsymbol{\beta}'\mathbf{X})\;\mathbf{X}\right],\quad E_H\left[\rho_1''\left(\frac{y-\boldsymbol{\beta}'\mathbf{X}}{\boldsymbol{\sigma}}\right)\mathbf{X}\,\mathbf{X}'\right],$$
$$E_H\left[\rho_1''\left(\frac{y-\boldsymbol{\beta}'\mathbf{X}}{\boldsymbol{\sigma}}\right)\left(\frac{y-\boldsymbol{\beta}'\mathbf{X}}{\boldsymbol{\sigma}}\right)\mathbf{X}\right],\quad E_H\left[\rho_1'''\left(\frac{y-\boldsymbol{\beta}'\mathbf{X}}{\boldsymbol{\sigma}}\right)\|\mathbf{X}\|^2\;\mathbf{X}\right],$$

and

$$E_H\left[\rho_1'''\left(\frac{y-\boldsymbol{\beta}'\mathbf{X}}{\boldsymbol{\sigma}}\right)(y-\boldsymbol{\beta}'\mathbf{X})\;\mathbf{X}\,\mathbf{X}'\right].$$

Then

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}\,(H)\right) \xrightarrow[n\to\infty]{W} \mathcal{N}_p\Big(\mathbf{0},\boldsymbol{\Sigma}\Big). \tag{4.32}$$

The asymptotic variance-covariance matrix is given by the following. Let

$$\mathbf{A} = E_H\left[\rho_1''\left(\frac{y-\boldsymbol{\beta}'\mathbf{X}}{\boldsymbol{\sigma}}\right)\mathbf{X}\,\mathbf{X}'\right]^{-1}\boldsymbol{\sigma},$$

and

$$\mathbf{b} = \mathbf{A}\,\boldsymbol{\sigma} \frac{E_H\left[\rho_1''\left(\frac{y-\boldsymbol{\beta}'\mathbf{X}}{\boldsymbol{\sigma}}\right)\left(\frac{y-\boldsymbol{\beta}'\mathbf{X}}{\boldsymbol{\sigma}}\right)\mathbf{X}\right]}{E_H\left[\rho_0'\left(\frac{y-\boldsymbol{\beta}'\mathbf{X}}{\boldsymbol{\sigma}}\right)\left(\frac{y-\boldsymbol{\beta}'\mathbf{X}}{\boldsymbol{\sigma}}\right)\right]},$$

then

$$\begin{aligned}\boldsymbol{\Sigma} = \mathbf{A}\,E_H\left[{\rho_1'}^2\left(\frac{y-\boldsymbol{\beta}'\mathbf{X}}{\boldsymbol{\sigma}}\right)\mathbf{X}\,\mathbf{X}'\right]\mathbf{A}' + E_H\left[\left(\rho_0\left(\frac{y-\boldsymbol{\beta}'\mathbf{X}}{\boldsymbol{\sigma}}\right) - b\right)^2\right]\mathbf{b}\,\mathbf{b}' \\ + 2\,\mathbf{A}E_H\left[\rho_1'\left(\frac{y-\boldsymbol{\beta}'\mathbf{X}}{\boldsymbol{\sigma}}\right)\rho_0\left(\frac{y-\boldsymbol{\beta}'\mathbf{X}}{\boldsymbol{\sigma}}\right)\mathbf{X}\right]\mathbf{b}' \quad (4.33)\end{aligned}$$

Note that the form of $\boldsymbol{\Sigma}$ is very involved, and hence the empirical estimate obtained by replacing H with $\hat{H}_n$ will be numerically unstable. This will have a negative impact on the quality of statistical inferences which are based on it.

Chapter 5

Robust bootstrap for the linear regression model

In this chapter we extend the results of Chapter 3 to the linear regression model. We consider the problem of statistical inference based on robust regression estimates and describe the implementation of the robust bootstrap for this model. We illustrate its robustness properties with two real data sets. We show that under regularity conditions the robust bootstrap distribution estimate is consistent for the asymptotic distribution of the statistic of interest. We also study the breakdown point of the resulting quantile estimates and show that they improve upon the classical bootstrap quantile estimates. Finally, we compare the finite sample coverage and mean length of confidence intervals built with the empirical asymptotic variance estimate and the robust bootstrap.

There are several results in the literature concerning the application of the bootstrap principle to this model (Efron, 1979; Freedman, 1981; Wu, 1986; Efron and Tibshirani, 1993). Moreover, the bootstrapping of robust regression estimates has also been studied by Shorack (1982).

As in the location-scale model, two difficulties arise when bootstrapping robust regression estimates. The first is related to the computational complexity of robust regression estimates. The second is a consequence of the presence of outliers in the data.

Consider MM-regression estimates $\hat{\boldsymbol{\beta}}_n$ calculated with an initial S-estimate $\tilde{\boldsymbol{\beta}}_n$ and S-scale $\hat{\sigma}_n$ (see Definitions 4.2 and 4.1 on pages 143 and 141 respectively). These estimates have desirable robustness properties but are not easy to calculate. When we bootstrap these estimates, for each bootstrap sample we have to solve a non-convex minimization problem in $\mathbb{R}^p$ to determine $\tilde{\boldsymbol{\beta}}_n$ and the scale estimate $\hat{\sigma}_n$. Then we have to find a local extreme of another non-convex function in $\mathbb{R}^p$ to determine $\hat{\boldsymbol{\beta}}_n$. The number of bootstrap samples needed to obtain reliable distribution estimates naturally grows with the dimension of the statistic and hence makes the problem computationally even more expensive to solve. In the context of data-mining and other applications with extremely large data sets (both in the number of cases and the number of covariates) straightforward re-calculation of such robust estimates is rarely a feasible option.

The presence of outliers in the data can have an unduly large effect on the final distribution estimate. As described in Chapter 3 the reason lies in the re-sampling scheme: in many bootstrap samples the proportion of outliers can be significantly higher than in the original data set. This may in turn produce extreme re-calculated estimates and affect the bootstrap distribution estimate.

We will show that with our method, each re-calculation only involves solving a linear system of equations. Hence there is a very important gain in speed, and consequently in feasibility. We also obtain more stable and robust quantile estimates than the classical bootstrap method. To quantify this property we extend Singh's concept of quantile breakdown point (Singh, 1998) to the linear regression model.

To illustrate the magnitude of the gain in speed obtained with our method consider this simple example. We generated an artificial data set following model (4.1) with $\boldsymbol{\beta} = \mathbf{0}$ with $n = 50$ and $p = 5$. We applied both the classical and robust bootstrap to an MM-regression estimate, with 10,000 bootstrap samples. The classical bootstrap took 2735 CPU seconds while our method used less than 7 CPU seconds. The computations were done with C code designed by the author and run on a dual-CPU Sun Sparc Ultra 4 (each CPU a 296 Megahertz SUNW UltraSPARC-II) with 1.1 Gigabytes of RAM memory and using SunOS 5.7.

The rest of this chapter is organized as follows. Section 5.1 presents the method and the notation. Section 5.2 contains two examples that illustrate the use of our method. Sections 5.3 and 5.4 discuss its asymptotic and robustness properties respectively. Finally, Section 5.5 contains the results of a Monte Carlo study on the coverage and average mean of confidence intervals obtained with MM-regression estimates and different methods of estimating their distribution.

5.1 Definitions

Let $(y_1, \mathbf{x}_1), \ldots, (y_n, \mathbf{x}_n)$ be a random sample following model (4.1). We consider MM-regression estimates with initial S-regression estimates and S-scales (see Definitions 4.2 and 4.1). To simplify the notation let $\tilde{\boldsymbol{\beta}}_n$ be the initial S-regression estimate, $\hat{\sigma}_n$ the associated S-scale estimate and let $\hat{\boldsymbol{\beta}}_n$ be the final MM-regression estimate. We will consider the case of random explanatory variables in detail, but briefly discuss the fixed design case in Remark 5.3 below.

As discussed in Chapter 3, we are interested in making statistical inferences about the regression parameter $\boldsymbol{\beta}$. We can use the result of Theorem 4.3 on the asymptotic normality of the sequence $\sqrt{n}\,(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})$. To use this method we only have to estimate the asymptotic variance of $\hat{\boldsymbol{\beta}}_n$. The second option is to directly estimate the distribution function of $\sqrt{n}\,(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})$. We can then use this distribution estimate to approximate the quantiles needed to construct confidence intervals.

We propose to use the following computer intensive method to generate a large number of re-calculated $\hat{\boldsymbol{\beta}}_n^*$'s. These re-computed statistics can be used to estimate both the asymptotic covariance matrix and the distribution function of the statistics $\hat{\boldsymbol{\beta}}_n$. For the first objective we can use the empirical covariance matrix of the re-calculated $\hat{\boldsymbol{\beta}}_n^*$'s. To estimate the distribution function F_n of $\hat{\boldsymbol{\beta}}_n$ we use the empirical distribution function of the re-computed statistics. In what follows we will focus on this last problem.

Theorem 4.3 shows that the asymptotic behaviour of the sequence $\hat{\boldsymbol{\beta}}_n$ depends on that of $\hat{\sigma}_n$. Hence, to obtain an estimate of the distribution of $\hat{\boldsymbol{\beta}}_n$ we take into account the behaviour of the scale estimate $\hat{\sigma}_n$.

For each pair $(y_i, \mathbf{x}_i)$ in the sample define the residuals associated with $\hat{\boldsymbol{\beta}}_n$ and $\tilde{\boldsymbol{\beta}}_n$: $r_i = y_i - \hat{\boldsymbol{\beta}}_n' \mathbf{x}_i$ and $\tilde{r}_i = y_i - \tilde{\boldsymbol{\beta}}_n' \mathbf{x}_i$. First note that $\hat{\boldsymbol{\beta}}_n$ and $\hat{\sigma}_n$ can be represented as a weighted least squares fit. Similarly to Chapter 3, define the weights ω_i and v_i as

$$\begin{aligned} \omega_i &= \rho_1' \left(r_i / \hat{\sigma}_n \right) / \, r_i \quad 1 \leq i \leq n, \\ v_i &= \frac{\hat{\sigma}_n}{n \, b} \, \rho_0 \left(\tilde{r}_i / \hat{\sigma}_n \right) / \, \tilde{r}_i \quad 1 \leq i \leq n. \end{aligned} \tag{5.1}$$

Simple computations yield the following weighted average representation of equations (4.7) and (4.8):

$$\hat{\boldsymbol{\beta}}_n = \left[\sum_{i=1}^n \omega_i \ \mathbf{x}_i \ \mathbf{x}_i' \right]^{-1} \ \sum_{i=1}^n \omega_i \ \mathbf{x}_i \ y_i \, , \tag{5.2}$$

$$\hat{\sigma}_n = \sum_{i=1}^n v_i \ \left(y_i - \tilde{\boldsymbol{\beta}}_n' \mathbf{x}_i \right) . \tag{5.3}$$

Let $(y_i^*, \mathbf{x}_i^*)$, $i = 1, \ldots, n$ be a bootstrap sample from the observations. Define the random variables $\hat{\boldsymbol{\beta}}_n^*$ and $\hat{\sigma}_n^*$ by

$$\hat{\boldsymbol{\beta}}_n^* = \left[\sum_{i=1}^n \omega_i^* \ \mathbf{x}_i^* \ \mathbf{x}_i^{*\prime} \right]^{-1} \ \sum_{i=1}^n \omega_i^* \ \mathbf{x}_i^* \ y_i^* \, , \tag{5.4}$$

$$\hat{\sigma}_n^* = \sum_{i=1}^n v_i^* \left(y_i^* - \tilde{\boldsymbol{\beta}}_n' \mathbf{x}_i^* \right) \tag{5.5}$$

where $\omega_i^* = \rho_1' \left(r_i^* / \hat{\sigma}_n \right) / \, r_i^*$, $v_i^* = \hat{\sigma}_n \, \rho_0 \left(\tilde{r}_i^* / \hat{\sigma}_n \right) / \left(n \, b \, \tilde{r}_i^* \right)$, $r_i^* = y_i^* - \hat{\boldsymbol{\beta}}_n' \mathbf{x}_i^*$, and $\tilde{r}_i^* = y_i^* - \tilde{\boldsymbol{\beta}}_n' \mathbf{x}_i^*$ for $1 \leq i \leq n$. Note that $\hat{\boldsymbol{\beta}}_n$, $\hat{\sigma}_n$ and $\tilde{\boldsymbol{\beta}}_n$ are not re-calculated from each bootstrap sample $(y_i^*, \mathbf{x}_i^*)$, $i = 1, \ldots, n$.

We now apply a linear correction to the estimates obtained in (5.4) and (5.5) and combine them. Intuitively the correction is needed to account for the loss in variability due to the fixed weights. Let

$$\mathbf{M}_n = \hat{\sigma}_n \left[\sum_{i=1}^{n} \rho_1'' \left(r_i / \hat{\sigma}_n, \mathbf{x}_i \right) \mathbf{x}_i \, \mathbf{x}_i' \right]^{-1} \sum_{i=1}^{n} \omega_i \, \mathbf{x}_i \, \mathbf{x}_i' \,, \tag{5.6}$$

$$\mathbf{d}_n = a_n^{-1} \left[\sum_{i=1}^{n} \rho_1'' \left(r_i / \hat{\sigma}_n, \mathbf{x}_i \right) \mathbf{x}_i \mathbf{x}_i' \right]^{-1} \sum_{i=1}^{n} \rho_1'' \left(r_i / \hat{\sigma}_n, \mathbf{x}_i \right) r_i \, \mathbf{x}_i \,, \tag{5.7}$$

$$a_n = \hat{\sigma}_n^2 \, \frac{1}{n} \frac{1}{b} \sum_{i=1}^{n} \left[\rho_0' \left(\tilde{r}_i / \hat{\sigma}_n \right) \tilde{r}_i / \hat{\sigma}_n \right] \,. \tag{5.8}$$

The robust bootstrap $\hat{\boldsymbol{\beta}}_n^{R*} - \hat{\boldsymbol{\beta}}_n$ is given by

$$\hat{\boldsymbol{\beta}}_n^{R*} - \hat{\boldsymbol{\beta}}_n = \mathbf{M}_n \left(\hat{\boldsymbol{\beta}}_n^* - \hat{\boldsymbol{\beta}}_n \right) + \mathbf{d}_n \left(\hat{\sigma}_n^* - \hat{\sigma}_n \right),$$

Remark 5.1 - *Computational Ease*: Note that to recalculate $\hat{\boldsymbol{\beta}}_n^{R*}$ we do not solve (4.8) and (4.7). For each bootstrap sample we only solve the linear system of equations (5.4) and calculate the weighted average (5.5). The correction factors $\mathbf{M}_n$, $\mathbf{d}_n$ and a_n arise from two linear systems and a weighted average respectively and are computed only once with the full sample.

Remark 5.2 - *Robustness*: For MM-regression estimates $\hat{\boldsymbol{\beta}}_n$ with a re-descending score function ρ_1' (i.e., $\rho_1'(r) \equiv 0$ for $|r| \geq c > 0$), the weights ω_i give the method stability in the presence of outliers. Outlying points will be associated with small weights in equations (5.2) and (5.3). Note that extreme outliers (those with an associated residual $|r_i| > c\,\hat{\sigma}_n$) will receive a zero weight, and hence will have no effect at all on the recalculated coefficients. Note that the weights v_i^* used in recalculating the scale are also decreasing in the absolute value of the residuals. This also makes the outlying points less influential in the recalculated $\hat{\sigma}_n^*$.

Remark 5.3 - *Fixed design*: In the case of a linear regression model with fixed design we propose to adapt our method as follows. The main difference lies in the re-sampling procedure, so that it best resembles the randomness of the model (Freedman, 1981). Let $e_j = y_j - \hat{\boldsymbol{\beta}}_n' \mathbf{x}_j$, $1 \leq j \leq n$ be the residuals of the MM-estimate. The bootstrapped y_i^*'s are

$$y_i^* = \hat{\boldsymbol{\beta}}_n' \mathbf{x}_i + e_i^*$$

where e_i^*, $1 \leq i \leq n$ is a random sample from the residuals. Now $\hat{\boldsymbol{\beta}}_n^*$ and $\hat{\sigma}_n^*$ are defined by

$$\hat{\boldsymbol{\beta}}_n^* = \left[\sum_{i=1}^{n} \omega_i^* \; \mathbf{x}_i \; \mathbf{x}_i' \right]^{-1} \left[\sum_{i=1}^{n} \omega_i^* \; \mathbf{x}_i \; y_i^* \right], \tag{5.9}$$

$$\hat{\sigma}_n^* = \sum_{i=1}^{n} v_i^* \, (y_i^* - \tilde{\boldsymbol{\beta}}_n' \mathbf{x}_i) \tag{5.10}$$

where $\omega_i^* = \rho_1' \, (r_i^* / \hat{\sigma}_n) / r_i^*$, $v_i^* = \hat{\sigma}_n \, \rho_0 \, (\tilde{r}_i^* / \hat{\sigma}_n) / (n \; b \; \tilde{r}_i^*)$, $r_i^* = y_i^* - \hat{\boldsymbol{\beta}}_n' \mathbf{x}_i$, and $\tilde{r}_i^* = y_i^* - \tilde{\boldsymbol{\beta}}_n' \mathbf{x}_i$ for $1 \leq i \leq n$. The correction factors $\mathbf{M}_n$, $\mathbf{d}_n$ and a_n are defined as before, and so is $\hat{\boldsymbol{\beta}}_n^{R*} - \hat{\boldsymbol{\beta}}_n$. See Chapter 6 for a discussion on future work regarding this case.

5.2 Examples

5.2.1 Body and Brain Weights

Consider the brain and body weight of 28 animals published in Rousseeuw and Leroy (1987, page 57). The model considered in the literature is

$$\log(\text{Brain weight (g)}) = \alpha_0 + \beta_0 \log(\text{Body weight (g)}) + \epsilon \, ;$$

where $\alpha_0 \in \mathbb{R}$ and $\beta_0 \in \mathbb{R}$ are the parameters of interest and ϵ are independent and identically distributed errors with mean zero and constant variance σ^2. We used an MM-regression estimate obtained with $\psi = \rho'_{4.685}$ in Tukey's family. The S-scale was obtained with $\rho_{1.54764}$ also in Tukey's family. This choice yields an estimate with simultaneous 50% breakdown point and 95% efficiency if the data are normally distributed. Figure 5.1 contains a scatter plot of the transformed data with the least squares and MM-regression fits. The question of interest is whether larger brains are required to govern heavier bodies. In particular, the magnitude of the slope is relevant: a slope larger than 1 would indicate that the required brain weight increases faster than the body weight. On the other hand, a slope smaller than 1 would indicate the opposite. We are interested on a confidence interval for the slope of this model.

We obtained 10,000 bootstrap samples to re-calculate the estimates $(\hat{\alpha}_n, \hat{\beta}_n)$. We used these re-computed estimates to approximate the distribution of the vector of parameters

$$\sqrt{n}\ \left((\hat{\alpha}_n, \hat{\beta}_n)' - (\alpha_0, \beta_0)' \right) , \tag{5.11}$$

in order to perform statistical inference on the parameters of interest (α_0, β_0). We used both the classical and robust bootstrap and concentrated on inferences for the slope β_0. Note that given the empirical distribution in $\mathbb{R}^2$ of the re-computed $(\hat{\alpha}_n^*, \hat{\beta}_n^*)$ we can use its projection on the β-axis to obtain an estimate of the distribution of the β projection of (5.11).

With the classical bootstrap, the 99% confidence interval for β_0 is $(0.66, 1.49)$. The p-value for the null hypothesis $\beta_0 \leq 1$ is between 0.01 and 0.05. The robust bootstrap yields the following 99% confidence interval for β_0: $(0.67, 0.84)$. The p-value for the same null hypothesis is $p < 0.0001$.

The reason for this difference lies with the tails of the bootstrap distribution of the regression parameters. Three observations in this data set correspond to dinosaurs and do not follow the same pattern as the other observations. A certain proportion of the bootstrap samples may contain enough outlying observations to breakdown the estimate. These samples can yield extreme values for the estimate that produce unduly large quantiles. The robust point estimate and the robust bootstrap re-calculated estimates down-weight these three observations and hence are less sensitive to them.

