Models where the Least Trimmed Squares and Least Median of Squares estimators are maximum likelihood

Vanessa Berenguer-Rico[*], Søren Johansen[†] & Bent Nielsen[‡]

1 September 2019

Abstract

The Least Trimmed Squares (LTS) and Least Median of Squares (LMS) estimators are popular robust regression estimators. The idea behind the estimators is to find, for a given h, a sub-sample of h 'good' observations among n observations and estimate the regression on that sub-sample. We find models, based on the normal or the uniform distribution respectively, in which these estimators are maximum likelihood. We provide an asymptotic theory for the location-scale case in those models. The LTS estimator is found to be $h^{1/2}$ consistent and asymptotically standard normal. The LMS estimator is found to be h consistent and asymptotically Laplace.

Keywords: Chebychev estimator, LMS, Uniform distribution, Least squares estimator, LTS, Normal distribution, Regression, Robust statistics.

1 Introduction

The Least Trimmed Squares (LTS) and the Least Median of Squares (LMS) estimators suggested by Rousseeuw (1984) are popular robust regression estimators. They are defined as follows. Consider a sample with n observations, where some are 'good' and some are 'outliers'. The user chooses a number h and searches for a sub-sample of h 'good' observations. The idea is to find the sub-sample with the smallest residual sum of squares – for LTS – or the least maximal squared residual – for LMS.

In this paper, we find models in which these estimators are maximum likelihood. In these models, we first draw h 'good' regression errors from a normal distribution, for LTS, or a uniform distribution, for LMS. Conditionally on these 'good' errors, we draw $n-h$ 'outlier' errors from a distribution with support outside the range of the drawn 'good' errors. The models are therefore semi-parametric, so we apply a general notion of maximum likelihood

[*]Department of Economics, University of Oxford. E-mail: vanessa.berenguer-rico@economics.ox.ac.uk.

[†]Department of Economics, University of Copenhagen and CREATES, Department of Economics and Business, Aarhus University, DK-8000 Aarhus C. E-mail: soren.johansen@econ.ku.dk.

[‡]Department of Economics, University of Oxford. Address for correspondence: Nuffield College, Oxford OX1 1NF, UK. E-mail: bent.nielsen@nuffield.ox.ac.uk.

that involves pairwise comparison of probabilities of small hyper-cubes around the data vector. We provide an asymptotic theory for location-scale models for the case where we know the number h of 'good' observations. We find that the LTS estimator is $h^{1/2}$-consistent and asymptotically normal, while the LMS estimator is h-consistent and asymptotically Laplace. More than 50% contamination can be allowed under mild regularity conditions on the distribution function for the 'outliers'. The associated scale estimators do not require consistency factors.

The approach of asking in which models the LTS and LMS estimators are maximum likelihood is similar to that taken by Gauss in 1809, following the principles set out by Laplace in 1774, see Hald (2007, §5.5, 7.2). In the terminology of Fisher (1922), Gauss asked in which continuous i.i.d. location model is the arithmetic mean the maximum likelihood estimator and found the answer to be the normal model. Maximum likelihood often brings a host of attractive features, such as: the model reveals the circumstances under which an estimator works well, provides insight from interpreting the model, produces nice distributional results, yields optimality properties, and we have a framework for testing the goodness-of-fit, which leads to the possibility of first refuting and then improving a model.

To take advantage of these attractive features of the likelihood framework, we follow Gauss and suggest models in which the LTS and LMS estimators are maximum likelihood. The models for LTS and LMS presented here are distinctive in that the errors are not i.i.d.. Rather, the h 'good' errors are i.i.d. and normal, in the LTS model, or uniform, in the LMS model, whereas the $n - h$ 'outlier' errors are i.i.d., conditionally on the 'good' errors, with a distribution assigning zero probability to the range of the 'good' errors. When $h = n$, the models are standard i.i.d. normal or uniform models, just as the LTS and LMS estimators reduce to the least squares estimator and the Chebychev estimator, respectively. The models are semi-parametric, so we use an extension of the traditional likelihoods, in which we carry out pairwise comparison of probabilities of small hypercubes around the data point, following suggestions by Kiefer and Wolfowitz (1956) and Scholz (1980).

The LTS estimator is widely used. At first, the LMS estimator seemed numerically more attractive, but that changed with the fast LTS algorithm approximation to LTS by Rousseeuw and van Driessen (2000). The LTS estimator is often used in its own right and sometimes as a starting point for algorithms such as the Forward Search (Atkinson, Riani, Cerioli, 2010). Many variants of LTS have been developed: non-linear regression in time series (Čižek, 2005), algorithms for fraud detection (Rouseeuw, Perrotta, Riani and Hubert, 2019) and sparse regression (Alfons, Croux and Gelper, 2013).

Rousseeuw (1984) developed the LTS and LMS estimators in the tradition of Huber (1964) and Hampel (1971). Both Huber and Hampel were instrumental in formalizing robust statistics. Huber suggested a framework of i.i.d. errors from an ϵ-contaminated normal distribution, where errors are normal with probability $1 - \epsilon$ and otherwise sampled from a contamination distribution, G say. He developed M-estimators for location as a generalization of maximum likelihood, where the most robust M-estimators would typically rely on a criterion function that would not stem from a distribution. This focused attention on finding estimators rather than providing models. Hampel defined robustness and breakdown points in terms of Prokhorov distances within a measure theoretic framework. Loosely speaking the definitions are as follows. A sequence of estimators from i.i.d. models is robust, if it is bounded in probability in a wide class of distributions. An estimator from an i.i.d. model

has a breakdown point $b \in [0, 1]$, if it is bounded in probability within a class of distribution functions where the maximal distance to the reference distribution function is b. The least squares estimator has breakdown point 0, hence, is severely affected by outliers. The LTS and LMS are some of the first high-breakdown point estimators in regression being suggested. As long as h is chosen larger than $n/2$ plus the dimension of the regressors, the breakdown point is $1 - h/n$ (Rousseeuw 1984, 1985).

An important aspect of the model based framework for analyzing the LTS and LMS estimators is a conceptual shift from the notion of robustness and breakdown point by Hampel (1971) to the notion of consistency by Fisher (1922). There is no doubt that Hampel's ideas have been extraordinarily fruitful in finding new robust estimators. However, when it comes to applying the estimators, classically trained statisticians will look for consistency and reliable inference. It is therefore of interest to describe classes of distributions under which this is achieved. The idea of bounded influence to outliers is a good starting point, but it is insufficient to complete the journey. This view shines through in the discussion of robust regression estimators by Huber and Ronchetti (2009).

To fulfill the task of establishing consistency and deriving reliable inference, we conduct an asymptotic analysis in the location-scale case, where the 'outliers' follow an arbitrary distribution subject to the regularity condition, that 'good' and 'outlier' errors are sufficiently separated. In the LTS model, the 'good' errors are normal. Since the 'outliers' are placed outside the range of the 'good' observations, and the normal distribution has thin tails, the 'good' observations and the outliers' are well-separated. The asymptotic theory shows that the set of 'good' observations is estimated consistently. The rate is so fast, that the estimation error in selecting the 'good' observations does not influence the asymptotic distribution of the LTS estimators. In the LMS model, the 'good' errors are uniform, so that they have thick tails. We separate the 'good' and the 'outlier' errors by requiring that the 'outliers' follow a distribution that is thin for small 'outliers'. This feature is not present in the ϵ-contaminated i.i.d. models of Rousseeuw (1984) and Kim and Pollard (1990), which then results in a rather complicated asymptotic theory, compared to the present case. It also throws light on the discussion regarding estimators' ability to separate overlapping populations of 'good' and 'outlier' observations, see Riani, Atkinson and Perrotta (2014), Doornik (2016).