In Figure 5.2 we show scatter plots of $(\hat{\boldsymbol{\beta}}_u^* - \hat{\boldsymbol{\beta}}_n)$ for $1 \leq u \leq 10,000$, where $\hat{\boldsymbol{\beta}}_u^*$ denotes either the classical or the robust bootstrap estimate for the u-th bootstrap sample. We see that the tails of the classical bootstrap estimate are highly influenced by the outliers present in the data. This causes the estimates for the 99.5% and 0.5% quantiles to be highly inaccurate. The quantile estimates based on the robust bootstrap do not have this problem.

5.2.2 Belgium International Phone Calls

Consider the Belgium International Calls data set (see Rousseeuw and Yohai, 1984). These data consist of the number of international phone calls (in tens of millions) originating in Belgium between 1950 and 1973. From 1964 to 1969 the observations

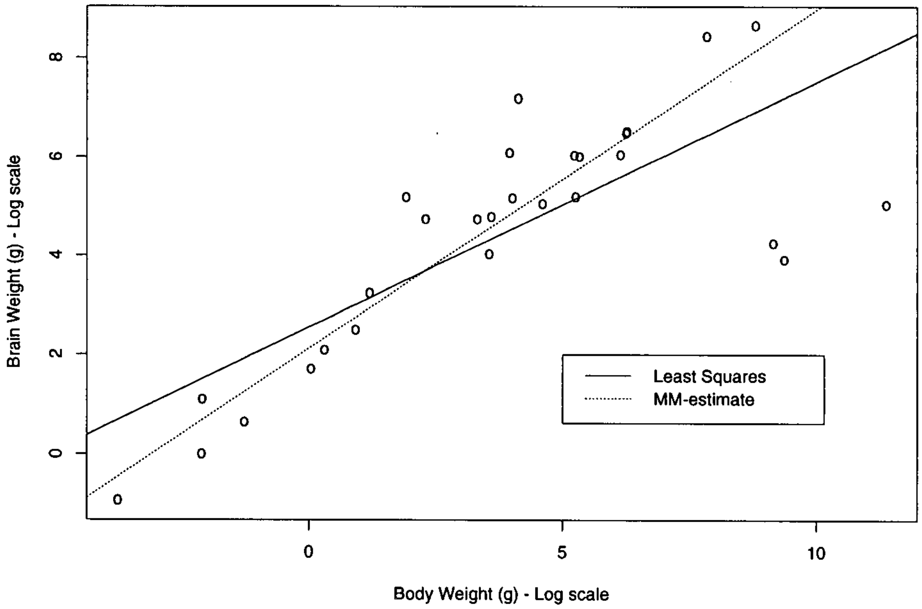
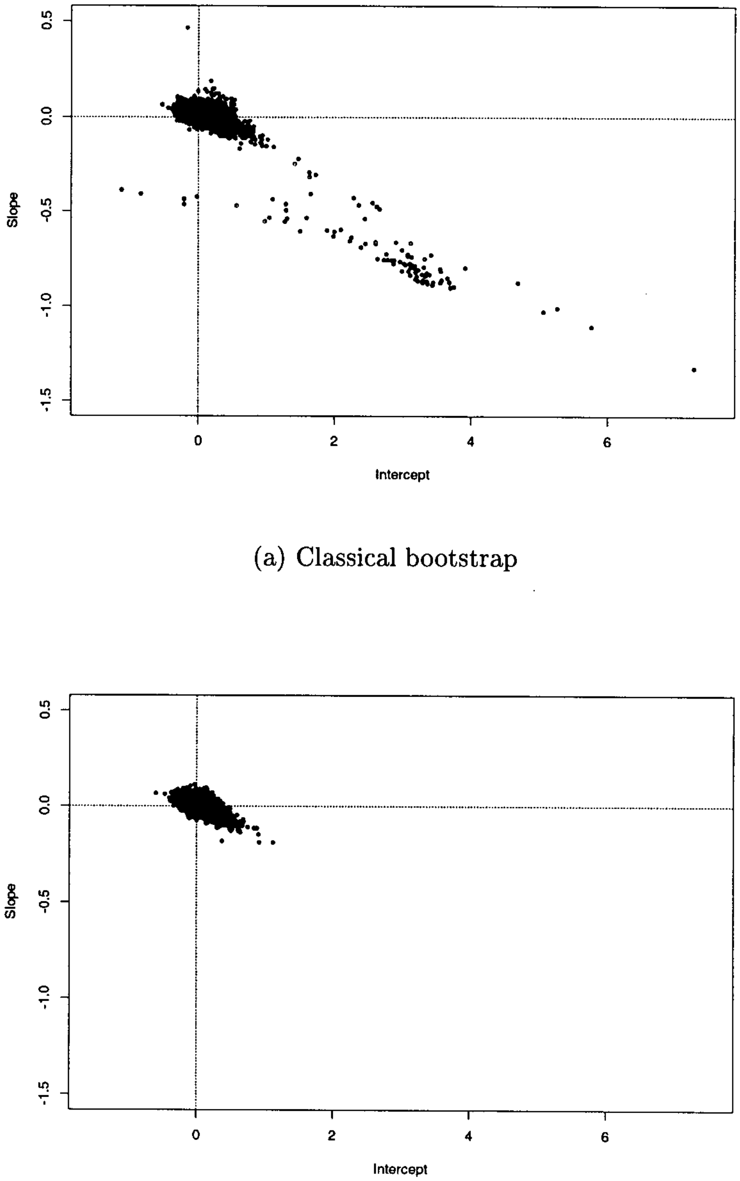


Figure 5.1: Least squares and robust regression fits to the Brain and Body Weight data



(a) Classical bootstrap

(b) Robust bootstrap

Figure 5.2: Classical and robust bootstrap distribution estimates for the Brain and Body Weight data - 10,000 bootstrap samples

Parameter	Method	95% Confidence Interval
α_0	Robust Bootstrap	$(-10.32, -3.20)$
	Classical Bootstrap	$(-17.74, 0.35)$
β_0	Robust Bootstrap	$(0.08, 0.20)$
	Classical Bootstrap	$(0.00, 0.28)$

Table 5.1: Belgium International Calls - Bootstrap and robust bootstrap 95% confidence intervals

were mistakenly recorded. Instead of the number of calls, their total duration in minutes was registered. The figure for 1970 is partly contaminated; some calls were recorded with their duration, others were registered according to the old convention. The linear regression model considered in the literature is

$$\#\,\text{Calls (in tens of millions)} = \alpha_0 + \beta_0\,\text{Year} + \epsilon\,, \tag{5.12}$$

where α_0 and β_0 are the parameters of interest, and the errors ϵ are assumed to be independent and identically distributed with mean zero and unknown but constant variance σ^2. The MM-regression estimate with an S-scale gives $\hat{\alpha}_0 = -5.23$ and $\hat{\beta}_0 = 0.11$. Figure 5.3 displays the data with the robust and least squares fits. To obtain confidence intervals for the regression parameters $\boldsymbol{\beta}$ we use the classical and robust bootstrap. We performed 10,000 bootstrap re-calculations. Scatter plots of $\hat{\boldsymbol{\beta}}_u^{R*} - \hat{\boldsymbol{\beta}}_n$ for the robust bootstrap and of $\hat{\boldsymbol{\beta}}_u^{*} - \hat{\boldsymbol{\beta}}_n$ for the classical bootstrap are presented in Figure 5.4. We clearly see that the robust bootstrap estimates are more stable. This is reflected in the length of the confidence intervals. Table 5.1 contains the lower and upper limits of 95% confidence intervals for the slope and intercept calculated with both the classical and robust bootstrap.

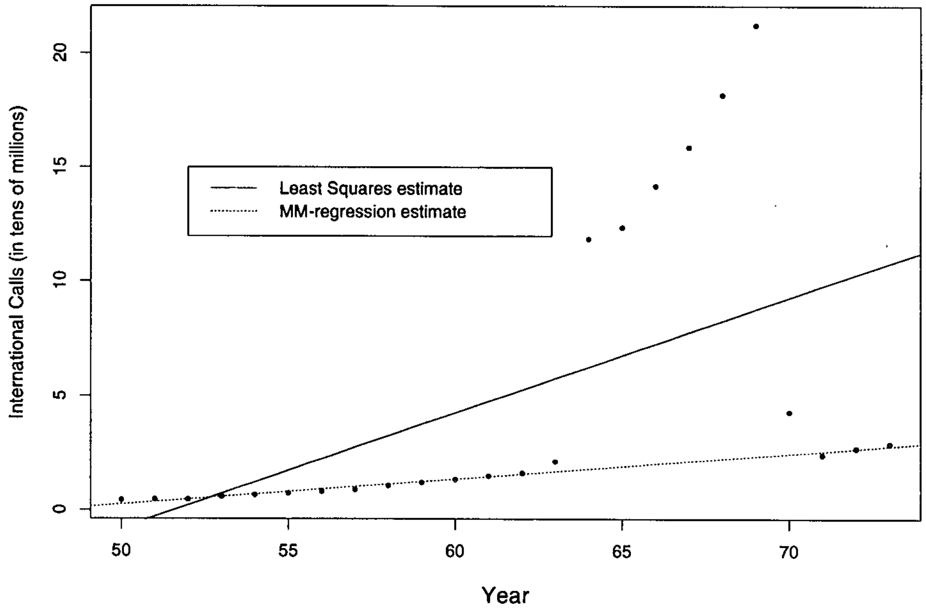
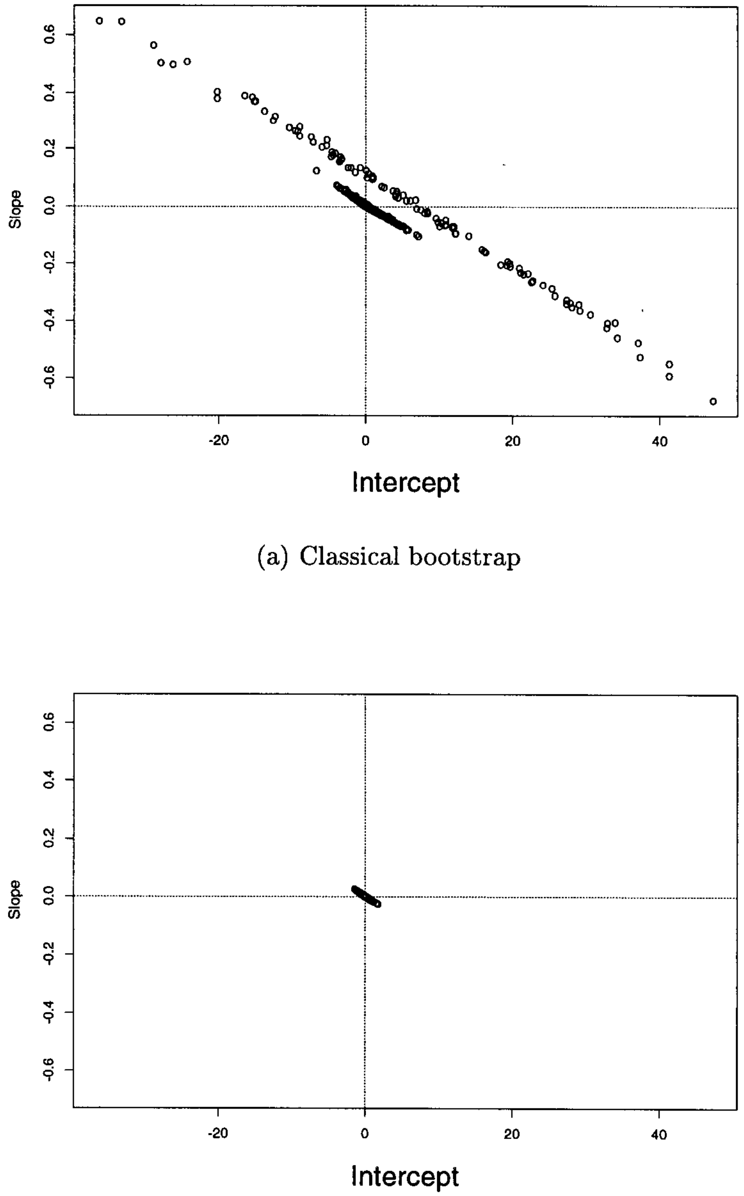


Figure 5.3: Least squares and robust regression fits to the Belgium International Phone Calls data

Note that when we estimate the variability of the robust estimates with the robust bootstrap we conclude that the estimates of the regression coefficients are significantly different from zero at the 5% level. The artificial variability introduced by the outliers in the classical bootstrap re-calculated $\hat{\beta}_u^*$'s inflates the standard deviation estimates. As a consequence, if we use these standard deviation estimates we conclude that, at the 5% level, there is no significant linear relationship between the response and the predictor variable. The conclusion obtained with the robust bootstrap analysis is intuitively in agreement with the linear trend observed in the scatter plot of the data (see Figure 5.3).



(a) Classical bootstrap

(b) Robust bootstrap

Figure 5.4: Comparison of classical and robust bootstrap distribution estimates for the Belgium International Phone Calls data - 10,000 bootstrap samples

5.3 Asymptotic properties

The following theorem shows that the asymptotic distribution of the robust bootstrap is the same as that of the MM-regression estimator.

Theorem 5.1 - Convergence of the robust bootstrap distribution - *Let ρ_0 and ρ_1 be real functions as in Definition 4.2. Assume that they have continuous third derivatives. Let $\hat{\boldsymbol{\beta}}_n$ be the MM-regression estimator, $\hat{\sigma}_n$ the S-scale and $\tilde{\boldsymbol{\beta}}_n$ the associated S-regression estimator. Assume that they are consistent, that is: $\hat{\boldsymbol{\beta}}_n \xrightarrow{P} \boldsymbol{\beta}$, $\hat{\sigma}_n \xrightarrow{P} \sigma$ and $\tilde{\boldsymbol{\beta}}_n \xrightarrow{P} \tilde{\boldsymbol{\beta}}$. If the following conditions hold:*

1. *the following matrices exist and are finite:*

$$E\left[\rho_1'(r)/r\ \mathbf{XX}'\right]^{-1}, \quad E\left[\rho_0'(r)/r\ \mathbf{XX}'\right]^{-1}, \quad E\left[\rho_1'(r)\mathbf{XX}'\right], \quad E\left[\rho_1'(r)\, r\ \mathbf{XX}'\right],$$
$$E\left[\rho_0''(r)\mathbf{XX}'\right], \quad E\left[\rho_1''(r)\mathbf{XX}'\right]^{-1}, \quad E\left[\rho_0''(r)\, r\mathbf{X}\right], E\left[\rho_1''(r)\, r\mathbf{X}\right];$$

2. *$E\left[\rho_0'(r)\, r\right] \neq 0$ and finite,*

3. *$\rho_0'(u)/u$, $\rho_1'(u)/u$, $(\rho_0'(u) - \rho_0''(u)\, u)/u^2$ and $(\rho_1'(u) - \rho_1''(u)\, u)/u^2$ are continuous;*

then along almost all sample sequences, conditional on the first n pairs, $\sqrt{n}\,(\hat{\boldsymbol{\beta}}_n^{R*} - \hat{\boldsymbol{\beta}}_n)$ converges weakly, as n goes to infinity, to the same limit distribution as $\sqrt{n}\,(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})$.

Remark 5.4 We refer to Remark 3.3 on page 102 to verify that assumption 3 above is satisfied for functions ρ_d in Tukey's family (2.8).

Proof: First note that the estimates $\hat{\boldsymbol{\beta}}_n$, $\hat{\sigma}_n$ and $\tilde{\boldsymbol{\beta}}_n$ satisfy the following equations

$$\frac{1}{n}\sum_{i=1}^{n}\rho_1'\left(\frac{r_i\left(\hat{\boldsymbol{\beta}}_n\right)}{\hat{\sigma}_n}\right)\mathbf{x}_i=\mathbf{0}$$

$$\frac{1}{n}\sum_{i=1}^{n}\rho_0\left(\frac{r_i\left(\tilde{\boldsymbol{\beta}}_n\right)}{\hat{\sigma}_n}\right)=b$$

$$\frac{1}{n}\sum_{i=1}^{n}\rho_0'\left(\frac{r_i\left(\tilde{\boldsymbol{\beta}}_n\right)}{\hat{\sigma}_n}\right)\mathbf{x}_i=\mathbf{0}.$$

Simple calculations yield the following re-weighted version of the estimates

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_n &= \mathbf{A}_n\left(\hat{\boldsymbol{\beta}}_n,\hat{\sigma}_n\right)^{-1}\ \mathbf{v}_n\left(\hat{\boldsymbol{\beta}}_n,\hat{\sigma}_n\right)\\
\hat{\sigma}_n &= \hat{\sigma}_n\ u_n\left(\tilde{\boldsymbol{\beta}}_n,\hat{\sigma}_n\right)\\
\tilde{\boldsymbol{\beta}}_n &= \mathbf{B}_n\left(\tilde{\boldsymbol{\beta}}_n,\hat{\sigma}_n\right)^{-1}\ \mathbf{w}_n\left(\tilde{\boldsymbol{\beta}}_n,\hat{\sigma}_n\right),
\end{aligned}\tag{5.13}$$

where

$$\begin{aligned}
\mathbf{A}_n\left(\boldsymbol{\beta}_1,\sigma\right) &= \frac{1}{n}\sum_{i=1}^{n}\omega_i\left(\boldsymbol{\beta}_1,\sigma\right)\mathbf{x}_i\mathbf{x}_i',\\
\mathbf{v}_n\left(\boldsymbol{\beta}_1,\sigma\right) &= \frac{1}{n}\sum_{i=1}^{n}\omega_i\left(\boldsymbol{\beta}_1,\sigma\right)y_i\mathbf{x}_i,\\
u_n\left(\boldsymbol{\beta}_2,\sigma\right) &= \sum_{i=1}^{n}v_i\left(\boldsymbol{\beta}_2,\sigma\right)\tilde{r}_i,\\
\mathbf{B}_n\left(\boldsymbol{\beta}_2,\sigma\right) &= \frac{1}{n}\sum_{i=1}^{n}w_i\left(\boldsymbol{\beta}_2,\sigma\right)\mathbf{x}_i\mathbf{x}_i',\\
\mathbf{w}_n\left(\boldsymbol{\beta}_2,\sigma\right) &= \frac{1}{n}\sum_{i=1}^{n}w_i\left(\boldsymbol{\beta}_2,\sigma\right)y_i\mathbf{x}_i,\\
\omega_i\left(\boldsymbol{\beta}_1,\sigma\right) &= \rho_1'\left(r_i/\sigma\right)/\,r_i,\\
v_i\left(\boldsymbol{\beta}_2,\sigma\right) &= \rho_0\left(\tilde{r}_i/\sigma\right)/\left(n\ b\ \tilde{r}_i\right),
\end{aligned}$$

and

$$w_i\left(\boldsymbol{\beta}_2, \sigma\right) = \rho_0'\left(\tilde{r}_i / \sigma\right) / \tilde{r}_i .$$

Equations (5.13) can be expressed as the fixed point of a conveniently chosen function. Consider $\mathbf{f} : \mathbb{R}^{2p+1} \to \mathbb{R}^{2p+1}$ defined for $\boldsymbol{\beta}_1 \in \mathbb{R}^p$, $\sigma \in \mathbb{R}$ and $\boldsymbol{\beta}_2 \in \mathbb{R}^p$ by

$$\mathbf{f}\left(\boldsymbol{\beta}_1, \sigma, \boldsymbol{\beta}_2\right) = \begin{pmatrix} \mathbf{A}_n\left(\boldsymbol{\beta}_1, \sigma\right)^{-1} \ \mathbf{v}_n\left(\boldsymbol{\beta}_1, \sigma\right) \\ \sigma \ u_n\left(\boldsymbol{\beta}_2, \sigma\right) \\ \mathbf{B}_n\left(\boldsymbol{\beta}_2 \sigma\right)^{-1} \ \mathbf{w}_n\left(\boldsymbol{\beta}_2, \sigma\right) \ . \end{pmatrix}$$