The LTS and LMS estimators for the regression parameter have the virtue that their derivation does not depend on the scale of the errors. However, the scale is needed for conducting inference. The LTS estimator had previously been analyzed by Butler (1982) for the location-scale case with symmetric contamination. He found it to be $n^{1/2}$-consistent, see also Rousseeuw (1985). Rousseeuw (1984) suggested to estimate the scale by the residual sum of squares for the selected h observations, normalized by a consistency factor found as the variance in a truncated normal distribution. Víšek (2006) has analyzed the asymptotic distribution of the LTS estimator for regression. Regarding the LMS regression estimator, Rousseeuw (1984) suggested that it would be $n^{1/3}$-consistent under symmetric i.i.d. errors, as confirmed by Kim and Pollard (1990). As scale estimator, Rousseeuw (1984) suggested the largest absolute residual among the selected h observations, normalized by a consistency factor found as the $(1 + h/n)/2$ quantile of the standard normal distribution. The maximum likelihood estimators for the scale in the models analyzed in this paper are those suggested by Rousseeuw, but without consistency factors.

The number, h, of 'good' observations is assumed known throughout this paper to match

the data generating process with the LTS/LMS estimators. At first, this may be seen to be very restrictive. In the concluding remarks, we will argue that this assumption is testable and outline a broader context of different types of testable distributional assumptions for the 'good' and the 'outlying' observations.

The technical analysis in the paper includes the following ingredients. As maximum likelihood concept, we use pair-wise comparison of probability measures, as suggested by Kiefer and Wolfowitz (1956), and consider small hyper-cubes around the data point following Fisher (1922) and Scholz (1980). For the asymptotic analysis of the LTS model, we apply marked and weighted empirical processes of residuals (Johansen and Nielsen, 2016a; Berenguer-Rico, Johansen and Nielsen, 2019), quantile processes (Csörgő, 1983), and extreme value theory for the normal distribution (Leadbetter, Lindgren and Rootzén, 1982, Watts 1980). For the asymptotic analysis of the LMS model, we apply results for uniform spacings (Pyke, 1965).

We start by presenting the LTS and LMS estimators and the associated least squares and Chebychev estimators in §2. The general maximum likelihood concept is introduced in §3. The models for LTS and LMS are given in §4, and §6 provides an asymptotic analysis for the location-scale case with proofs given in the Appendix. The supplementary material contains some derivations that appear to be well-known in the literature, but where we could not find a convenient reference, along with some simple, but tedious algebraic manipulations.

2 The LTS and LMS estimators

The available data are a scalar y_i and a p-vector of regressors x_i for $i = 1, \ldots, n$. We consider a regression model $y_i = \beta' x_i + \sigma\varepsilon_i$ with regression parameter β and scale σ. For the moment, the distribution of x_i, ε_i is not specified.

2.1 Definitions of estimators

We will consider four estimators for β. First, the ordinary least squares estimator

$$\hat{\beta}_{OLS} = (\sum_{i=1}^{n} x_i x_i')^{-1} \sum_{i=1}^{n} x_i y_i,$$

which minimizes the least squares criteria $\sum_{i=1}^{n}(y_i - \beta' x_i)^2$. It is known to be sensitive to unusual combinations of x_i, ε_i.

Second, the Chebychev regression estimator, also referred to as the L_∞ estimator or the minimax estimator, has the form

$$\hat{\beta}_{Cheb} = \arg\min_{\beta} \max_{1\leq i\leq n} |y_i - \beta' x_i|.$$

Knight (2017) studied the asymptotic theory of this estimator. Wagner (1959) pointed out that the Chebychev estimator can be found as a regular linear programming problem with $p + 1$ relations, where p is the dimension of x_i. This implies that the maximum absolute residual will be attained at $p + 1$ points. Harter (1953) and Schechtman and Schechtman (1986) found that the Chebychev estimator is a non-unique maximum likelihood estimator in a model with uniform errors with known range. We show in Theorem 5.2, that it is the

unique maximum likelihood, when the range of the uniform errors is unknown, see also §B.1 in the supplementary material.