To simplify the notation we do not explicitly indicate the dependence of $\mathbf{f}$ on n. We have

$$\mathbf{f}\left(\hat{\boldsymbol{\beta}}_n, \hat{\sigma}_n, \tilde{\boldsymbol{\beta}}_n\right) = \left(\hat{\boldsymbol{\beta}}_n, \hat{\sigma}_n, \tilde{\boldsymbol{\beta}}_n\right)' .$$

Using the differentiability of ρ_0 and ρ_1 we can calculate a Taylor expansion of $\mathbf{f}$ about the limiting values of the estimates $(\boldsymbol{\beta}, \sigma, \tilde{\boldsymbol{\beta}})$,

$$\begin{pmatrix} \hat{\boldsymbol{\beta}}_n \\ \hat{\sigma}_n \\ \tilde{\boldsymbol{\beta}}_n \end{pmatrix} = \mathbf{f}\left(\boldsymbol{\beta}, \sigma, \tilde{\boldsymbol{\beta}}\right) + \nabla \mathbf{f}\left(\boldsymbol{\beta}, \sigma, \tilde{\boldsymbol{\beta}}\right) \begin{pmatrix} \hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta} \\ \hat{\sigma}_n - \sigma \\ \tilde{\boldsymbol{\beta}}_n - \tilde{\boldsymbol{\beta}} \end{pmatrix} + R_n, \tag{5.14}$$

where R_n is the remainder term and $\nabla \mathbf{f}\left(\cdot\right) \in \mathbb{R}^{(2p+1)\times(2p+1)}$ is the matrix of partial derivatives,

$$\begin{array}{c|ccc} & p & 1 & p \\ \hline p & \partial\left[\mathbf{A}_n^{-1} \ \mathbf{v}_n\right] / \partial\boldsymbol{\beta} & \partial\left[\mathbf{A}_n^{-1} \ \mathbf{v}_n\right] / \partial\sigma & \partial\left[\mathbf{A}_n^{-1} \ \mathbf{v}_n\right] / \partial\tilde{\boldsymbol{\beta}} \\ 1 & \partial\left[\sigma \ u_n\right] / \partial\boldsymbol{\beta} & \partial\left[\sigma \ u_n\right] / \partial\sigma & \partial\left[\sigma \ u_n\right] / \partial\tilde{\boldsymbol{\beta}} \\ p & \partial\left[\mathbf{B}_n^{-1} \ \mathbf{w}_n\right] / \partial\boldsymbol{\beta} & \partial\left[\mathbf{B}_n^{-1} \ \mathbf{w}_n\right] / \partial\sigma & \partial\left[\mathbf{B}_n^{-1} \ \mathbf{w}_n\right] / \partial\tilde{\boldsymbol{\beta}} \end{array}$$

Tedious but straightforward calculations show that each entry in R_n is a linear combination of quadratic forms $\mathbf{x}_n' \, H_n \, \mathbf{x}_n$ where $\mathbf{x}_n = \hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}$ or $\mathbf{x}_n = \hat{\sigma}_n - \sigma$ or $\mathbf{x}_n = \tilde{\boldsymbol{\beta}}_n - \tilde{\boldsymbol{\beta}}$. Note that $\|\mathbf{x}_n\| = O_P\,(1/\sqrt{n})$. The regularity conditions on ρ_0 and ρ_1 and Lemma 7.1 show that $\|H_n\| = O_P\,(1)$. We have $|\mathbf{x}_n' \, H_n \, \mathbf{x}_n| = o_P\,(1/\sqrt{n})$. Hence $\|R_n\| = o_p\,(1/\sqrt{n})$ in (5.14).

To simplify the notation let $\boldsymbol{\tau}_n = (\hat{\boldsymbol{\beta}}_n, \hat{\sigma}_n, \tilde{\boldsymbol{\beta}}_n)'$ and $\boldsymbol{\tau} = (\boldsymbol{\beta}, \sigma, \tilde{\boldsymbol{\beta}})'$. Equation (5.14) becomes

$$\sqrt{n}\,(\boldsymbol{\tau}_n - \boldsymbol{\tau}) = [\mathbf{I} - \nabla \mathbf{f}\,(\boldsymbol{\tau})]^{-1} \; \sqrt{n}\,[\mathbf{f}\,(\boldsymbol{\tau}) - \boldsymbol{\tau}] + o_P\,(1)\;. \tag{5.15}$$

We will now show that the correction factors $\mathbf{M}_n$, and $\mathbf{d}_n$ in (5.6) and (5.7) are the corresponding first p rows of the estimate $[\mathbf{I} - \nabla \mathbf{f}\,(\boldsymbol{\tau}_n)]^{-1}$ of the matrix $[\mathbf{I} - \nabla \mathbf{f}\,(\boldsymbol{\tau})]^{-1}$ in (5.15). It is easy to see that $\mathbf{I} - \nabla \mathbf{f}\,(\boldsymbol{\tau}_n)$ has the following form

$$\mathbf{I} - \nabla \mathbf{f}\,(\boldsymbol{\tau}_n) = \begin{array}{|c|c|c|} \hline \mathcal{A} & v & \begin{matrix} 0 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & 0 \end{matrix} \\ \hline 0 \;\cdots\; 0 & a & 0 \;\cdots\; 0 \\ \hline \begin{matrix} 0 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & 0 \end{matrix} & w & \mathcal{B} \\ \hline \end{array} \;,$$

where

$$\mathcal{A} = \mathbf{I} - \frac{\partial}{\partial \boldsymbol{\beta}}\left[\mathbf{A}_n^{-1}\,\mathbf{v}_n\right], \quad v = -\frac{\partial}{\partial \sigma}\left[\mathbf{A}_n^{-1}\,\mathbf{v}_n\right], \quad a = 1 - \frac{\partial}{\partial \sigma}\left[\sigma\; u_n\right],$$

$$w = -\frac{\partial}{\partial \sigma}\left[\mathbf{B}_n^{-1}\,\mathbf{w}_n\right] \quad \text{and} \quad \mathcal{B} = \mathbf{I} - \frac{\partial}{\partial \tilde{\boldsymbol{\beta}}}\left[\mathbf{B}_n^{-1}\,\mathbf{w}_n\right].$$

That

$$\frac{\partial}{\partial\tilde{\boldsymbol{\beta}}}\left[u_n\right]=(0,\ldots,0)$$

follows from the fact that $\hat{\sigma}_n$ attains the minimum of the S-scale.

Now note that the estimate of the correction factor in (5.15) has the following form:

$$\left[\mathbf{I}-\nabla\mathbf{f}\left(\boldsymbol{\tau}_n\right)\right]^{-1}=\begin{array}{|c|c|c|}\hline \mathcal{A}^{-1} & -\mathcal{A}^{-1}v/\,a & \begin{matrix}0 & \cdots & 0\\ \vdots & & \vdots\\ 0 & \cdots & 0\end{matrix}\\ \hline 0\ \cdots\ 0 & 1/\,a & 0\ \cdots\ 0\\ \hline \begin{matrix}0 & \cdots & 0\\ \vdots & & \vdots\\ 0 & \cdots & 0\end{matrix} & -\mathcal{B}^{-1}w/\,a & \mathcal{B}^{-1}\\ \hline\end{array}\,. \tag{5.16}$$

Note that in (5.15) we are only interested in the first $p+1$ coordinates of $\boldsymbol{\tau}_n$ (the remaining p correspond to the S-regression estimate). From $\left[\mathbf{I}-\nabla\mathbf{f}\left(\boldsymbol{\tau}_n\right)\right]^{-1}$ in (5.16) we see that the last p coordinates of $\mathbf{f}$ are not involved in determining the first $p+1$ coordinates of $\boldsymbol{\tau}_n-\boldsymbol{\tau}$. Hence, when we apply this method in practice we do not need to bootstrap $\tilde{\boldsymbol{\beta}}_n$.

It also follows from (5.16) that we only need to calculate $\mathcal{A}$, v and a. We need to find

$$\frac{\partial}{\partial\boldsymbol{\beta}}\left[\mathbf{A}_n\left(\boldsymbol{\beta},\sigma\right)^{-1}\ \mathbf{v}_n\left(\boldsymbol{\beta},\sigma\right)\right]\Big|_{\hat{\boldsymbol{\beta}}_n,\hat{\sigma}_n}\ ,$$

and

$$\frac{\partial}{\partial\sigma}\left[\mathbf{A}_n\left(\boldsymbol{\beta},\sigma\right)^{-1}\ \mathbf{v}_n\left(\boldsymbol{\beta},\sigma\right)\right]\Big|_{\hat{\boldsymbol{\beta}}_n,\hat{\sigma}_n}\ .$$

One way to calculate them is to differentiate the vector $\boldsymbol{\alpha}_n$ defined implicitly by

$$\mathbf{A}_n(\boldsymbol{\beta},\sigma)\ \boldsymbol{\alpha}_n(\boldsymbol{\beta},\sigma) = \mathbf{v}_n(\boldsymbol{\beta},\sigma)\ .$$

Drop the arguments $(\boldsymbol{\beta},\sigma)$ and the subscripts to simplify the notation. We differentiate on both sides of the equation

$$\frac{\partial}{\partial\boldsymbol{\beta}}[\mathbf{A}\,\boldsymbol{\alpha}] = \frac{\partial}{\partial\boldsymbol{\beta}}\mathbf{v}\,.$$

Note that

$$\begin{aligned}\frac{\partial}{\partial\boldsymbol{\beta}}[\mathbf{A}\ \boldsymbol{\alpha}] &= \mathbf{A}\ \frac{\partial}{\partial\boldsymbol{\beta}}\boldsymbol{\alpha} + \begin{pmatrix} | & \vdots\ \vdots & | \\ \left(\frac{\partial}{\partial\beta_1}\mathbf{A}\right)\boldsymbol{\alpha} & \vdots\ \vdots & \left(\frac{\partial}{\partial\beta_p}\mathbf{A}\right)\boldsymbol{\alpha} \\ | & \vdots\ \vdots & | \end{pmatrix} \\ &= \mathbf{A}\ \frac{\partial}{\partial\boldsymbol{\beta}}\boldsymbol{\alpha} + \tilde{\mathbf{A}},\end{aligned}$$

say. So

$$\frac{\partial}{\partial\boldsymbol{\beta}}\boldsymbol{\alpha} = \mathbf{A}^{-1}\ \left[\frac{\partial}{\partial\boldsymbol{\beta}}\mathbf{v} - \tilde{\mathbf{A}}\right]\ .$$

Simple calculations show that

$$\left.\frac{\partial}{\partial\boldsymbol{\beta}}\mathbf{v}\right|_{\hat{\boldsymbol{\beta}}_n,\hat{\sigma}_n} = \sum_{i=1}^{n} y_i \frac{\rho_1'(r_i/\,\hat{\sigma}_n) - \rho_1''(r_i/\,\hat{\sigma}_n)\,r_i/\,\hat{\sigma}_n}{r_i^2}\mathbf{x}_i\mathbf{x}_i'\ ,$$

and

$$\begin{aligned}\left.\tilde{\mathbf{A}}\right|_{\hat{\boldsymbol{\beta}}_n,\hat{\sigma}_n} &= \frac{\partial}{\partial\boldsymbol{\beta}}\mathbf{v} - \sum_{i=1}^{n}\rho_1'(r_i/\,\hat{\sigma}_n)/\,r_i\mathbf{x}_i\mathbf{x}_i' + \sum_{i=1}^{n}\rho_1''(r_i/\,\hat{\sigma}_n)/\,\hat{\sigma}_n\mathbf{x}_i\mathbf{x}_i'\,, \\ &= \frac{\partial}{\partial\boldsymbol{\beta}}\mathbf{v} - \mathbf{A} + \sum_{i=1}^{n}\rho_1''(r_i/\,\hat{\sigma}_n)/\,\hat{\sigma}_n\mathbf{x}_i\mathbf{x}_i'\,,\end{aligned}$$

which yields

$$\frac{\partial}{\partial\boldsymbol{\beta}}\boldsymbol{\alpha} = \mathbf{A}^{-1}\ \left[\mathbf{A} - \sum_{i=1}^{n}\rho_1''(r_i/\,\hat{\sigma}_n)/\,\hat{\sigma}_n\mathbf{x}_i\mathbf{x}_i'\right]\ .$$

It follows that

$$\mathcal{A} = \mathbf{I} - \frac{\partial}{\partial \boldsymbol{\beta}} \boldsymbol{\alpha} \bigg|_{\hat{\boldsymbol{\beta}}, \hat{\sigma}_n} = \mathbf{A}^{-1} \frac{1}{\hat{\sigma}_n} \sum_{i=1}^{n} \rho_1'' \left(r_i / \hat{\sigma}_n \right) \mathbf{x}_i \mathbf{x}_i' \, .$$

Then we have

$$\mathcal{A}^{-1} = \hat{\sigma}_n \left(\sum_{i=1}^{n} \rho_1'' \left(r_i / \hat{\sigma}_n \right) \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \mathbf{A} \, , \tag{5.17}$$

and

$$-\mathcal{A}^{-1} \; v / \, a = \frac{b \; n \; \hat{\sigma}_n \left[\sum_{i=1}^{n} \rho_1'' \left(r_i / \hat{\sigma}_n \right) \mathbf{x}_i \mathbf{x}_i' \right]^{-1} \; \sum_{i=1}^{n} \rho_1'' \left(r_i / \hat{\sigma}_n \right) r_i / \hat{\sigma}_n \mathbf{x}_i}{\sum_{i=1}^{n} \rho_0' \left(r_i / \hat{\sigma}_n \right) r_i / \hat{\sigma}_n} \, . \tag{5.18}$$

It is easy to see that $\mathbf{M}_n$ in (5.6) is equal to (5.17) above, and that $\mathbf{d}_n$ in (5.7) is $-\mathcal{A}^{-1} \; v / \, a$ in (5.18).

We will now show that the bootstrap distribution of $\sqrt{n} \left[\mathbf{f}^* \left(\boldsymbol{\tau}_n \right) - \boldsymbol{\tau}_n \right]$ converges to the same limiting distribution as that of the sequence $\sqrt{n} \left[\mathbf{f}^* \left(\boldsymbol{\tau} \right) - \boldsymbol{\tau} \right]$.

First, note that

$$\left[\mathbf{f}^* \left(\boldsymbol{\tau}_n \right) - \boldsymbol{\tau}_n \right] = \begin{pmatrix} \hat{\boldsymbol{\beta}}_n^* - \hat{\boldsymbol{\beta}}_n \\ \hat{\sigma}_n^* - \hat{\sigma}_n \\ \tilde{\boldsymbol{\beta}}_n^* - \tilde{\boldsymbol{\beta}}_n \end{pmatrix} = \begin{pmatrix} \mathbf{A}_n^{*\,-1} \; \mathbf{v}_n^* - \hat{\boldsymbol{\beta}}_n \\ \hat{\sigma}_n \; u_n^* - \hat{\sigma}_n \\ \mathbf{B}_n^{*\,-1} \; \mathbf{w}_n^* - \tilde{\boldsymbol{\beta}}_n \end{pmatrix}$$

where $*$ denotes the bootstrap version of these quantities. It is easy to see that

$$\mathbf{v}_n^* \left(\hat{\boldsymbol{\beta}}_n, \hat{\sigma}_n \right) = \sum_{i=1}^{n} \rho_1' \left(r_i^* / \hat{\sigma}_n \right) \mathbf{x}_i^* \; + \; \mathbf{A}_n^* \left(\hat{\boldsymbol{\beta}}_n, \hat{\sigma}_n \right) \; \hat{\boldsymbol{\beta}}_n$$

and

$$\mathbf{w}_n^* \left(\hat{\boldsymbol{\beta}}_n, \hat{\sigma}_n \right) = \sum_{i=1}^{n} \rho_0' \left(\tilde{r}_i^* / \hat{\sigma}_n \right) \mathbf{x}_i^* \; + \; \mathbf{B}_n^* \left(\hat{\boldsymbol{\beta}}_n, \hat{\sigma}_n \right) \; \tilde{\boldsymbol{\beta}}_n \, .$$

Then

$$\begin{pmatrix} \hat{\boldsymbol{\beta}}_n^* - \hat{\boldsymbol{\beta}}_n \\ \hat{\sigma}_n^* - \hat{\sigma}_n \\ \tilde{\boldsymbol{\beta}}_n^* - \tilde{\boldsymbol{\beta}}_n \end{pmatrix} = \begin{pmatrix} \mathbf{A}_n^{*-1} \ \sum_{i=1}^n \rho_1' \left(r_i^* / \hat{\sigma}_n \right) \mathbf{x}_i^* \\ \hat{\sigma}_n \ u_n^* - \hat{\sigma}_n \\ \mathbf{B}_n^{*-1} \ \sum_{i=1}^n \rho_0' \left(\tilde{r}_i^* / \hat{\sigma}_n \right) \mathbf{x}_i^* \end{pmatrix} . \tag{5.19}$$

This last expression can be expressed as a function of means. Consider the function $\mathbf{g} : \mathbb{R}^{p \times p} \times \mathbb{R}^p \times \mathbb{R}^p \times \mathbb{R}^{p \times p} \times \mathbb{R}^p \to \mathbb{R}^p \times \mathbb{R}^p \times \mathbb{R}^p$,

$$\mathbf{g} \left(\overline{\mathbf{A}}, \overline{\mathbf{v}}, \overline{u}, \overline{\mathbf{B}}, \overline{\mathbf{w}} \right) = \left(\overline{\mathbf{A}}^{-1} \overline{\mathbf{v}}, \overline{u}, \overline{\mathbf{B}}^{-1} \overline{\mathbf{w}} \right).$$

Then (5.19) can be written as $\mathbf{g} \left(\mathbf{A}_n^*, \overline{\mathbf{z}}^*, u_n^*, \mathbf{B}_n^*, \overline{\mathbf{w}}^* \right)$ where $\mathbf{A}_n^*$, u_n^* and $\mathbf{B}_n^*$ are as before, $\mathbf{z}_i = \rho_1' \left(r_i^* / \hat{\sigma}_n \right) \mathbf{x}_i^*$, and $\mathbf{w}_i = \rho_0' \left(\tilde{r}_i^* / \hat{\sigma}_n \right) \mathbf{x}_i^*$ for $1 \leq i \leq n$. This function is differentiable (this can be seen by thinking it as a composition of differentiable functions). We have that the statistic we are bootstrapping is of the form

$$\mathbf{g} \left(\bar{\mathbf{y}}_n \left(\boldsymbol{\tau}_n \right) \right) - \mathbf{g} \left(\boldsymbol{\mu} \left(\boldsymbol{\tau}_n \right) \right),$$

where $\mathbf{y}_i$ for $1 \leq i \leq n$ is a vector of the bootstrapped dimension and $\boldsymbol{\tau}_n$ is a consistent estimate of the vector of parameters $\boldsymbol{\tau}$. As in the proof of Theorem 3.1 we have to show that the asymptotic distribution of

$$\sqrt{n} \left(\bar{\mathbf{y}}_n \left(\boldsymbol{\tau}_n \right) - \boldsymbol{\mu} \left(\boldsymbol{\tau}_n \right) \right) \tag{5.20}$$

is the same as that of

$$\sqrt{n} \left(\bar{\mathbf{y}}_n \left(\boldsymbol{\tau} \right) - \boldsymbol{\mu} \left(\boldsymbol{\tau} \right) \right) . \tag{5.21}$$

The proof of this last statement uses the same idea as that used in Theorem 3.1 to prove the corresponding statement for the location-scale model. It is based on bounding the distance d_2 (see Bickel and Freedman, 1981) between the distribution functions of (5.20) and (5.21), using the fact that $\boldsymbol{\tau}_n \to \boldsymbol{\tau}$ almost surely. Lemma 8.1 of Bickel and Freedman (1981) and the regularity conditions of $\mathbf{g}$ show that the bootstrap distribution of $\mathbf{g}(\bar{\mathbf{y}}_n(\boldsymbol{\tau}_n)) - \mathbf{g}(\boldsymbol{\mu}(\boldsymbol{\tau}_n))$ converges to the same limit as that of the sequence $\mathbf{g}(\bar{\mathbf{y}}_n(\boldsymbol{\tau})) - \mathbf{g}(\boldsymbol{\mu}(\boldsymbol{\tau}))$. ■

5.4 Robustness properties

We are also interested in the robustness properties of the quantile estimates of our robust bootstrap. Let $t \in (0,1)$, and let q_t be the t-th upper quantile of a statistic $\hat{\boldsymbol{\theta}}_n$, that is, q_t satisfies $P[\hat{\boldsymbol{\theta}}_n > q_t] = t$. As in Section 3.4 we define the upper breakdown point (UB) of a bootstrap estimate $\hat{q}_t$ as the smallest proportion of arbitrarily large outliers such that we expect $\hat{q}_t$ to be driven above any bound in at least $t \times 100$ % of the bootstrap samples.

For the linear regression model we need an extra assumption. To fix ideas consider a linear regression model with a single explanatory variable ($x_i \in \mathbb{R}$). We will require that no two x's are equal. If the design is random then this event has probability one. If the data contain two observations at the same x the breakdown point of the robust bootstrap quantile estimates will decrease. Intuitively this is due to the fact that in this case we would not need to introduce outliers in the sample to get an unbounded recalculated slope: a bootstrap sample consisting only of these vertically aligned points will result in an arbitrarily large set of coefficients. The following definition can also be found in Rousseeuw and Leroy (1987).

Definition 5.1 - **General position** - *We say that k points in $\mathbb{R}^p$ are in general position if no subset of size $p+1$ of them determines an affine subspace of dimension p. In other words, for every subset $\mathbf{x}_{i_1}, \ldots, \mathbf{x}_{i_{p+1}}$, $1 \leq i_j \leq k$, $i_j \neq i_l$ if $j \neq l$, there is no vector $\mathbf{v}_0 \in \mathbb{R}^p \setminus \{\mathbf{0}\}$ and scalar $\alpha \in \mathbb{R}$ such that*

$$\mathbf{x}_{i_j}' \mathbf{v}_0 = \alpha \quad \textit{for} \quad j = 1, \ldots, p+1.$$

The main result of this section is the following theorem that establishes the breakdown point of the quantile estimates based on the robust bootstrap.

Theorem 5.2 - **Breakdown point of the robust bootstrap quantiles for the regression model** - *Let $(y_1, \mathbf{x}_1), \ldots, (y_n, \mathbf{x}_n) \in \mathbb{R}^{p+1}$ be a random sample following the linear model (4.1). Assume that the explanatory variables $\mathbf{x}_1$, ..., $\mathbf{x}_n$ in $\mathbb{R}^p$ are in general position (see Definition 5.1). Let $\hat{\boldsymbol{\beta}}_n$ be an MM-regression estimate and let ϵ^* be its breakdown point. Then the breakdown point of the t-th robust bootstrap quantile estimate of the regression parameters β_j, $j = 1, \ldots, p$ is given by $\min(\epsilon^*, \epsilon)$, where ϵ is the smallest solution in δ of the equation*

$$P\,[\ \textit{Binomial}(n, 1-\delta) < p\,] \geq t\,.$$

The following lemma is needed for the proof of Theorem 5.2.

Lemma 5.1 *Let $(y_1, \mathbf{x}_1), \ldots, (y_n, \mathbf{x}_n)$ be $n \geq p$ points in $\mathbb{R}^p$ such that if*

$$\mathbf{X}_n = \begin{bmatrix} \mathbf{x}_1' \\ \vdots \\ \mathbf{x}_n' \end{bmatrix}, \tag{5.22}$$

then $\mathbf{X}_n'\mathbf{X}_n$ *has full rank. For a given* $(y_{n+1}, \mathbf{x}_{n+1})$ *let* $\hat{\boldsymbol{\beta}}_{n+1}$ *be the least squares regression coefficients determined by the* $n+1$ *points. There exists a finite constant* K *such that* $\left\|\hat{\boldsymbol{\beta}}_{n+1}\right\| \leq K$ *for any* $(y_{n+1}, \mathbf{x}_{n+1})$ *with* $|y_{n+1}| \leq c$. *(The constant* K *only depends on the first* n *points and on the constant* c*)*

Proof: We will show that for any set of $n \geq p$ points, the regression parameters $\hat{\boldsymbol{\beta}}_{n+1}$ obtained when adding a new point $(y_{n+1}, \mathbf{x}_{n+1})$ are bounded for any $\mathbf{x}_{n+1}$ if y_{n+1} is bounded. Let $\mathbf{X}_n \in \mathbb{R}^{n \times p}$ be the design matrix in (5.22). Note that $\mathbf{X}_n$ has rank p by hypothesis. As a consequence both $(\mathbf{X}_n' \mathbf{X}_n)$ and its inverse are positive definite. Let $\mathbf{A}$ be a non-singular matrix in $\mathbb{R}^{p \times p}$ and let $\mathbf{x} \in \mathbb{R}^p$. Use the following formula (see for example Seber (1984), page 519)

$$(\mathbf{A} + \mathbf{x}\,\mathbf{x}')^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{x}\,\mathbf{x}'\,\mathbf{A}^{-1}\left(1 + \mathbf{x}'\,\mathbf{A}^{-1}\,\mathbf{x}\right)^{-1}$$

to obtain

$$\hat{\boldsymbol{\beta}}_{n+1} = \left[\mathbf{I} - \frac{\mathbf{V}\,\mathbf{x}_{n+1}\,\mathbf{x}_{n+1}'}{1 + \mathbf{x}_{n+1}'\,\mathbf{V}\,\mathbf{x}_{n+1}}\right]\hat{\boldsymbol{\beta}}_n + \left[\mathbf{V} - \frac{\mathbf{V}\,\mathbf{x}_{n+1}\,\mathbf{x}_{n+1}'\mathbf{V}}{1 + \mathbf{x}_{n+1}'\,\mathbf{V}\,\mathbf{x}_{n+1}}\right]\mathbf{x}_{n+1}\,y_{n+1},$$

where $\mathbf{V} = (\mathbf{X}_n'\,\mathbf{X}_n)^{-1}$ is positive definite and $(y_{n+1}, \mathbf{x}_{n+1})$ is the new point to be added to the regression. To simplify the notation let $\mathbf{u} = \mathbf{x}_{n+1}$,

$$\mathbf{A} = \mathbf{I} - \frac{\mathbf{V}\,\mathbf{u}\,\mathbf{u}'}{1 + \mathbf{u}'\,\mathbf{V}\,\mathbf{u}}, \qquad \text{and} \qquad \mathbf{B} = \mathbf{V} - \frac{\mathbf{V}\,\mathbf{u}\,\mathbf{u}'\mathbf{V}}{1 + \mathbf{u}'\,\mathbf{V}\,\mathbf{u}}.$$

The last equation can then be written as

$$\boldsymbol{\beta}_{n+1} = \mathbf{A}\,\boldsymbol{\beta}_n + \mathbf{B}\,\mathbf{u}\,y_{n+1}.$$

First we will show that every entry in $\mathbf{A}$ is bounded for $\|\mathbf{u}\| \to \infty$. The (i, j) element is given by

$$\begin{aligned}\mathbf{A}_{(i,j)} &= \delta_{i,j} - \frac{u_j \sum_k v_{i\,k} u_k}{1+\sum_k \sum_l v_{k\,l} u_k u_l} \\ &= \delta_{i,j} - \frac{v_{i\,j} u_j^2 + u_j \left(\sum_{k \neq j} v_{i\,k} u_k\right)}{1+\sum_k \sum_l v_{k\,l} u_k u_l}\,.\end{aligned}$$

It is easy to see (for example by dividing both the numerator and denominator by $\|\mathbf{u}\|^2$) that the denominator has the same order as the numerator, so that the fraction will remain bounded as $\|\mathbf{u}\| \to \infty$. Note that the denominator is bounded away from zero, so that the whole expression is bounded for any $\mathbf{u}$. We now show that the rth element of $\mathbf{B}\,\mathbf{u}$ goes to zero as $\|\mathbf{u}\| \to \infty$. Note that

$$\mathbf{B}\,\mathbf{u} = \frac{\mathbf{V}\,\mathbf{u}}{1+\mathbf{u}'\,\mathbf{V}\,\mathbf{u}}.$$

The rth element is then

$$\frac{\sum_i v_{r\,i}\, u_i}{1+\sum_{i\,j} v_{i\,j}\, u_i\, u_j}.$$

Divide both numerator and denominator by $\|\mathbf{u}\|^2$ and use that

$$\frac{|\mathbf{u}_j|}{\|\mathbf{u}\|} \leq 1 \qquad \text{for } 1 \leq j \leq p\,, \tag{5.23}$$

to conclude that the denominator is bounded, and that the numerator goes to zero because (5.23) implies

$$\frac{|\mathbf{u}_j|}{\|\mathbf{u}\|^2} \to 0 \qquad \text{for } 1 \leq j \leq p\,.$$

■

Proof of Theorem 5.2: Let $(y_1, \mathbf{x}_1), \ldots, (y_n, \mathbf{x}_n) \in \mathbb{R}^{p+1}$ be n observations following model (4.1). We assume that $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$ are in general position (see

Definition 5.1 above). This assumption guarantees that any subset of size p of them will determine a bounded least squares estimate.

We assume that there is a certain proportion of observations that do not necessarily follow the linear regression model (4.1). We will show that any bootstrap sample that contains at least p points that are not outliers yields a bounded $\hat{\boldsymbol{\beta}}_n^{R*}$. It follows that the only samples that can produce unbounded robust bootstrap coefficients are those that contain at most $p-1$ points that are not outliers. The robust bootstrap $\hat{\boldsymbol{\beta}}_n^{R*}$ is given by

$$\hat{\boldsymbol{\beta}}_n^{R*} = \mathbf{M}_n \left(\hat{\boldsymbol{\beta}}_n^* - \hat{\boldsymbol{\beta}}_n \right) + \mathbf{d}_n \left(\hat{\sigma}_n^* - \hat{\sigma}_n \right) .$$

Note that the matrix $\mathbf{M}_n$ and the vector $\mathbf{d}_n$ are not re-calculated with each bootstrap sample, and as long as the robust regression estimate $\hat{\boldsymbol{\beta}}_n$ does not breakdown, they remain bounded. It is also easy to see that $\hat{\sigma}_n^*$ also remains bounded for any bootstrap sample. Hence, the problem becomes determining under which circumstances $\hat{\boldsymbol{\beta}}_n^*$ can be driven beyond any finite bound. Recall that

$$\hat{\boldsymbol{\beta}}_n^* = \left[\sum_{i=1}^n \omega_i^* \, \mathbf{x}_i^* \, \mathbf{x}_i^{*\prime} \right]^{-1} \sum_{i=1}^n \omega_i^* \, \mathbf{x}_i^* \, y_i^* ,$$

where the weights $\omega_i^* = \rho_1' \left(r_i^* / \hat{\sigma}_n \right) / r_i^*$ are bounded. The above expression can be re-written as

$$\hat{\boldsymbol{\beta}}_n^* = \left[\sum_{i=1}^n \tilde{\mathbf{x}}_i^* \, \tilde{\mathbf{x}}_i^{*\prime} \right]^{-1} \sum_{i=1}^n \tilde{\mathbf{x}_i}^* \, \tilde{y_i}^* ,$$

where $\tilde{\mathbf{x}_i}^* = \sqrt{\omega_i^*}\, \mathbf{x}_i^*$ and $\tilde{y_i}^* = \sqrt{\omega_i^*}\, y_i^*$. We consider the case of having at least p data points that are not outliers. It is enough to have a bound on the effect of one outlier and that that bound does not depend on the outlier. In what follows we show

how to obtain such a bound. To simplify the notation we use the same symbols $\mathbf{x}_i$ and y_i for the weighted points $\tilde{\mathbf{x}}_i$ and $\tilde{y}_i$.

Let $(y_1, \mathbf{x}_1), \ldots, (y_n, \mathbf{x}_n)$, be a bootstrap sample of $n \geq p$ good data points, and let $(y_{n+1}, \mathbf{x}_{n+1})$ be an arbitrary outlier included in this sample. Let $\hat{\boldsymbol{\beta}}_n$ be the MM-estimate based on the full data. Without loss of generality assume that $\hat{\boldsymbol{\beta}}_n = \mathbf{0} \in \mathbb{R}^p$. The data can always be transformed to satisfy this assumption. In particular if

$$\tilde{y}_i = y_i - \hat{\boldsymbol{\beta}}_n' \mathbf{x}_i \qquad i = 1, \ldots, n,$$

then the points $(\tilde{y}_1, \mathbf{x}_1), \ldots, (\tilde{y}_n, \mathbf{x}_n)$ have a zero regression estimate.

We now show that the outlier $(y_{n+1}, \mathbf{x}_{n+1})$ will only have an effect on $\hat{\boldsymbol{\beta}}^*_{n+1}$ for a bounded range of y_{n+1}. Let $c > 0$ be the constant of the function ψ_c used for the MM-estimate in (4.7), and let $\sigma_n^+ = \sup \hat{\sigma}_n$ be the largest possible value of $\hat{\sigma}_n$ for a sample of size n. Any point $(y_{n+1}, \mathbf{x}_{n+1})$ such that $|y_{n+1}| > \sigma_n^+ c$ will not affect $\hat{\boldsymbol{\beta}}_n$ and will receive a null weight in the robust bootstrap re-calculations. Hence it is not possible to upset $\hat{\boldsymbol{\beta}}^*_{n+1}$ with this type of contamination. In what follows we consider the case $|y_{n+1}| \leq \sigma_n^+ c$. Lemma 5.1 gives a bound for the effect of $(y_{n+1}, \mathbf{x}_{n+1})$ on $\hat{\boldsymbol{\beta}}^*_{n+1}$. This bound only depends on the first n pairs.

Given a bootstrap sample of size n, assume that the first k observations are "good" and the remaining $n - k$ are arbitrary outliers. Applying Lemma 5.1 $n - k$ times we see that the new $\hat{\boldsymbol{\beta}}^*_{n+1}$ can only be modified by a finite amount. This amount depends on the k first observations of this bootstrap sample, but it does not depend on the values of the $n - k$ outliers. Considering all the possible bootstrap samples that contain at least p points that are not outliers we find a bound that only depends on

		Robust Bootstrap			Classical Bootstrap		
p	n	$\hat{q}_{0.005}$	$\hat{q}_{0.025}$	$\hat{q}_{0.05}$	$\hat{q}_{0.005}$	$\hat{q}_{0.025}$	$\hat{q}_{0.05}$
1	10	0.500	0.500	0.500	0.191	0.262	0.304
	20	0.500	0.500	0.500	0.257	0.315	0.347
	30	0.500	0.500	0.500	0.293	0.343	0.370
2	10	0.456	0.500	0.500	0.128	0.187	0.222
	20	0.500	0.500	0.500	0.217	0.272	0.302
	30	0.500	0.500	0.500	0.265	0.313	0.339
5	10	0.191	0.262	0.304	0.011	0.025	0.036
	20	0.500	0.500	0.500	0.114	0.154	0.177
	30	0.500	0.500	0.500	0.185	0.226	0.249
10	20	0.257	0.315	0.347	0.005	0.012	0.018
	50	0.500	0.500	0.500	0.180	0.212	0.230
	100	0.500	0.500	0.500	0.294	0.322	0.336

Table 5.2: Comparison of quantile upper breakdown points for MM-regression estimates with 50% breakdown point.

the original data set. To drive the t-th robust bootstrap quantile estimate above any bound we need to have at least t% of the bootstrap samples containing less than p "good" points. The proportion ϵ of outliers in the original sample should then satisfy $P\big[\text{ Binomial}\,(n, 1-\epsilon) < p\ \big] \geq t.$ ■

The following table compares the breakdown point of the robust bootstrap quantile estimates with the equivalent classical bootstrap estimates. We considered an MM-regression estimate with 50% breakdown point and 95% efficiency when the data are normally distributed. We compared the quantiles needed to construct 90%, 95% and 99% confidence intervals for different sample sizes (n) and number of explanatory variables (p). Note that the only cases where the upper breakdown point for the robust bootstrap quantiles is significantly smaller than the breakdown point of the regression estimate (50%) is for $n = 10$, $p = 5$, and for $n = 20$, $p = 10$. These cases are not of interest from a practical point of view due to the extremely large dimension of the model relative to the number of observations available. Also note that our upper breakdown points are significantly larger than those of the classical bootstrap quantiles estimate.

5.5 Inference

5.5.1 Empirical coverage levels of confidence intervals

In this section we report the results of a Monte Carlo study on the finite sample properties of confidence intervals for the parameters $\boldsymbol{\beta}$ in the linear regression model (4.1). We considered sample sizes n = 30, 50 and 100 with 2 and 5 explanatory variables. These independent variables included an intercept: $x_1 \equiv 1$, and $x_i \sim N(0,1)$ for $i = 2, \ldots, p$. Finally, the errors followed the gross-error contamination model with distributions $F_\epsilon = (1-\epsilon)\,\Phi\,(x) + \epsilon V\,(x)$ where $V\,(x) = 0.5\,\Phi\,(\,(x-x_0)/\,0.1\,) + 0.5\,\Phi\,(\,(x+x_0)/\,0.1\,)$ and $\Phi\,(x)$ denotes the standard normal cumulative distribution function. We used $\epsilon = 0.00$, 0.10 and 0.20. The contamination point x_0 was set at 3, 4 and 10. We report the results obtained for $x_0 = 4$, the others being very similar.

We generated 5,000 data sets from the above distributions with $\epsilon = 0.00$, 0.10 and 0.20. We build 95% and 99% confidence intervals for the parameters of the model. We used MM-regression estimates obtained with $\psi = \rho'_{4.685}$ in Tukey's family. The S-scale was obtained with $\rho_{1.54764}$ also in Tukey's family. This choice yields estimates with simultaneous 50% breakdown point and 95% efficiency when the data are normally distributed.

We considered two methods to obtain confidence intervals for the regression coefficients. The first was the robust bootstrap as discussed above. We generated many re-calculated $\hat{\boldsymbol{\beta}}_n^{R*} - \hat{\boldsymbol{\beta}}_n$ and used the empirical distribution of each projection to obtain estimates of the distribution of $\hat{\boldsymbol{\beta}}_{n\,(j)} - \boldsymbol{\beta}_{(j)}$ for each coordinate $j = 1, \ldots, p$. With these distribution estimates we obtained the quantiles needed to build the confidence intervals of interest.

The second approach used the normal approximation (4.18) where we estimated the asymptotic variance with its empirical version $\hat{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma}\,(F_n, G_n, \hat{\boldsymbol{\beta}}_n, \hat{\sigma}_n)$, where F_n is the empirical distribution of the observed errors and G_n is the empirical distribution of the observed design.

Note that this estimate of the asymptotic variance is simpler (and hence numerically more stable) than the one given in Theorem 4.3 for an arbitrary distribution H. Because we know that the distribution generating the data is symmetric, we are confident that formula (4.18) is correct. Because of its numerical simplicity it is the best competitor to the robust bootstrap among (4.33) and (4.18).

In this context the classical bootstrap demands so much computer time that it becomes almost unfeasible; hence we did not include it in our study.

Tables 5.3 and 5.4 tabulate the results for $p = 2$. Tables 5.5 and 5.6 contain the corresponding findings for $p = 5$. Figures 5.5, 5.6, 5.7 and 5.8 display part of the results in a graphical form.