Third, the Least Trimmed Squares (LTS) estimator was suggested by Rousseeuw (1984). It is computed as follows. Given a value of β compute the squared residuals $r_i^2(\beta) = (y_i - \beta' x_i)^2$. The ordered residuals are denoted $r^2_{(1)}(\beta) \leq \cdots \leq r^2_{(n)}(\beta)$. The user chooses an integer $h \leq n$. Given that choice, the sum of the h smallest residual squares is computed. Minimizing over β gives the LTS estimator

$$\hat{\beta}_{LTS} = \arg\min_{\beta} \sum_{i=1}^{h} r^2_{(i)}(\beta). \tag{2.1}$$

The LTS minimization classifies the observations as 'good' or 'outliers'. The set of indices of the h 'good' observations is denoted ζ, which is an h-subset of $(1, \ldots, n)$. The estimator $\hat{\zeta}_{LTS}$ is, therefore, the indices of the observations corresponding to $r^2_{(i)}(\hat{\beta}_{LTS})$ for $i \leq h$. Rousseeuw and van Driessen (2000) point out that we can compute $\hat{\beta}_{LTS}$ as a minimizer over least squares estimators, that is

$$\hat{\zeta}_{LTS} = \arg\min_{\zeta} \sum_{i \in \zeta} (y_i - \hat{\beta}'_{\zeta} x_i)^2 \quad \text{where} \quad \hat{\beta}_{\zeta} = (\sum_{i \in \zeta} x_i x_i')^{-1} \sum_{i \in \zeta} x_i y_i. \tag{2.2}$$

Fourth, the Least Median of Squares (LMS) estimator was also suggested by Rousseeuw (1984). Rousseeuw was concerned with the case where $h = n/2$ or its integer part, but we consider any choice of an integer $h \leq n$. Given that choice, the LMS estimator is

$$\hat{\beta}_{LMS} = \arg\min_{\beta} r^2_{(h)}(\beta). \tag{2.3}$$

Rousseeuw has $h = n/2$, so that $r^2_{(h)}(\beta)$ is a median, but other quantiles are routinely used.

As for the LTS, the LMS minimization divides the observations into 'good' observations and 'outliers'. The indices of the 'good' observations are estimated by $\hat{\zeta}_{LMS}$, which consists of the indices of the observations corresponding to $r^2_{(i)}(\hat{\beta}_{LMS})$ for $i \leq h$. Thus, we can compute $\hat{\beta}_{LMS}$ as a minimizer over Chebychev estimators, that is

$$\hat{\zeta}_{LMS} = \arg\min_{\zeta} \max_{i \in \zeta} |y_i - \hat{\beta}'_{\zeta} x_i| \quad \text{where} \quad \hat{\beta}_{\zeta} = \arg\min_{\beta} \max_{i \in \zeta} |y_i - \beta' x_i|. \tag{2.4}$$

When calculating the LTS and LMS estimators it is possible that observations are repeated. In the models we introduce later, continuity of the distribution of the 'good' observations is important and repetitions are ruled out. The idea is essentially the same as in ordinary least squares theory. The least squares estimator can be applied to data with repetitions of the dependent variable, but under normality, repetitions happen with probability zero.

The LTS and LMS estimators are computationally demanding. They require binomial searches, which are infeasible except for a very small dimensions. Computational approximations have been suggested in the literature, such as the fast LTS by Rousseeuw and van Driessen (2000). The optimization problem simplifies considerably for location-scale models.