These pictures show at a glance that the levels obtained with the robust bootstrap are better than the ones yielded by the empirical asymptotic variance estimate. The difference in performance is more important for $p = 5$. Both methods are very close only for the case of $n = 100$ and $\epsilon \leq 0.10$ or for $n = 50$ and $\epsilon = 0.00$. The observed behaviour for the first scenario was expected because both methods are asymptotically equivalent, and hence behave similarly for large sample sizes. It is also reasonable not to observe large differences for $\epsilon = 0.00$ and moderate to large sample sizes, as the empirical asymptotic variance estimate is asymptotically correct when the errors follow the central model. Our method is more stable for smaller sample sizes, and yields more reasonable coverage levels for positive values of ϵ.

n	ϵ	Parameter	Robust bootstrap	Empirical AV
30	0.00	β_0	0.927 (0.748)	0.918 (0.720)
		β_1	0.925 (0.796)	0.921 (0.739)
	0.10	β_0	0.930 (0.953)	0.915 (0.720)
		β_1	0.933 (1.057)	0.908 (0.892)
	0.20	β_0	0.945 (1.408)	0.921 (1.215)
		β_1	0.943 (1.550)	0.914 (1.247)
50	0.00	β_0	0.932 (0.572)	0.928 (0.562)
		β_1	0.928 (0.591)	0.935 (0.571)
	0.10	β_0	0.938 (0.716)	0.930 (0.684)
		β_1	0.939 (0.755)	0.926 (0.697)
	0.20	β_0	0.953 (1.037)	0.938 (0.962)
		β_1	0.950 (1.119)	0.925 (0.977)
100	0.00	β_0	0.943 (0.404)	0.942 (0.400)
		β_1	0.936 (0.410)	0.939 (0.404)
	0.10	β_0	0.951 (0.501)	0.946 (0.490)
		β_1	0.942 (0.514)	0.941 (0.495)
	0.20	β_0	0.954 (0.713)	0.948 (0.690)
		β_1	0.949 (0.742)	0.938 (0.695)

Table 5.3: Average coverage and length of 5,000 95% confidence intervals for the linear regression model with $p = 2$

n	ϵ	Parameter	Robust bootstrap	Empirical AV
30	0.00	β_0	0.975 (0.984)	0.972 (0.947)
		β_1	0.974 (1.047)	0.972 (0.973)
	0.10	β_0	0.979 (1.255)	0.973 (1.143)
		β_1	0.977 (1.391)	0.970 (1.174)
	0.20	β_0	0.984 (1.854)	0.977 (1.599)
		β_1	0.980 (2.040)	0.966 (1.641)
50	0.00	β_0	0.980 (0.752)	0.979 (0.740)
		β_1	0.976 (0.777)	0.981 (0.751)
	0.10	β_0	0.984 (0.942)	0.982 (0.901)
		β_1	0.983 (0.994)	0.978 (0.917)
	0.20	β_0	0.989 (1.365)	0.986 (1.266)
		β_1	0.989 (1.473)	0.977 (1.286)
100	0.00	β_0	0.986 (0.531)	0.986 (0.527)
		β_1	0.984 (0.540)	0.987 (0.532)
	0.10	β_0	0.989 (0.659)	0.987 (0.645)
		β_1	0.984 (0.677)	0.985 (0.651)
	0.20	β_0	0.991 (0.938)	0.989 (0.908)
		β_1	0.990 (0.977)	0.983 (0.915)

Table 5.4: Coverage and length of 5,000 99% confidence intervals for the linear regression model with $p = 2$

Table 5.5: Coverage and length of 5,000 95% confidence intervals for the linear regression model with $p = 5$

n	ϵ	Parameter	Robust bootstrap	Empirical AV
30	0.00	β_0	0.924 (1.018)	0.829 (0.684)
		β_1	0.920 (1.070)	0.833 (0.702)
		β_2	0.921 (1.071)	0.835 (0.702)
		β_3	0.917 (1.077)	0.831 (0.701)
		β_4	0.915 (1.060)	0.828 (0.702)
30	0.10	β_0	0.940 (1.248)	0.837 (0.780)
		β_1	0.936 (1.332)	0.835 (0.797)
		β_2	0.932 (1.330)	0.829 (0.801)
		β_3	0.935 (1.328)	0.828 (0.798)
		β_4	0.932 (1.337)	0.822 (0.800)
30	0.20	β_0	0.950 (1.917)	0.825 (1.058)
		β_1	0.930 (2.006)	0.809 (1.085)
		β_2	0.929 (1.983)	0.806 (1.082)
		β_3	0.935 (2.042)	0.812 (1.084)
		β_4	0.936 (2.057)	0.808 (1.084)
50	0.00	β_0	0.946 (0.644)	0.912 (0.559)
		β_1	0.945 (0.669)	0.918 (0.567)
		β_2	0.938 (0.671)	0.911 (0.567)
		β_3	0.938 (0.669)	0.906 (0.568)
		β_4	0.940 (0.672)	0.911 (0.566)
50	0.10	β_0	0.957 (0.814)	0.912 (0.653)
		β_1	0.950 (0.851)	0.901 (0.664)
		β_2	0.952 (0.864)	0.895 (0.664)
		β_3	0.950 (0.858)	0.900 (0.663)
		β_4	0.949 (0.848)	0.900 (0.664)
50	0.20	β_0	0.960 (1.306)	0.904 (0.930)
		β_1	0.951 (1.387)	0.877 (0.947)
		β_2	0.952 (1.384)	0.880 (0.942)
		β_3	0.949 (1.366)	0.888 (0.944)

continued on next page

continued from previous page

n	ϵ	Parameter	Robust bootstrap	Empirical AV
		β_4	0.955 (1.375)	0.890 (0.941)
100	0.00	β_0	0.944 (0.421)	0.935 (0.400)
		β_1	0.945 (0.428)	0.935 (0.402)
		β_2	0.939 (0.427)	0.935 (0.403)
		β_3	0.948 (0.428)	0.939 (0.403)
		β_4	0.941 (0.427)	0.938 (0.403)
100	0.10	β_0	0.954 (0.531)	0.932 (0.482)
		β_1	0.950 (0.544)	0.927 (0.486)
		β_2	0.948 (0.546)	0.925 (0.485)
		β_3	0.950 (0.547)	0.927 (0.485)
		β_4	0.950 (0.545)	0.926 (0.485)
100	0.20	β_0	0.960 (0.797)	0.935 (0.686)
		β_1	0.968 (0.828)	0.936 (0.692)
		β_2	0.956 (0.832)	0.919 (0.691)
		β_3	0.960 (0.828)	0.916 (0.691)
		β_4	0.957 (0.830)	0.919 (0.691)

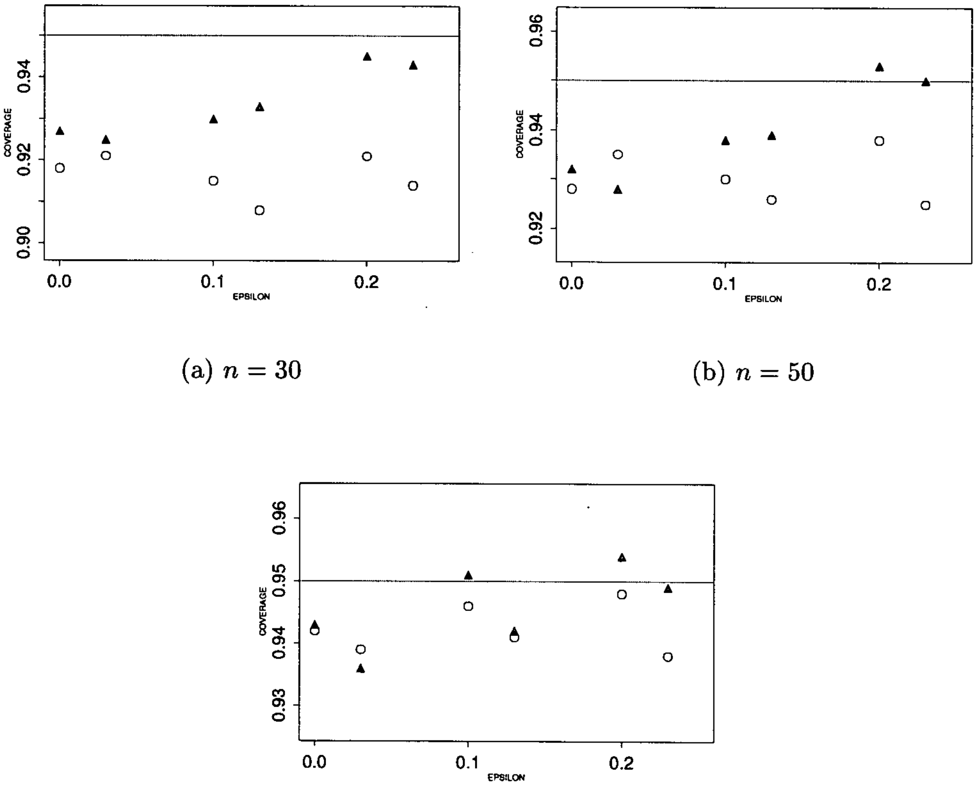
Table 5.6: Coverage and length of 5,000 99% confidence intervals for the linear regression model with $p = 5$

n	ϵ	Parameter	Robust bootstrap	Empirical AV
30	0.00	β_0	0.967 (1.340)	0.911 (0.901)
		β_1	0.963 (1.408)	0.912 (0.924)
		β_2	0.963 (1.410)	0.913 (0.923)
		β_3	0.963 (1.417)	0.907 (0.923)
		β_4	0.963 (1.395)	0.908 (0.924)
30	0.10	β_0	0.979 (1.643)	0.923 (1.027)
		β_1	0.974 (1.753)	0.917 (1.049)
		β_2	0.973 (1.751)	0.913 (1.054)
		β_3	0.973 (1.748)	0.910 (1.050)
		β_4	0.971 (1.760)	0.908 (1.052)
30	0.20	β_0	0.983 (2.523)	0.917 (1.393)
		β_1	0.973 (2.641)	0.895 (1.429)
		β_2	0.973 (2.610)	0.901 (1.424)
		β_3	0.978 (2.688)	0.901 (1.427)
		β_4	0.974 (2.707)	0.898 (1.427)
50	0.00	β_0	0.985 (0.847)	0.973 (0.735)
		β_1	0.984 (0.881)	0.971 (0.747)
		β_2	0.983 (0.883)	0.970 (0.747)
		β_3	0.981 (0.880)	0.970 (0.748)
		β_4	0.985 (0.885)	0.974 (0.745)
50	0.10	β_0	0.988 (1.072)	0.970 (0.860)
		β_1	0.986 (1.121)	0.965 (0.874)
		β_2	0.987 (1.138)	0.963 (0.874)
		β_3	0.989 (1.130)	0.971 (0.873)
		β_4	0.986 (1.116)	0.964 (0.874)
50	0.20	β_0	0.992 (1.719)	0.968 (1.224)
		β_1	0.990 (1.826)	0.954 (1.246)
		β_2	0.987 (1.821)	0.949 (1.239)
		β_3	0.987 (1.798)	0.955 (1.242)

continued on next page

continued from previous page

n	ϵ	Parameter	Robust bootstrap	Empirical AV
		β_4	0.988 (1.810)	0.957 (1.238)
100	0.00	β_0	0.988 (0.555)	0.983 (0.526)
		β_1	0.986 (0.563)	0.986 (0.530)
		β_2	0.984 (0.562)	0.982 (0.530)
		β_3	0.987 (0.564)	0.985 (0.530)
		β_4	0.988 (0.562)	0.985 (0.531)
100	0.10	β_0	0.990 (0.699)	0.983 (0.635)
		β_1	0.989 (0.717)	0.981 (0.640)
		β_2	0.988 (0.719)	0.981 (0.639)
		β_3	0.990 (0.720)	0.980 (0.639)
		β_4	0.990 (0.718)	0.980 (0.638)
100	0.20	β_0	0.994 (1.050)	0.984 (0.903)
		β_1	0.993 (1.090)	0.982 (0.911)
		β_2	0.990 (1.095)	0.978 (0.910)
		β_3	0.992 (1.090)	0.974 (0.910)
		β_4	0.992 (1.093)	0.978 (0.910)

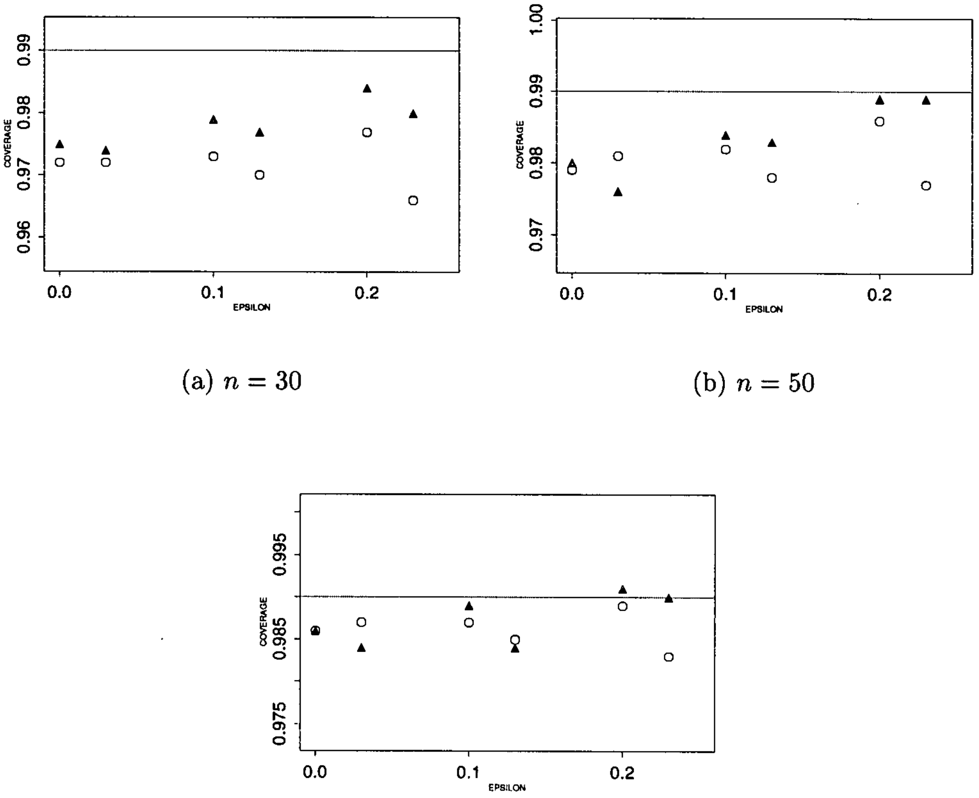


(a) $n = 30$

(b) $n = 50$

(c) $n = 100$

Figure 5.5: Average coverage of 95% confidence intervals for the linear regression model with $p = 2$. Solid triangles are levels of the confidence intervals for the intercept and the coefficient of $\mathbf{x}_1$ calculated with the robust bootstrap; circles represent the corresponding levels for the confidence intervals obtained with the empirical asymptotic variance estimate. Across the horizontal axis, the three groups correspond to $\epsilon = 0.0$, 0.1 and 0.2 respectively. The horizontal line indicates the nominal level.

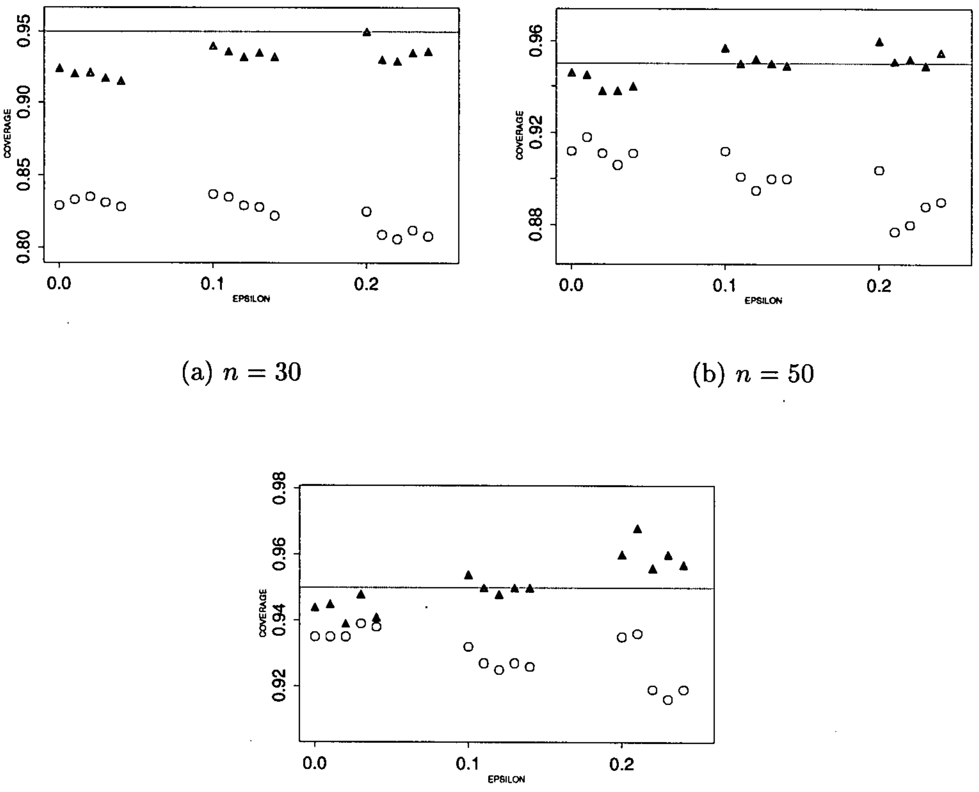


(a) $n = 30$

(b) $n = 50$

(c) $n = 100$

Figure 5.6: Average coverage of 99% confidence intervals for the linear regression model with $p = 2$. Solid triangles are levels of the confidence intervals for the intercept and the coefficient of $\mathbf{x}_1$ calculated with the robust bootstrap; circles represent the corresponding levels for the confidence intervals obtained with the empirical asymptotic variance estimate. Across the horizontal axis, the three groups correspond to $\epsilon = 0.0$, 0.1 and 0.2 respectively. The horizontal line indicates the nominal level.

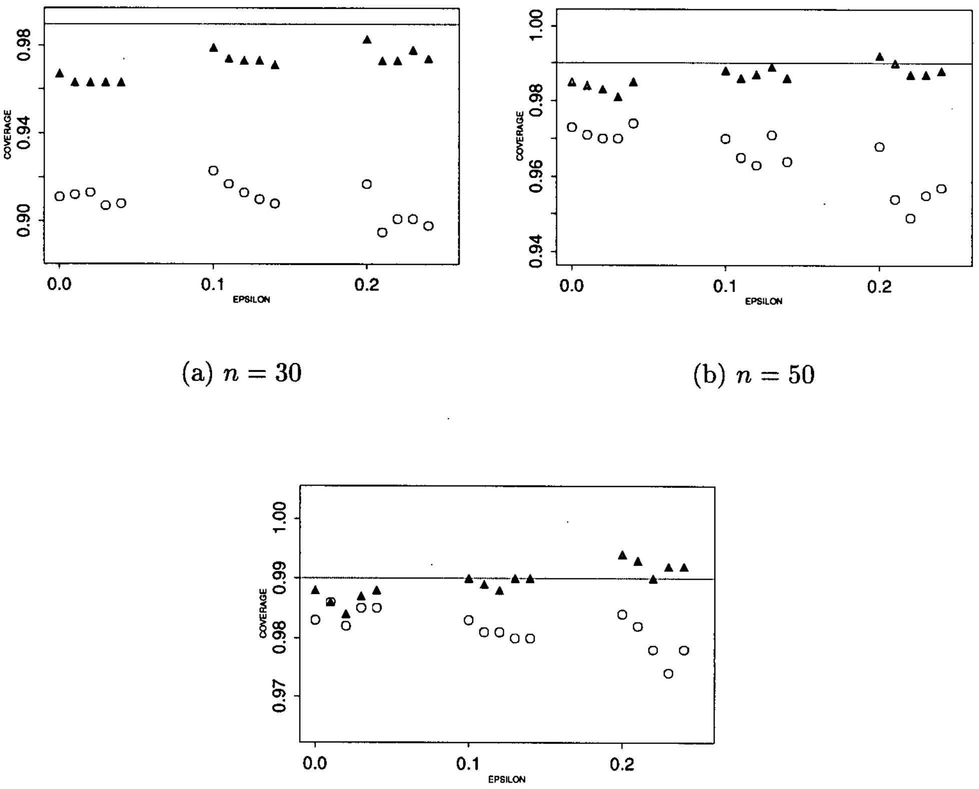


(a) $n = 30$

(b) $n = 50$

(c) $n = 100$

Figure 5.7: Average coverage of 95% confidence intervals for the linear regression model with $p = 5$. Solid triangles are levels of the confidence intervals for the intercept and the coefficients of $\mathbf{x}_1, \ldots, \mathbf{x}_4$ calculated with the robust bootstrap; circles represent the corresponding levels for the confidence intervals obtained with the empirical asymptotic variance estimate. Across the horizontal axis, the three groups correspond to $\epsilon = 0.0$, 0.1 and 0.2 respectively. The horizontal line indicates the nominal level.



(a) $n = 30$

(b) $n = 50$

(c) $n = 100$

Figure 5.8: Average coverage of 99% confidence intervals for the linear regression model with $p = 5$. Solid triangles are levels of the confidence intervals for the intercept and the coefficients of $\mathbf{x}_1, \ldots, \mathbf{x}_4$ calculated with the robust bootstrap; circles represent the corresponding levels for the confidence intervals obtained with the empirical asymptotic variance estimate. Across the horizontal axis, the three groups correspond to $\epsilon = 0.0$, 0.1 and 0.2 respectively. The horizontal line indicates the nominal level.

Chapter 6

Conclusion

This chapter contains an outline of the results obtained in the thesis, the problems that we encountered, and the directions we foresee for future work.

- **Global asymptotic properties of robust estimates**:

  ❑ We established the consistency and asymptotic normality of robust location and regression estimates for an arbitrary distribution function in the contamination neighbourhood. Under additional regularity conditions, we showed that the consistency and asymptotic normality of the robust estimates for the location model hold uniformly on the contamination neighborhood for certain proportions of contamination. There seems to be a trade-off between the breakdown point of the estimate and the size of the neighbourhood where it is uniformly consistent.

- ❑ It is desirable to find regularity conditions on the loss function ρ that will ensure a unique minimum of the functional that defines the estimate for any distribution in the contamination neighbourhood.

- ❑ It remains to study whether the asymptotic properties of the robust regression estimates hold uniformly on the contamination neighborhood. We conjecture that this is the case and we anticipate some technical difficulties due to the multivariate nature of the problem. We also expect that the required regularity conditions on the function ρ will be more strict than those found for the location model.

- **Maximum asymptotic bias calculation for location estimates**:

  - ❑ As a byproduct of our computations regarding the uniform consistency of the S-location estimate (see Section 2.3.3), we derived a method to determine the maximum asymptotic bias for location estimates calculated with a re-descending function ψ. To our knowledge there are no results in the literature regarding the maximum asymptotic bias of this type of estimates.

- **A stable and feasible computer intensive inference method**:

  - ❑ We introduced a new computer intensive method to perform statistical inference based on robust estimates. This method, which we call the robust bootstrap, can in principle be used on any statistical model where residuals are well defined. We studied its theoretical and practical properties when it is applied it to estimate the variability and the sampling distribution of robust estimates for the location and regression models. We found that the robust bootstrap is computationally simpler than the classical bootstrap and that it is more stable when the data are contaminated. In particular, it yields estimates of the quantiles of the distribution of the location and regression estimators that have a higher breakdown point than those obtained from the classical bootstrap.

❑ Robust regression estimates for the linear model with fixed design remain to be studied in detail. We expect the robust bootstrap to apply as described in Section 3.5. Note that if the design is fixed then the contamination model does not contemplate outliers in the covariates. Hence, we can use a M-regression estimate with a monotone function ψ. This change may modify the robustness properties of the quantile estimates.

❑ We have not been able to show that the Studentized robust bootstrap converges faster than $O_P\left(1/\sqrt{n}\right)$. Our proof seems to fail because the correction factor is of that order. A possible solution is to perform a second order Taylor expansion when deriving the correction factor for the robust bootstrap. This deserves further study.

❑ We are interested in extending the robust bootstrap to estimators that are defined by estimating equations of the form

$$\sum_{i=1}^{n} g_i\left(y_i, \mathbf{x}_i, \hat{\boldsymbol{\theta}}_n, \hat{\boldsymbol{\nu}}_n\right) = \mathbf{0}\,,$$

where y_i are the response variables, $\mathbf{x}_i$ are vectors of covariates, $\hat{\boldsymbol{\theta}}_n$ is the estimator of interest and $\hat{\boldsymbol{\nu}}_n$ is an estimator of nuisance parameters.

We expect the robust bootstrap to be computationally simpler than the classical bootstrap. Its robustness properties (stability and breakdown point of the quantile estimates) will depend on the form of the functions g_i.

❑ It is also of interest to determine whether the convergence of the distribution of the robust bootstrap holds uniformly on the contamination neighbourhood. If this is the case we will have an estimation method for the sampling distribution of the estimate of interest that is uniformly close to its target. This would be a very interesting result to obtain.

Chapter 7

Appendix - Auxiliary results

This chapter contains auxiliary results used throughout this thesis. Proofs are presented for those that could not be found in the literature.

Auxiliary results found in the literature

Definition 7.1 - Big O in probability : *Let $a_n, n = 1, \ldots$ be a sequence of real numbers and let $X_n, n = 1, \ldots$ be a sequence of random variables. We say that $X_n = O_P(a_n)$ if*

$$\lim_{k\to\infty} \lim_{n\to\infty} P\left[\left| \frac{X_n}{a_n} \right| > k \right] = 0.$$

That is, the sequence $|X_n / a_n|$ is bounded in probability.

Definition 7.2 - Small o in probability: *Let $a_n, n = 1, \ldots$ be a sequence of real* *numbers and let $X_n, n = 1, \ldots$ be a sequence of random variables. We say that $X_n = o_P(a_n)$ if $\forall\, \delta > 0$*

$$\lim_{n\to\infty} P\left[\left|\frac{X_n}{a_n}\right| > \delta\right] = 0.$$

That is, the sequence $|X_n/\,a_n|$ converges to zero in probability.

It is easy to see that the following three implications hold

$$X_n = O_P(1)\,, \quad Y_n = o_P(1) \quad \Rightarrow \quad X_n + Y_n = O_P(1)\,, \tag{7.1}$$

$$X_n = O_P(1)\,, \quad Y_n = O_P(1) \quad \Rightarrow \quad X_n \times Y_n = O_P(1)\,, \tag{7.2}$$

and

$$X_n = O_P(1)\,, \quad Y_n = o_P(1) \quad \Rightarrow \quad X_n \times Y_n = o_P(1)\,. \tag{7.3}$$

Remark 7.1 - The above definition of $O_P(a_n)$ is equivalent to the following definition, also found in the literature (see Davison and Hinkley, 1997, page 39): A sequence of random variables X_n is said to be $O_P(a_n)$ if, for each $\epsilon > 0$ we have $\lim_{n\to\infty} P(|X_n/\,a_n| > \epsilon) = p$, a constant. It is clear that our first definition implies the latter. To see the other implication first note that $p = p(\epsilon)$ above is a non-increasing function of $\epsilon > 0$. This is a consequence of the following inequality that holds for each $n \in \mathbb{N}$

$$P\left(\left|\frac{X_n}{a_n}\right| > \epsilon_2\right) \leq P\left(\left|\frac{X_n}{a_n}\right| > \epsilon_1\right) \qquad \epsilon_1 < \epsilon_2\,.$$

We will now show that $\lim_{\epsilon\to\infty} p(\epsilon) = 0$. Because for any ϵ we have $p(\epsilon) \geq 0$ it is enough to show that we cannot have $p(\epsilon) \geq \tilde{p} > 0$ for some fixed $\tilde{p}$. If such a $\tilde{p}$ existed we would have that for any $\delta > 0$ there exists a $n_0 \in \mathbb{N}$ such that for any $n \geq n_0$,

$$P\left(\left|\frac{X_n}{a_n}\right| > \epsilon\right) > \tilde{p} - \delta \qquad \forall \epsilon > 0\,.$$

Because for each fixed $n \in \mathbb{N}$ the left hand side converges to zero as ϵ increases, this is a contradiction. We conclude that $p\left(\epsilon\right) \to 0$ as ϵ goes to infinity. Hence both definitions of $O_P\left(a_n\right)$ are equivalent.

Definition 7.3 - **Infinitely often** - *Let A_n be a sequence of subsets of a probability space Ω. The event $\{A_n$ infinitely often$\}$ is defined by*

$$\{A_n \textit{ infinitely often}\} = \bigcap_{m=1}^{\infty} \bigcup_{n=m}^{\infty} A_n\,.$$

We will also write $\{A_n$ i.o.$\}$

Lemma 7.1 - **Serfling** - *(Serfling, 1980, page 253) - Let $X_1, \ldots, X_n$ be a sequence of independent identically distributed random variables and let $g\left(x,t\right) : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ be continuous in t uniformly on $x \in \mathbb{R}$. Let θ_n be a sequence of random variables such that $\theta_n \xrightarrow[n\to\infty]{P} \theta$, a constant, then*

$$\frac{1}{n}\sum_{i=1}^{n} g\left(X_i, \theta_n\right) \xrightarrow[n\to\infty]{P} E\left[g\left(X, \theta\right)\right]. \tag{7.4}$$

If $\theta_n \xrightarrow[n\to\infty]{a.s.} \theta$ then (7.4) holds a.s. as well.

Lemma 7.2 - **Bernstein** - *(Chow and Teicher, 1988, page 111) Let $S_n = \sum_{i=1}^{n} X_i$ where X_i are independent random variables with $E\left(X_i\right) = 0$, $E\left(X_i^2\right) = \sigma_i^2$. Let $s_n = \sum_{i=1}^{n} \sigma_i^2 > 0$. Assume that $E\left|X_i\right|^k \leq \left(k!/2\right) \sigma_i^2\, c^{k-2}$ for $k > 2$ and $0 < c < \infty$, then*

$$P\left[S_n > x\right] \leq \exp\left(\frac{-x^2}{2\left(s_n^2 + c\,x\right)}\right), \qquad x > 0. \tag{7.5}$$

The condition $E\left|X_i\right|^k \leq (k!/2)\, \sigma_i^2\, c^{k-2}$ *holds if, for example,* $P\left[\left|X_i\right| \leq c\right] = 1$.

Theorem 7.1 - **Berry-Esseen for i.i.d. random variables** - *(Chow and Teicher, 1988, page 305) If* $\{X_n, n \geq 1\}$ *are i.i.d. random variables with* $EX_i = 0$, $EX_i^2 = \sigma^2$, $E\left|X_i\right|^{2+\delta} = \gamma^{2+\delta} < \infty$ *for some* $\delta \in (0,1]$, $S_n = \sum_{i=1}^n X_i$ *and* Φ *is the standard normal distribution function, then there exists a universal constant* c_δ *such that*

$$\sup_{x\in\mathbb{R}} \left|P\left\{S_n < x\,\sigma\, n^{1/2}\right\} - \Phi\left(x\right)\right| \leq \frac{c_\delta}{n^{\delta/2}} \left(\frac{\gamma}{\sigma}\right)^{2+\delta}.$$

Lemma 7.3 - **Borel-Cantelli** *(see for example Chung, 1974, page 73) Let* $\left\{A_n\right\}_{n\in\mathbb{N}}$ *be a sequence of events. Then*

$$\sum_{i=1}^n P\left(A_i\right) < \infty \quad \Rightarrow \quad P\left(A_n \text{ } i.o.\right) = 0.$$

Lemma 7.4 - *Lemma 3.4 in Yohai (1985) - If* $P\left(\left|\boldsymbol{\theta}'\mathbf{X}\right| > 0\right) > \lambda$ *for all* $\boldsymbol{\theta} \in \mathbb{R}^p$, *then there exist* $\phi > 0$, $\delta > 0$, $\gamma > 0$ *and a finite collection of compact sets* $\mathcal{C}_1, \ldots, \mathcal{C}_s$ *such that*

$$\bigcup_{i=1}^s \mathcal{C}_i \supset \left\{\boldsymbol{\theta} \in \mathbb{R}^p : \|\boldsymbol{\theta}\| = 1\right\}$$

and

$$P\left(\inf_{\boldsymbol{\theta}\in\mathcal{C}_i} \left|\boldsymbol{\theta}'\mathbf{X}\right| \geq \phi\right) \geq \lambda + \gamma\,.$$

Lemma 7.5 - *Lemma 4.2 in Yohai (1985) - Let* $g : \mathbb{R}^k \times \mathbb{R}^h \to \mathbb{R}$ *be continuous and let* Q *be a probability measure on* $\mathbb{R}^k$ *such that for some* $\delta > 0$ *we have*

$$E_Q\left[\sup_{\|\boldsymbol{\gamma}-\boldsymbol{\gamma}_0\|\leq\delta} \left|g\left(\mathbf{z}, \boldsymbol{\gamma}\right)\right|\right] < \infty\,.$$

Let $\hat{\gamma}_n$ be a sequence of estimates in $\mathbb{R}^k$ such that $\hat{\gamma}_n \to \gamma_0$ almost surely. Then if $\mathbf{z}_1, \ldots, \mathbf{z}_n$ are independent identically distributed random variables in $\mathbb{R}^k$ with distribution Q, we have

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} g\left(\mathbf{z}_i, \hat{\gamma}_n\right) = E_Q\left[g\left(\mathbf{z}, \gamma_0\right)\right], \qquad a.s.$$

Auxiliary results not found in the literature

The following lemma is an immediate consequence of Lemma 7.2. For completeness we state and prove it here.

Lemma 7.6 *Assume that X_i for $i = 1, \ldots, n$ are independent random variables with zero mean such that there exists a constant c with $P\left(|X_i| \le c\right) = 1$. Let $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Then, for any $\delta > 0$ we have*

$$P\left(\left|\overline{X}_n\right| > \delta\right) \le 2\exp\left(-n\gamma\right) \qquad \text{for some } \gamma = \gamma\left(c, \delta\right) > 0\,.$$

Proof: Let $\sigma_i^2 = V\left(X_i\right)$, $s_n^2 = \sum_{i=1}^{n} \sigma_i^2$, and $S_n = \sum_{i=1}^{n} X_i$. By Lemma 7.2 we have

$$\begin{aligned} P\left(\left|\overline{X}_n\right| > \delta\right) &= P\left(\overline{X}_n > \delta\right) + P\left(\overline{X}_n < -\delta\right) \\ &= P\left(S_n > n\delta\right) + P\left(S_n < -n\delta\right) \\ &\le \exp\left(\frac{-n^2\delta^2}{2\left(s_n^2 + cn\delta\right)}\right) + P\left(-S_n > n\delta\right) \\ &\le 2\exp\left(\frac{-n^2\delta^2}{2\left(s_n^2 + cn\delta\right)}\right), \end{aligned}$$

and because $s_n^2 \leq nc^2$ we have

$$\leq 2\exp\left(-n\,\frac{\delta^2}{2\,(c^2+c\delta)}\right) = 2\exp\left(-n\,\gamma\right)\,,$$

where $\gamma = \gamma\,(c,\delta)$. ■

Lemma 7.7 *Let $\rho : \mathbb{R} \longrightarrow \mathbb{R}^+$ be a continuous real function such that there exists a finite constant c with $\rho\,(u) = 1$ for $|u| \geq c$. Let $t \in \mathcal{T}$ and $s \in \mathcal{S}$, where $\mathcal{T}$ and $\mathcal{S}$ are bounded real intervals, with* $\inf\{s \in \mathcal{S}\} > 0$. *Then the function*

$$f\,(u,t,s) = \rho\left(\frac{u-t}{s}\right), \quad u \in \mathbb{R},\ t \in \mathcal{T}\ s \in \mathcal{S}$$

is continuous in s and t uniformly in u. In other words, for any $\epsilon > 0$, there exist $\delta_t > 0$ and $\delta_s > 0$ such that

$$|s_1 - s_2| < \delta_s, \quad |t_1 - t_2| < \delta_t \quad \Rightarrow\ |f\,(u,t_1,s_1) - f\,(u,t_2,s_2)| < \epsilon,\ \forall u \in \mathbb{R}.$$

Proof: The idea is to show that there exists a closed and bounded interval $\mathcal{U}$ such that for any $t_1, t_2 \in \mathcal{T}$, $s_1, s_2 \in \mathcal{S}$ we have

$$\rho\left(\frac{u-t_1}{s_1}\right) = \rho\left(\frac{u-t_2}{s_2}\right), \quad u \notin \mathcal{U}, \tag{7.6}$$

whereas for $u \in \mathcal{U}$ we will use a standard ϵ-δ argument. By hypothesis we have $\underline{t} \leq t \leq \bar{t}$ and $\underline{s} \leq s \leq \bar{s}$. Consider $u \geq \bar{t} + \bar{s}\,c$. It is clear that, for any $t \leq \bar{t}$ and $s \leq \bar{s}$ we have $(u-t)/\,s \geq c$. Hence (7.6) holds. Similarly, for $u \leq \underline{t} - \underline{s}\,c$ we have that for any $t \geq \underline{t}$ and $s \geq \underline{s}$ we have $(u-t)/\,s \leq -c$ and (7.6) holds as well. So,

$\mathcal{U} = [\underline{t} - \underline{s}\, c\ ,\ \bar{t} + \bar{s}\, c]$. In $\mathcal{U}$ ρ is uniformly continuous and hence it is enough to bound

$$\left| \frac{u - t_1}{s_1} - \frac{u - t_2}{s_2} \right| = |u| \, \frac{|s_2 - s_1|}{s_1\, s_2} + \frac{|t_2\, s_1 - t_1\, s_2|}{s_1\, s_2} = |u| \, \frac{|s_2 - s_1|}{s_1\, s_2} + $$
$$+ |t_2| \, \frac{|s_2 - s_1|}{s_1\, s_2} + s_2 \frac{|t_2 - t_1|}{s_1\, s_2} \leq K_{\mathcal{U}} \frac{|s_2 - s_1|}{\underline{s}^2} + \bar{t} \frac{|s_2 - s_1|}{\underline{s}^2} + \frac{|t_2 - t_1|}{\underline{s}^2} < \delta,$$

if $|s_1 - s_2|$ and $|t_2 - t_1|$ are sufficiently small ($K_{\mathcal{U}} = \sup\{|u| : u \in \mathcal{U}\}$). ■

Lemma 7.8 *If*

$$E_F \left| \rho' \left(\frac{X - t}{s} \right) \left(\frac{X - t}{s} \right) \right| \leq K, \qquad \forall\, t \in \mathbb{R}, \quad \forall\, s \in \mathcal{K}_s, \quad \forall\, F \in \mathcal{H}_\epsilon,$$

then $\gamma\,(F, t, s)$ is continuous in s uniformly in t and F, i.e., for any $\epsilon > 0$, there exists $\delta > 0$ independent of t and F such that

$$|s_1 - s_2| < \delta \quad \Rightarrow \quad |\gamma\,(F, t, s_1) - \gamma\,(F, t, s_2)| < \epsilon, \qquad \forall\, t \in \mathbb{R}, \quad \forall\, F \in \mathcal{H}_\epsilon.$$

Proof: A Taylor expansion yields

$$\rho \left(\frac{X - t}{s_1} \right) - \rho \left(\frac{X - t}{s_2} \right) = -\frac{1}{\tilde{s}} \rho' \left(\frac{X - t}{\tilde{s}} \right) \left(\frac{X - t}{\tilde{s}} \right) (s_1 - s_2),$$

where $s_1 \leq \tilde{s} \leq s_2$. Hence,

$$\left| E_F \rho \left(\frac{X - t}{s_1} \right) - E_F \rho \left(\frac{X - t}{s_2} \right) \right| \leq E_F \left| \rho \left(\frac{X - t}{s_1} \right) - \rho \left(\frac{X - t}{s_2} \right) \right|$$
$$= \frac{1}{\tilde{s}} E_F \left| \rho' \left(\frac{X - t}{\tilde{s}} \right) \left(\frac{X - t}{\tilde{s}} \right) \right| |s_1 - s_2| \leq \frac{K\, |s_1 - s_2|}{s^-} < \epsilon$$

if $|s_1 - s_2| < s^- \epsilon / K$. Note that by hypothesis K does not depend on t or F. ■