2.2 The special case of location-scale models

The LTS and LMS estimators simplify for the special case of a location-scale model, so that $y_i = \mu + \sigma\varepsilon_i$. In both cases, we need to find sets ζ of 'good' data. The parameter ζ has $\binom{n}{h}$ possible values. However, we will argue that to find the optimal set $\hat{\zeta}$, it suffices to study the order statistics $y_{(1)} \leq \cdots \leq y_{(n)}$. The optimal set $\hat{\zeta}$ is then the indices of observations of the form $y_{(\delta+1)}, \ldots, y_{(\delta+h)}$ for some $\delta = 0, \ldots, n-h$, where the optimal $\hat{\delta}$ is

$$\hat{\delta}_{LTS} = \underset{0\leq\delta\leq n-h}{\arg\min} \sum_{i=1}^{h} \{y_{(\delta+i)} - \hat{\mu}_\delta\}^2 \qquad \text{where} \qquad \hat{\mu}_\delta = \frac{1}{h}\sum_{i=1}^{h} y_{(\delta+i)},$$

in the case of LTS, while in the case of LMS,

$$\hat{\delta}_{LMS} = \underset{0\leq\delta\leq n-h}{\arg\min} \{y_{(\delta+h)} - y_{(\delta+1)}\} \qquad \text{while} \qquad \hat{\mu}_{LMS} = \{y_{(\delta+h)} + y_{(\delta+1)}\}/2.$$

This replaces the problem of finding the optimal $\hat{\zeta}$ by the problem of finding an optimal $\hat{\delta}$.

For the LMS estimator, Rousseuw (1984, Theorem 2) proves the equivalence of the optimality problems. The idea is as follows. For an arbitrary ζ, the Chebychev estimator is $\hat{\mu}_\zeta = (\max_{i\in\zeta} y_i + \min_{i\in\zeta} y_i)/2$ and the square root of the criterion function in (2.4) is then $\max_{i\in\zeta} |y_i - \hat{\mu}_\zeta| = (\max_{i\in\zeta} y_i - \min_{i\in\zeta} y_i)/2$. This value equals $\{y_{(\delta+r)} - y_{(\delta+1)}\}/2$ where $\delta+1$ and $\delta+r$ are the ranks of $\min_{i\in\zeta} y_i$ and $\max_{i\in\zeta} y_i$, such that $r \geq h$. If $r > h$, the criterion is bounded below by $\{y_{(\delta+h)} - y_{(\delta+1)}\}/2$.

For the LTS estimator and an arbitrary ζ, the argument is slightly tedious and left to §**??** in the supplementary material.

2.3 Scale estimation in i.i.d. models

The problem of estimating the scale is intricately linked to the choice of model for the innovations ε_i. Proposals have been given for regression models with i.i.d. innovations ε_i with distribution function F.

For the LTS estimator, Croux and Rousseeuw (1992) proposed to estimate the scale by the residual sum of squares of selected observations divided by a consistency factor defined as the conditional variance of ε_i given that $|\mathsf{F}(\varepsilon_i) - 1/2| \leq h/(2n)$.

For the LMS estimator, Rousseuw (1984) was concerned with a model where ε_i are i.i.d. normal with $\mathsf{F} = \Phi$ and $h = n/2$. Evaluated at the true parameter, $r^2_{(h)}(\beta_0)$ equals the hth smallest value of $\sigma^2\varepsilon_i^2$. Under normality this converges to $\sigma^2\{\Phi^{-1}(0.75)\}^2$. Thus, he suggests to estimate the scale by $r_{(h)}(\hat{\beta}_{LMS})/\Phi^{-1}(0.75)$ where $1/\Phi^{-1}(0.75) = 1.483$.

2.4 Asymptotics for i.i.d. models

Butler (1982) proved that in a location-scale model, where ε_i are i.i.d. with a symmetric and strongly unimodal distribution, then $n^{1/2}(\hat{\beta}_{LTS} - \beta)$ is asymptotically normal distributed with a variance depending on the density of ε_i. Víšek (2006) analyzed $\hat{\beta}_{LTS}$ in the regression case. Johansen and Nielsen (2016b, Theorem 5) proved consistency and provided an asymptotic expansion of the above scale estimator under normality.

Rousseeuw (1984) argues heuristically that in a location-scale model, where ε_i are i.i.d. with a symmetric and strongly unimodal distribution, $n^{1/3}(\hat{\beta}_{LMS} - \beta)$ converges in distribution. This has been studied formally by Kim and Pollard (1990).

3 A general definition of maximum likelihood

Traditional parametric maximum likelihood is defined in terms of densities, which are not well-defined here. Thus, we follow the generalization proposed by Scholz (1980), which has two ingredients. First, it uses pairwise comparison of measures, as suggested by Kiefer and Wolfowitz (1956), see Johansen (1978) and Gissibl, Klüppelberg, and Lauritzen (2019) for applications. This way, a dominating measure is avoided. Second, it compares probabilities of small sets that include the data point, following the informality of Fisher (1922). This way, densities are not needed. Scholz' approach is suited to the present situation, where the candidate maximum likelihood estimator is known and we are searching for a model.

We consider data in $\mathbb{R}^n$ and can therefore simplify the approach of Scholz. Let $\mathcal{P}$ be a family of probability measures on the Borel sets of $\mathbb{R}^n$. Given a (data) point $y \in \mathbb{R}^n$ and a distance ϵ define the hypercube $C_y^\epsilon = (y_1 - \epsilon, y_1] \times \cdots \times (y_n - \epsilon, y_n]$, which is a Borel set.

Definition 3.1 *For* $\mathsf{P}, \mathsf{Q} \in \mathcal{P}$ *write* $\mathsf{P} <_y \mathsf{Q}$ *if* $\limsup_{\epsilon \to 0}\{\mathsf{P}(C_y^\epsilon)/\mathsf{Q}(C_y^\epsilon)\} < 1$ *and* $\mathsf{P} \leq_y \mathsf{Q}$ *if* $\limsup_{\epsilon \to 0}\{\mathsf{P}(C_y^\epsilon)/\mathsf{Q}(C_y^\epsilon)\} \leq 1$, *where by convention* $0/0 = 1$.

Following Scholz (1980), define P, Q to be equivalent at y and write $\mathsf{P} =_y \mathsf{Q}$ if $\mathsf{P} \leq_y \mathsf{Q}$ and $\mathsf{Q} \leq_y \mathsf{P}$. As Scholz, we get that (i) $\mathsf{P} =_y \mathsf{Q}$ if and only if $\lim_{\epsilon \to 0}\{\mathsf{P}(C_y^\epsilon)/\mathsf{Q}(C_y^\epsilon)\}$ exists and equals 1; (ii) $\mathsf{P} <_y \mathsf{Q}$ and $\mathsf{Q} <_y \mathsf{R}$ implies $\mathsf{P} <_y \mathsf{R}$ (transitivity); and (iii) $\mathsf{P} =_y \mathsf{P}$ for all $\mathsf{P} \in \mathcal{P}$ (reflexivity).

Definition 3.2 *The probability measure* $\hat{\mathsf{P}} \in \mathcal{P}$ *is a maximum likelihood estimator of* $\mathsf{P} \in \mathcal{P}$ *at* y *if* $\mathsf{P} \leq_y \hat{\mathsf{P}}$ *for all* $\mathsf{P} \in \mathcal{P}$. *It is unique if* $\mathsf{P} <_y \hat{\mathsf{P}}$ *for all* $\mathsf{P} \neq \hat{\mathsf{P}}$. *We will say that* $\mathsf{L}^\epsilon(\mathsf{P}) = \mathsf{P}(C_y^\epsilon)$ *is the* ϵ*-likelihood for the data point* y.

Scholz provides two examples that we will use here. Detailed derivations are provided in the supplementary material.