Lemma 7.9 *If ρ belongs to Tukey's family and $\mathcal{H}_\epsilon$ is a contamination neighbourhood around the standard normal distribution, then Lemma 7.8 holds.*

Proof: Note that $\rho_d^{'}(u)\, u \geq 0$. Hence we have to find a uniform bound for

$$\int_{\left|\frac{X-t}{s}\right|\leq d} \frac{1}{d}\left(\frac{X-t}{s}\right)^2 \left(1-\left(\frac{X-t}{s}\right)^2\right)^2 f(X)\, dX.$$

It is enough to bound

$$\int_{-d}^{d} \frac{1}{d} u^2 \left(1-u^2\right)^2 \phi\left(t+s\,u\right) du,$$

and

$$\int_{-d}^{d} \frac{1}{d} u^2 \left(1-u^2\right)^2 dH\left(t+s\,u\right).$$

The first integral is bounded because ϕ is. Also note that

$$\begin{aligned}\int_{-d}^{d} \frac{1}{d} u^2 \left(1-u^2\right)^2 dH\left(t+s\,u\right) &\leq K\,\frac{1}{d}\int_{-d}^{d} dH\left(t+s\,u\right) \\ &= \frac{s}{d} K \int_{t-s\,d}^{t+s\,d} dH\left(x\right) \leq \frac{\bar{s}}{d} K.\end{aligned}$$

■

Lemma 7.10 *If $\rho(x,t,\sigma)$ is continuous in t, then, for fixed x and σ we have*

$$\inf_{t\in B(t_0)} \rho\left(x,t,\sigma\right) \xrightarrow[B(t_0)\searrow\{t_0\}]{} \rho\left(x,t_0,\sigma\right),$$

where $B(t_0)$ is an open ball around t_0. If, in addition, $\rho(x,t,\sigma)$ is bounded, we have

$$E_F\left[\inf_{t\in B(t_0)} \rho\left(X,t,\sigma\right)\right] \xrightarrow[B(t_0)\searrow\{t_0\}]{} E_F\rho\left(X,t_0,\sigma\right).$$

Proof: Fix $\epsilon > 0$. By continuity there exists $\delta = \delta\left(x,\sigma,t_0\right) > 0$ such that

$$\left|t-t_0\right| < \delta \;\Rightarrow\; \left|\rho\left(x,t,\sigma\right)-\rho\left(x,t_0,\sigma\right)\right| < \epsilon.$$

Hence, $\rho(x,t,\sigma) < \rho(x,t_0,\sigma)+\epsilon$ for all t in a sufficiently small neighbourhood $B(t_0)$ of t_0. Immediately we obtain

$$\inf_{t\in B(t_0)} \rho(x,t,\sigma) \leq \rho(x,t_0,\sigma)+\epsilon.$$

Similarly we have

$$\inf_{t\in B(t_0)} \rho(x,t,\sigma) \geq \rho(x,t_0,\sigma)-\epsilon,$$

and the proof of the first claim is complete. The second claim is a consequence of the Dominated Convergence Theorem. ■

Lemma 7.11 *Let $x_1,\ldots,x_n$ be i.i.d. random variables, $x_i \sim F$. If ρ satisfies the conditions of Lemma 7.7 and $\hat{\sigma}_n \to \boldsymbol{\sigma}$ a.s. [F], then for any t_0 and bounded neighbourhood $B(t_0)$ we have*

$$\frac{1}{n}\sum_{i=1}^{n} \inf_{t'\in B(t_0)} \rho(x_i,t',\hat{\sigma}_n) \xrightarrow[n\to\infty]{a.s.} E_F\left[\inf_{t'\in B(t_0)} \rho(X,t',\boldsymbol{\sigma})\right]. \tag{7.7}$$

Proof: By Lemma 7.1 it is enough to show that the function

$$f(x,\sigma) = \inf_{t'\in B(t_0)} \rho(x,t',\sigma)$$

is continuous in σ uniformly in x. Fix $\epsilon > 0$. The proof of Lemma 7.7 shows that there exists $\delta > 0$ such that if $|s_1 - s_2| < \delta$ then

$$\left|\rho\left(\frac{x-t}{\sigma_1}\right) - \rho\left(\frac{x-t}{\sigma_2}\right)\right| < \epsilon, \qquad \text{if } |\sigma_1-\sigma_2|<\delta, \quad \forall t \in \overline{B(t_0)}, \quad \forall x \in \mathbb{R},$$

where $\overline{A}$ denotes the completion of A. Note that δ does not depend on t (although it does depend on $\overline{B(t_0)}$). We have

$$\rho\left(\frac{x-t}{s_1}\right) \leq \rho\left(\frac{x-t}{s_2}\right) + \epsilon.$$

It follows then that $\inf_{t\in B(t_0)} \rho\left(\frac{x-t}{s_1}\right) \leq \inf_{t\in B(t_0)} \rho\left(\frac{x-t}{s_2}\right) + \epsilon$. The same argument can be applied to the other inequality to obtain

$$\left| \inf_{t\in B(t_0)} \rho\left(\frac{x-t}{s_1}\right) - \inf_{t\in B(t_0)} \rho\left(\frac{x-t}{s_2}\right) \right| \leq \epsilon,$$

for all x. Hence (7.7) holds. ■

Lemma 7.12 *Let X_n and Y_n, $n = 1, \ldots$ be two sequences of random variables such that $X_n = O_P(1)$ and $Y_n = O_P(1)$. Then*

$$\frac{X_n}{Y_n + o_P(1)} = \frac{X_n}{Y_n} + o_P(1)$$

Proof: The result follows immediately from (7.1)-(7.3) and

$$\frac{X_n}{Y_n + o_P(1)} - \frac{X_n}{Y_n} = \frac{-X_n\, o_P(1)}{(Y_n + o_P(1))\, Y_n}.$$

■

The following lemma is an elemental result. We state and prove it here for completeness of the presentation.

Lemma 7.13 *Let $f : \mathbb{R} \to \mathbb{R}$ be a continuous function. If $\lim_{|x|\to\infty} f(x) = 0$, then f is uniformly continuous in $\mathbb{R}$.*

Proof: Let $\epsilon > 0$. Choose $K(\epsilon)$ such that if $|x| > K$, $|f(x)| < \epsilon/4$. For the compact set $\{|x| \leq K\}$ choose $\delta = \delta(K)$ such that if $|u - v| < \delta$, $|u| \leq K$, $|v| \leq K$ then $|f(u) - f(v)| < \epsilon/4$. Note that $|f(\pm K)| \leq \epsilon/4$. To see that $|f(K)| \leq \epsilon/4$, take a sequence $x_n \searrow K$, with $|x_n| > K$ and use the continuity of f to conclude that $|f(K)| = \lim_n |f(x_n)| \leq \epsilon/4$ because $|f(x_n)| < \epsilon/4$ for all $n \in \mathbb{N}$. Now, take any two real numbers x, y such that $|x - y| < \delta$. If $|x| \leq K$ and $|y| \leq K$ then $|f(x) - f(y)| < \epsilon/4 < \epsilon$. If both $|x| > K$ and $|y| > K$, then $|f(x) - f(y)| < |f(x)| + |f(y)| < \epsilon/2 < \epsilon$. If $|x| \leq K$ and $|y| > K$, then $|f(x) - f(y)| < |f(x) - f(K)| + |f(K) - f(y)| < \epsilon/4 + \epsilon/2 < \epsilon$. ■

The following lemma is an elemental result. We state and prove it here for completeness of the presentation.

Lemma 7.14 *Assume that the sequence of random variables X_n, $n \in \mathbb{N}$ converges in probability to the random variable X. Then $X_n = O_P(1)$.*

Proof: Let $\delta > 0$ be arbitrary.

$$\begin{aligned} P\Big(|X_n| \geq k\Big) &\leq P\Big(k \leq |X_n - X| + |X|\Big) = P\Big(|X_n - X| \geq k - |X|\Big) = \\ &P\Big(|X_n - X| \geq k - |X|, |X| \leq k - \delta\Big) + P\Big(|X_n - X| \geq k - |X|, |X| > k - \delta\Big) \\ &\leq P\Big(|X_n - X| \geq \delta, |X| \leq k - \delta\Big) + P\Big(|X| > k - \delta\Big) \\ &\leq P\Big(|X_n - X| \geq \delta\Big) + P\Big(|X| > k - \delta\Big) \xrightarrow[n\to\infty]{} P\Big(|X| > k - \delta\Big). \end{aligned}$$

Hence, $\limsup_{n\to\infty} P\Big(|X_n| \geq k\Big) \leq P\Big(|X| > k - \delta\Big) \quad \forall \delta > 0$, then

$$\limsup_{n\to\infty} P\Big(|X_n| \geq k\Big) \leq P\Big(|X| \geq k\Big) \xrightarrow[k\to\infty]{} 0.$$

■

Lemma 7.15 - **Uniform Slutzky** - *Assume that $X_n(\theta)$ is a sequence of random variables indexed by a parameter $\theta \in \Theta$. Assume that*

$$\sup_{\theta\in\Theta}\sup_{x\in\mathbb{R}}\left|P\Big(X_n(\theta)\le x\Big)-P\Big(X(\theta)\le x\Big)\right| \xrightarrow[n\to\infty]{} 0\,,$$

and that

$$P\Big(X(\theta)\le x+\eta\Big) \xrightarrow[\eta\to 0]{} P\Big(X(\theta)\le x\Big) \tag{7.8}$$

uniformly on $x\in\mathbb{R}$ and $\theta\in\Theta$. Let $a_n(\theta)$ be a sequence of real random variables indexed by the same set Θ, such that for any $\delta>0$

$$\sup_{\theta\in\Theta} P\Big(\big|a_n(\theta)-a(\theta)\big|>\delta\Big) \xrightarrow[n\to\infty]{} 0\,.$$

Then,

$$\sup_{\theta\in\Theta}\sup_{x\in\mathbb{R}}\left|P\Big(a_n(\theta)\;X_n(\theta)\le x\Big)-P\Big(a(\theta)\;X(\theta)\le x\Big)\right| \xrightarrow[n\to\infty]{} 0\,.$$

Proof: To simplify the notation we will write X_n, a_n, X and a instead of $X_n(\theta)$, $a_n(\theta)$, $X(\theta)$ and $a(\theta)$ respectively. Let $\delta>0$ be arbitrary. Note that

$$\begin{aligned} P(X_n\,a_n\le x) &\le P(|a_n-a|>\delta)+P(X_n\,a_n\le x, |a_n-a|\le\delta) = \\ &= P(|a_n-a|>\delta)+P(X_n\,a_n\le x, |a_n-a|\le\delta)\le \\ &\le P(|a_n-a|>\delta)+P(X_n\le x/(a-\delta))\,. \end{aligned}$$

Let

$$u_n(\theta,\delta)=P(|a_n-a|>\delta)\ ,$$

and

$$t_n(\theta)=\sup_{x\in\mathbb{R}}\left|P\Big(X_n(\theta)\le x\Big)-P\Big(X(\theta)\le x\Big)\right|.$$

Note that if $c \geq 0$ then for any $b \in \mathbb{R}$ we have $b \leq c + |b - c|$. Hence we have

$$P\left(X_n\, a_n \leq x\right) \leq u_n\left(\theta, \delta\right) + t_n\left(\theta\right) + P\left(X \leq x/\left(a - \delta\right)\right) .$$

Similarly we have

$$P\left(X_n > x/\, a_n\right) \leq P\left(\left|a_n - a\right| > \delta\right) + P\left(X_n > x/\left(a + \delta\right)\right) .$$

It follows that

$$P\left(X_n \leq x/\, a_n\right) \geq P\left(X_n \leq x/\left(a + \delta\right)\right) - P\left(\left|a_n - a\right| > \delta\right) .$$

Now the inequality $b \geq c - |b - c|$ implies

$$P\left(X_n \leq x/\, a_n\right) \geq P\left(X \leq x/\left(a + \delta\right)\right) - t_n\left(\theta\right) - u_n\left(\theta, \delta\right) .$$

Let $\epsilon > 0$, choose $n_0 = n_0\left(\epsilon\right)$ large enough such that for any θ, $\left|t_n\left(\theta\right)\right| \leq \epsilon$, for all $n \geq n_0$. Choose $\delta > 0$ such that

$$\left|P\left(X \leq x/\left(a + \delta\right)\right) - P\left(X \leq x/\, a\right)\right| < \epsilon\,,$$

and

$$\left|P\left(X \leq x/\left(a - \delta\right)\right) - P\left(X \leq x/\, a\right)\right| < \epsilon\,,$$

for all $\theta \in \Theta$ and $x \in \mathbb{R}$. For this $\delta = \delta\left(\epsilon\right)$ choose $n_1 = n_1\left(\delta\right) = n_1\left(\epsilon\right)$ such that for all $n \geq n_1$ we have $\sup_{\theta\in\Theta}\left|u_n\left(\theta, \delta\right)\right| \leq \epsilon$. It follows that if $n \geq \max\left(n_0\left(\epsilon\right), n_1\left(\epsilon\right)\right)$ then

$$\left|P\left(X_n \leq x/\, a_n\right) - P\left(X \leq x/\, a\right)\right| < 3\,\epsilon\,,$$

for all $\theta \in \Theta$ and $x \in \mathbb{R}$. ■

Corollary 7.1 *Assume that $X_n(\theta)$ is a sequence of random variables indexed by a parameter $\theta \in \Theta$. Assume that*

$$\sup_{\theta\in\Theta}\sup_{x\in\mathbb{R}}\left|P\Big(X_n(\theta) \le x\Big) - P\Big(X \le x\Big)\right| \xrightarrow[n\to\infty]{} 0,$$

and that X is a continuous random variable with bounded density function. Let $a_n(\theta)$ be a sequence of real random variables indexed by the same set Θ, such that for any $\delta > 0$

$$\sup_{\theta\in\Theta} P\Big(\big|a_n(\theta) - a(\theta)\big| > \delta\Big) \xrightarrow[n\to\infty]{} 0.$$

Then,

$$\sup_{\theta\in\Theta}\sup_{x\in\mathbb{R}}\left|P\Big(a_n(\theta)\; X_n(\theta) \le x\Big) - P\Big(a(\theta)\; X \le x\Big)\right| \xrightarrow[n\to\infty]{} 0.$$

Proof: The result follows immediately from Lemma 7.15 by noting that (7.8) is satisfied. ■

Lemma 7.16 *Assume that $f_n : \mathbb{R} \to \mathbb{R}$ is a sequence of real functions that converges uniformly to $g : \mathbb{R} \to \mathbb{R}$ on a set $\mathcal{K}$. Let a_n be the sequence of infimum of f_n on $\mathcal{K}$, i.e.*

$$a_n = \inf_{x\in\mathcal{K}} f_n(x).$$

and let $b = \inf_{x\in\mathcal{K}} g(x)$. Then $a_n \xrightarrow[n\to\infty]{} b$.

Proof: Assume $b > -\infty$ (the case $b = -\infty$ can be treated along the same lines). Fix $\epsilon > 0$. Let $n_0(\epsilon)$ be such that for all $n \ge n_0(\epsilon)$ we have

$$|f_n(x) - g(x)| < \epsilon, \qquad \forall x \in \mathcal{K}.$$

We have

$$a_n \leq f_n(x) < g(x) + \epsilon, \qquad \forall\, x \in \mathcal{K}. \tag{7.9}$$

It follows that $a_n \leq b + \epsilon$. To prove it, assume that $a_n > b + \epsilon$, i.e. $a_n - \epsilon > b$. By the definition of a_n, there must exist x_g such that $g(x_g) \leq a_n - \epsilon$, which contradicts (7.9). In the same way, we can show that $b - \epsilon < f_n(x)$ for all $x \in \mathcal{K}$, and hence $b - \epsilon \leq a_n$. Finally, we have that for $n \geq n_0(\epsilon)$, $|a_n - b| \leq \epsilon$. ■

Lemma 7.17 *Let $n \geq 1$ and $0 \leq k \leq n$ be integers. Then the function*

$$g(\delta) = P\left(\text{ Binomial}(n, \delta) \geq k\ \right)$$

for $\delta \in [0, 1]$ satisfies: $g(0) = 0$, $g(1) = 1$; g is continuous and non-decreasing.

Proof: That $g(\delta)$ is continuous, $g(0) = 0$, and $g(1) = 1$ is immediate. We now prove its monotonicity. First assume that $k > 1$. We will show that $h(\delta) = 1 - g(\delta)$ is non-increasing. We have

$$h(\delta) = \sum_{i=0}^{k-1} \binom{n}{i} \delta^i (1-\delta)^{n-i}.$$

Then

$$h'(\delta) = \sum_{i=0}^{k-1} \binom{n}{i} \left[i\delta^{i-1}(1-\delta)^{n-i} - (n-i)\,\delta^i (1-\delta)^{n-i-1} \right] = a - b,$$

where

$$a = \sum_{i=1}^{k-1} \binom{n}{i} i\, \delta^{i-1} (1-\delta)^{n-i}$$

and

$$b = \sum_{i=0}^{k-1} \binom{n}{i} (n-i)\, \delta^i (1-\delta)^{n-i-1} = \sum_{i=1}^{k} \binom{n}{i} i\, \delta^{i-1} (1-\delta)^{n-i},$$

where the last equality follows from

$$\binom{n}{i} (n-i) = \binom{n}{i+1} (i+1).$$

Then,

$$h'(\delta) = a - b = -k \binom{n}{k} \delta^{k-1} (1-\delta)^{n-k} \leq 0 \qquad \forall\, \delta \in [0,1]$$

and the proof is complete for $k > 1$. If $k = 1$

$$h(\delta) = P\,(\text{ Binomial}\,(n,\delta) = 0\ ) = (1-\delta)^n,$$

which is clearly decreasing for $\delta \in [0,1]$. Finally if $k = 0$

$$h(\delta) = P\,(\text{ Binomial}\,(n,\delta) < 0\ ) = 0 \qquad \forall\, \delta \in [0,1].$$

■

Consider the metric $d_2(F,G)$ for distribution functions defined by

$$d_2^2(F,G) = \inf E\left[(X-Y)^2\right] \tag{7.10}$$

where the infimum is taken over all the possible distributions of the random vector (X,Y) such that its marginal laws are F and G respectively. This metric was introduced in Mallows (1972) and Tanaka (1973). For a detailed discussion see Bickel and

Freedman, Section 8 (1981). d_2 metrizes weak convergence in the following sense:

$$d_2(F_\alpha, F) \to 0 \quad \text{iff} \quad F_\alpha \xrightarrow{\mathcal{W}} F \quad \text{and} \quad \lim_\alpha E_{F_\alpha} X^2 = E_F X^2$$

where $\xrightarrow{\mathcal{W}}$ denotes weak convergence.

Lemma 7.18 *Let* $\mathbf{X}_1, \ldots, \mathbf{X}_n$ *be independent and identically distributed random vectors on* $\mathbb{R}^p$. *Let* $\hat{\boldsymbol{\theta}}_n = \hat{\boldsymbol{\theta}}_n(\mathbf{X}_1, \ldots \mathbf{X}_n)$ *be a statistic in* $\mathbb{R}^p$ *such that*

$$\hat{\boldsymbol{\theta}}_n \longrightarrow \boldsymbol{\theta}_\infty,$$

to some vector $\boldsymbol{\theta}_\infty \in \mathbb{R}^p$ *almost surely. Let* $\mathbf{h}(\mathbf{x}, t) : \mathbb{R}^p \times \mathbb{R} \to \mathbb{R}^p$ *be a continuous function such that*

$$\|\mathbf{h}(\mathbf{X}, \boldsymbol{\theta})\|^2 \leq \mathbf{g}(\mathbf{X}) \quad \forall \boldsymbol{\theta} \in \Theta, \quad \textit{with } E[\mathbf{g}(\mathbf{X})] < \infty.$$

Let the random variables $\mathbf{Y}_i$ *and* $\mathbf{Z}_i$ *be*

$$\mathbf{Z}_i = \mathbf{h}(\mathbf{X}_i, \boldsymbol{\theta}_\infty) \quad \mathbf{Y}_i = \mathbf{h}(\mathbf{X}_i, \hat{\boldsymbol{\theta}}_n) \quad 1 \leq i \leq n.$$

If G_n *is the empirical cumulative distribution function of the* Y_i, *and* F_n *the corresponding empirical cumulative distribution of the* Z_i, *then we have*

$$\lim_{n \to \infty} d_2(G_n, F_n) = 0 \quad a.s.$$

Proof: Because d_2 in (7.10) is the infimum over all the joint distribution functions with the bootstrapped marginals, an upper bound is given by any joint distribution such that its marginals coincide with the distributions of X and Y. In particular consider the distribution in $\mathbb{R}^p \times \mathbb{R}^p$ that assigns mass $1/n$ to each of the "pairs"

$(\mathbf{h}(\mathbf{X}_i, \hat{\boldsymbol{\theta}}), \mathbf{h}(\mathbf{X}_i, \boldsymbol{\theta}_\infty))$. Choose an arbitrary ϵ. Let $\mathbf{X}$ be a random vector with the same distribution of the $\mathbf{X}_i$s. Let $I_{\mathcal{A}}(X)$ be the indicator function of the set $\mathcal{A}$, i.e. $I_{\mathcal{A}}(\mathbf{X}) = 1 \iff \mathbf{X} \in \mathcal{A}$ and $I_{\mathcal{A}}(\mathbf{X}) = 0$ otherwise. We know that there exists a compact set $\mathcal{K} = \mathcal{K}(\epsilon) \subset \mathbb{R}^p$ such that

$$2\ E\left[I_{\mathcal{K}^c}(\mathbf{X})\ \mathbf{g}(\mathbf{X})\right] < \epsilon/2.$$

For this set $\mathcal{K}(\epsilon)$ there exists a positive number $\delta = \delta(\epsilon, \mathcal{K})$ such that

$$\|\mathbf{h}(\mathbf{X}, \theta_1) - \mathbf{h}(\mathbf{X}, \theta_2)\|^2 < \epsilon/2$$

if $\mathbf{X} \in \mathcal{K}$ and $|\theta_1 - \theta_2| < \delta(\epsilon, \mathcal{K})$. Fix $\omega \in \Omega$ (the probability space) such that $\hat{\boldsymbol{\theta}}_n(\omega)$ converges to $\boldsymbol{\theta}_\infty$. Almost all ω satisfy this. There exists a $n_1 = n_1(\omega, \delta)$ such that $\forall n \geq n_1$

$$\left|\hat{\boldsymbol{\theta}}_n(\omega) - \boldsymbol{\theta}_\infty\right| < \delta/2\ .$$

On the other hand, for a fixed set $\mathcal{K}$ there exists an integer $n_2 = n_2(\epsilon, \omega, \mathcal{K})$ such that for $n \geq n_2$

$$\frac{2}{n}\sum_{i=1}^{n} \mathbf{I}_{\mathcal{K}^c}(\mathbf{X}_i)\, \mathbf{g}(\mathbf{X}_i) < \epsilon/2\ .$$

Take $n > \max(n_1, n_2)$. We have

$$\begin{aligned} d_2^2(F_n, G_n) &\leq \frac{1}{n}\sum_{i=1}^{n} \left\|\mathbf{h}\left(\mathbf{X}_i, \hat{\theta}\right) - \mathbf{h}(\mathbf{X}_i, \theta_\infty)\right\|^2 \\ &< \frac{1}{n}\sum_{\mathbf{X}_i \in \mathcal{K}} \epsilon/2 + \epsilon/2 \\ &< \epsilon\ . \end{aligned}$$

■

Lemma 7.19 *Let $\sigma(F,t)$ be as in (2.23). For any $K_1 > 0$, there exists $K_2 = K_2(K_1) > 0$ such that $\sigma(F,t) \geq K_1$, for all $|t| > K_2$, for all $F \in \mathcal{H}_\epsilon$.*

Proof: Let $F_t = (1-\epsilon)\,\Phi + \delta_t$ where δ_t is a point mass distribution function at t. Then $\sigma(F_t,t)$ satisfies $(1-\epsilon)\; E_\Phi \rho\left((X-t)/\,\sigma(F_t,t)\right) = b$. Hence, for any $F \in \mathcal{H}_\epsilon$ we have

$$E_F \rho\left((X-t)/\,\sigma(F_t,t)\right) = (1-\epsilon)\; E_\Phi \rho\left((X-t)/\,\sigma(F_t,t)\right) \\ + \epsilon\, E_H \rho\left((X-t)/\,\sigma(F_t,t)\right) = \\ b + \epsilon\, E_H \rho\left((X-t)/\,\sigma(F_t,t)\right) \geq b.$$

It follows that $\sigma(F,t) \geq \sigma(F_t,t)$. Also,

$$\lim_{|t|\to\infty} (1-\epsilon)\; E_\Phi \rho\left((X-t)/\,K_1\right) = (1-\epsilon) > b,$$

(because $\rho(\pm\infty) = 1$). Hence, there exists K_2 such that for all $|t| > K_2$,

$$(1-\epsilon)\; E_\Phi \rho\left((X-t)/\,K_1\right) \geq b.$$

Hence $\sigma(F_t,t) \geq K_1$ for $|t| > K_2$. It follows that

$$\sigma(F,t) \geq \sigma(F_t,t) \geq K_1 \qquad \text{for } |t| > K_2, \quad \forall\, F \in \mathcal{H}_\epsilon.$$

■

Lemma 7.20 *Let $\epsilon \in (0,1/2)$ be a fixed number, and let F_ϵ be a distribution function of the form $F_\epsilon = (1-\epsilon)\;\Phi + \epsilon\, H$, where Φ denotes the standard normal distribution function and H is an arbitrary distribution function. Let ψ_c be a function from*

Huber's family (see 2.5), ρ_k be as in (2.14) and let $b = E_\Phi(\rho_k)$. Then

$$\left|\frac{E_{F_\epsilon}\left[\psi_c'(u)\ u\right]}{E_{F_\epsilon}\left[\psi_c'(u)\right]\ E_{F_\epsilon}\left[\rho_k'(u)\ u\right]}\right| \leq \frac{\epsilon}{(1-\epsilon)^2}\,\frac{1}{2}\,\frac{1}{\left[2\Phi(c)-1\right]}\,\frac{1}{\left[b - P\left(|Z|>k\right)\right]},$$

where $Z \sim \Phi$.

Proof: First note that for F_ϵ and any real function h

$$E_{F_\epsilon}\left[h(u)\right] = (1-\epsilon)\ E_\Phi\left[h(u)\right] + \epsilon\, E_H\left[h(u)\right].$$

Also note that ψ_c satisfies

$$\psi_c'(u)\, u = \begin{cases} u/c & \text{if } |u| \leq c \\ 0 & \text{if } |u| > c \end{cases}.$$

Hence

$$\begin{aligned} \left|E_{F_\epsilon}\left[\psi_c'(u)\, u\right]\right| &= \left|(1-\epsilon)\ \frac{1}{c}\int_{-c}^{c} u\,\phi(u)\ du + \epsilon\, E_H\left[\psi_c'(u)\, u\right]\right| \\ &= \epsilon\ \left|E_H\left[\psi_c'(u)\, u\right]\right| \\ &\leq \epsilon, \end{aligned} \tag{7.11}$$

because $\int_{-c}^{c} u\phi(u)\, du = 0$ and $\left|\psi_c'(u)\, u\right| \leq 1$.

To control the denominator we will find upper bounds for $E_{F_\epsilon}\left[\psi_c'(u)\right]$ and $E_{F_\epsilon}\left[\rho_k'(u)\ u\right]$. First note that

$$\begin{aligned} \psi_c'(u) \geq 0,\ \forall\, u \quad \Rightarrow \quad E_{F_\epsilon}\left[\psi_c'(u)\right] &\geq (1-\epsilon)\ E_\Phi\left[\psi_c'(u)\right] \\ &= (1-\epsilon)\ \left(2\Phi(c) - 1\right). \end{aligned} \tag{7.12}$$

Let ρ_k be a function of the family described in (2.14). Then

$$\rho_k'(u)\ u = \begin{cases} 2\,(u/\,k)^2 & \text{if } |u| \leq k \\ 0 & \text{if } |u| > k \end{cases},$$

and then

$$\begin{aligned} E_{F_\epsilon}\left[\rho_k'(u)\ u\right] &\geq (1-\epsilon)\ E_\Phi\left[\rho_k'(u)\ u\right] \\ &= 2\ (1-\epsilon)\ (b - P\left(|Z| > k\right)), \end{aligned} \tag{7.13}$$

where Z denotes a random variable with a standard normal distribution. The latter equality is due to the fact that by hypothesis b satisfies

$$\begin{aligned} b = E_\Phi\left[\rho_k(u)\right] &= \int_{-k}^{k} (u/\,k)^2\,\phi(u) + \int_{-\infty}^{-k} \phi(u)\ du + \int_{k}^{\infty} \phi(u)\ du \\ &= E_\Phi\left[\rho_k'(u)\ u\right]/\,2 + P\left(|Z| > k\right). \end{aligned}$$

The result now follows from (7.11), (7.12) and (7.13). ■

Bibliography

1. Atkinson, A.C. (1985) *Plots, Transformations and Regression.* Oxford: Oxford University Press.

2. Berrendero Díaz, J.R. (1996). *Contribuciones a la teoría de robustez respecto al sesgo.* Tesis Doctoral, Universidad Carlos III de Madrid, Departamento de Estadística y Econometría, Madrid.

3. Berrendero, J.R., Mazzi, S., Romo, J. and Zamar, R. (1998). On the explosion rate of maximum-bias functions. *The Canadian Journal of Statistics*, **26**, 333-351.

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

5. Bickel, P.J. and Doksum, K.A. (1977). *Mathematical statistics: basic ideas and selected topics.* Oakland: Holden-Day.

6. Bickel, P.J. and Freedman, D.A. (1981). Some asymptotic theory for the boot-

strap. *The Annals of Statistics*, **9**, 1196-1217.

7. Boos, D.D. and Serfling, R.J. (1980). A note on differentials and the CLT and LIL for statistical functions, with applications to M-estimates. *The Annals of Statistics*, **8**, 618-624.

8. Brownlee, K.A. (1965). *Statistical Theory and Methodology in Science and Engineering.* Second Edition. New York: Wiley.

9. Carroll, R.J. (1979). On estimating variances of robust estimators when the errors are asymmetric. *Journal of the American Statistical Association*, **74**, 674-679.

10. Chow, Y.S. and Teicher, H. (1988). *Probability Theory. Independence, Interchangeability, Martingales.* Springer Texts in Statistics. New York: Springer-Verlag.

11. Chung, K.L. (1974). *A Course in Probability Theory.* New York: Academic Press.

12. Clarke, B.R. (1983). Uniqueness and Fréchet differentiability of functional solutions to maximum likelihood type equations. *The Annals of Statistics*, **11**, 1196-1205.

13. Clarke, B.R. (1984). Nonsmooth analysis and Fréchet differentiability of M-functionals. Research Report. Murdoch University, Murdoch, Western Australia.

14. Daniel, C. and Wood, F.S. (1980). *Fitting Equations to Data. Computer Analysis of Multifactor Data.* Second Edition. New York: John Wiley & Sons.

15. Davies, P.L. (1990). The asymptotics of S-estimators in the linear regression model. *The Annals of Statistics*, **18**, 1651-1675.

16. Davies, P.L. (1993). Aspects of robust linear regression. *The Annals of Statistics*, **21**, 1843-1899.

17. Davies, P.L. (1998). On locally uniformly linearizable high breakdown location and scale functionals. *The Annals of Statistics*, **26**, 1103-1125.

18. Davison, A.C. and Hinkley, D.V. (1997). *Bootstrap Methods and their Application.* Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge: Cambridge University Press.

19. DiCiccio, T.J. and Efron, B. (1996). Bootstrap Confidence Intervals. *Statistical Science*, **11**, 189-228.

20. 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, California.

21. Dupuis, D.J. and Hamilton, D.C. (2000). Regression residuals and test statistics: Assessing naive outlier deletion. To appear in *The Canadian Journal of Statistics.*

22. Efron, B. (1979). Bootstrap methods: another look at the jackknife. *The Annals of Statistics* **7**, 1-26.

23. Efron, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans.* Philadelphia: Society for Industrial and Applied Mathematics.

24. Efron, B. and Tibshirani, R.J. (1993). *An Introduction to the Bootstrap.* Monographs on Statistics and Applied Probability, 57. New York: Chapman & Hall.

25. Fraiman, R., Yohai, V.J., and Zamar, R. (2000). Optimal robust M-estimates of location. Unpublished manuscript.

26. Freedman, D. A. (1981). Bootstrapping regression models. *The Annals of Statistics*, **9**, 1218-1228.

27. Ghosh, M., Parr, W. C., Singh, K. and Babu, G.J. (1984). A note on bootstrapping the sample median. *The Annals of Statistics*, **12**, 1130-1135.

28. Hall, P. (1986). On the bootstrap and confidence intervals. *The Annals of Statistics*, **14**, 1431-1452

29. Hall, P. (1988). Theoretical comparison of bootstrap confidence intervals. *The Annals of Statistics*, **16**, 927-953.

30. Hall, P. (1992). *The bootstrap and Edgeworth expansion.* New York : Springer-Verlag.

31. Hampel, F.R. (1971). A General Qualitative Definition of Robustness. *The Annals of Mathematical Statistics*, **42**, 1887-1896.

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

33. He, X. (1991). A local breakdown property of robust tests in linear regression. *Journal of Multivariate Analysis*, **38**, 294-305.

34. He, X., Simpson, D.G. and Portnoy, S.L. (1990). Breakdown robustness of tests. *Journal of the American Statistical Association*, **85**, 446-452.

35. Hill, R.W. (1977). *Robust regression when there are outliers in the carriers.* Ph.D. thesis. Harvard University, Cambridge.

36. Huber, P.J. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, **35**, 73-101.

37. Huber, P.J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability.* Berkeley, California.

38. Huber, P.J. (1973). Robust regression: asymptotics, conjectures and Monte Carlo. *The Annals of Statistics*, **1**, 799-821.

39. Huber, P.J. (1981). *Robust Statistics.* New York: Wiley.

40. Krasker, W.S. (1980). Estimation in linear regression models with disparate data points. *Econometrica*, **48**, 1333-1346.

41. Krasker, W.S. and Welsch, R.E. (1982). Efficient bounded influence regression estimation. *Journal of the American Statistical Association*, **77**, 595-604.

42. Mallows, C.L. (1972). A note on asymptotic joint normality. *The Annals of Mathematical Statistics*, **43**, 508-515

43. Markatou, M. and Hettmansperger, T.P. (1990). Robust bounded-influence tests in linear models. *Journal of the American Statistical Association*, **85**, 187-190.

44. Markatou, M., Stahel, W.A., and Ronchetti, E. (1991). Robust M-type testing procedures for linear models. In *Directions in Robust Statistics and Diagnostics*, Part I, Stahel, W., Weisberg, S., eds. Springer-Verlag. pp. 201-220.

45. Maronna, R.A., Bustos, O.H. and Yohai, V.J. (1979). Bias- and efficiency-robustness of general M-estimators for regression with random carriers. In *Smoothing Techniques for Curve Estimation*, T. Gasser and M. Rosenblatt (eds.). Lecture Notes in Mathematics 757, 91-116. Berlin ; New York : Springer-Verlag.

46. Maronna, R.A. and Yohai, V.J. (1981). Asymptotic behavior of general M-estimates for regression and scale with random carriers. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, **58**, 7-20.

47. Martin, R.D. and Zamar, R. (1993). Bias robust estimation of scale. *The Annals of Statistics*, **21**, 991-1017.

48. Martin, R.D., Yohai, V.J. and Zamar, R. (1989). Min-max bias robust regression. *The Annals of Statistics*, **17**, 1608-1630.

49. Parr, W.C. (1985). The bootstrap: some large sample theory and connections with robustness. *Statistics and Probability Letters*, **3**, 97-100.

50. Relles, D.A. And Rogers, W.H. (1977). Statisticians are fairly robust estimators of location. *Journal of the American Statistical Association*, **72**, 107-111.

51. Rocke, D.M. and Downs, G.W. (1981). Estimating the variances of robust estimators of location: influence curve, jackknife and bootstrap. *Communications in Statistics, Part B – Simulation and Computation*, **10**, 221-248.

52. Ronchetti, E. (1982). *Robust Testing in Linear Models: The Infinitesimal Approach.* Unpublished Ph.D. thesis, Swiss Federal Institute of Technology, Zurich.

53. Rosner, B. (1977). Percentage points for the RST many outlier procedure. *Technometrics*, **19**, 307-312.

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

55. Rousseeuw, P.J. and Leroy, A.M. (1987). *Robust Regression and Outlier Detection.* New York: Wiley.

56. Rousseeuw, P.J. and Yohai, V.J. (1984). Robust regression by means of S-estimators. In *Robust and Nonlinear Time Series.* (J. Franke, W. Hardle and D. Martin, eds.). *Lecture Notes in Statist.*, **26** 256-272. Berlin: Springer-Verlag.

57. Seber, G.A.F. (1984). *Multivariate Observations.* New York: John Wiley and Sons.

58. Serfling, R.J. (1980). *Approximation Theorems of Mathematical Statistics.* New York: Wiley.

59. Shao, J. (1990). Bootstrap estimation of the Asymptotic variances of statistical functionals. *Annals of the Institute of Statistical Mathematics*, **42**, 737-752.

60. Shao, J. (1992). Bootstrap variance estimators with truncation. *Statistics and Probability Letters*, **15**, 95-101.

61. Shorack, G.R. (1982). Bootstrapping robust regression. *Communications in Statistics, Part A – Theory and Methods*, **11**, 961-972.

62. Simpson, A., and Eden (1975). A Bayesian analysis of a multiplicative treatment effect in weather modification. *Technometrics*, **17**, 161-166.

63. Singh, K. (1998). Breakdown theory for bootstrap quantiles. *The Annals of Statistics*, **26**, 1719-1732.

64. Tanaka, H. (1973). An inequality for a functional of probability distribution and its application to Kac's one-dimensional model of a Maxwellian gas. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, **27**, 47-52.

65. Venables, W.N. and Ripley, B.D. (1997). *Modern applied statistics with S-PLUS.* Second Edition. New York: Springer.

66. Weins, D.P. (1996). Asymptotics of generalized M-estimation of regression and scale with fixed carriers, in an approximately linear model. *Statistics and Probability Letters*, **30**, 271-285.

67. Weisberg, S. (1985). *Applied Linear Regression.* New York: Wiley.

68. Wu, C.F.J. (1986). Jackknife, bootstrap and other re-sampling methods in regression analysis. *The Annals of Statistics*, **14**, 1261-1295.

69. Yang, S-S (1985). On bootstrapping a class of differentiable statistical functionals with applications to L- and M-estimates. *Statistica Neerlandica*, **39**, 375-385.

70. Yohai, V.J. (1985). High breakdown-point and high efficiency robust estimates for regression. Technical Report No. 66, Deptartment of Statistics, University of Washington, Seattle.

71. Yohai, V.J. (1987). High breakdown-point and high efficiency robust estimates for regression. *The Annals of Statistics*, **15**, 642-656.

72. Yohai, V.J. and Maronna R.A. (1979). Asymptotic behavior of M-estimators for the linear model. *The Annals of Statistics*, **7**, 258-268.

73. Yohai, V.J. and Zamar, R. (1988). High breakdown point estimates of regression by means of the minimization of an efficient scale. *Journal of the American Statistical Association*, **83**, 406-413.