Example 3.1 (Traditional maximum likelihood) *Suppose* $\mathsf{P}, \mathsf{Q} \in \mathcal{P}$ *are dominated by a* σ*-finite measure* μ *with density versions* p *and* q *with respect to* μ. *Suppose* p *and* q *are continuous at* y *with* $q(y) > 0$. *Then* $\lim_{\epsilon \to 0} \mathsf{P}(C_y^\epsilon)/\mathsf{Q}(C_y^\epsilon) = p(y)/q(y)$.

Example 3.2 (Empirical distribution function) *Consider* $y_1, \ldots, y_n$ *that are i.i.d. with unknown distribution function* F *on* $\mathbb{R}$. *Let* $x_1 < \cdots < x_k$ *be the distinct outcomes with counts* $n_1, \ldots, n_k$ *so that* $\sum_{j=1}^k n_j = n$. *The empirical distribution function* $\mathsf{F}_n(x) = n^{-1} \sum_{i=1}^n 1_{(y_i \leq x)}$ *has support on* $x_1 < \cdots < x_k$ *with jumps of size* n_j/n. *Then* F_n *is the maximum likelihood estimator for* F *with* $\mathsf{P}_{\mathsf{F}_n}(C_y^\epsilon) = \prod_{j=1}^k (n_j/n)^{n_j}$ *for any* $\epsilon < \min_{1<j\leq k}(x_j - x_{j-1})$.

4 Location-scale models

We start by presenting the LTS location-scale model. Subsequently, it is used for the likelihood analysis. Finally, an LMS model and its likelihood are presented.

4.1 The LTS location-scale model

Model 1 (LTS location-scale model) *Consider the location-scale model $y_i = \mu + \sigma\varepsilon_i$ for data y_i with $i = 1, \ldots, n$. Let $h \leq n$ be given. Let ζ be a set with h elements from $1, \ldots, n$. For $i \in \zeta$, let ε_i be i.i.d. $\mathsf{N}(0,1)$ distributed.*
For $j \notin \zeta$, let ξ_j be i.i.d. with distribution function $\mathsf{G}(x)$ for $x \in \mathbb{R}$ where G is continuous at 0. The 'outlier' errors are defined by

$$\varepsilon_j = (\max_{i\in\zeta} \varepsilon_i + \xi_j)1_{(\xi_j>0)} + (\min_{i\in\zeta} \varepsilon_i + \xi_j)1_{(\xi_j<0)}. \tag{4.1}$$

The parameters are $\mu \in \mathbb{R}$, $\sigma > 0$, ζ which is any h-subset of $1, \ldots, n$ and G which is an arbitrary distribution on $\mathbb{R}$ assumed continuous at zero.

The observations y_i indexed by $i \in \zeta$ are the 'good' observations, while those indexed by $j \notin \zeta$ are the 'outliers'. We note that the 'good' observations have a continuous distribution, so they can not be repetitions. The 'outliers' may come from an arbitrary distribution that may have a discrete element. Moreover, the 'good' observations must have consecutive order statistics since the 'outliers' must take values outside the range of the 'good' observations. Randomly, according to the choice of G, some 'outliers' are to the left of the 'good' observations. The count of left 'outliers' is the random variable

$$\delta = \sum_{j\notin\zeta} 1_{(\xi_j<0)} = \sum_{j\notin\zeta} 1_{(\varepsilon_j<\min_{i\in\zeta}\varepsilon_i)}. \tag{4.2}$$

Thus, the ordered errors satisfy

$$\underbrace{\varepsilon_{(1)} \leq \cdots \leq \varepsilon_{(\delta)}}_{\delta \text{ left 'outliers'}} < \underbrace{\varepsilon_{(\delta+1)} < \cdots < \varepsilon_{(\delta+h)}}_{h \text{ 'good'}} < \underbrace{\varepsilon_{(\delta+h+1)} \leq \cdots \leq \varepsilon_{(n)}}_{\overline{n}=n-h-\delta \text{ right 'outliers'}}.$$

The set ζ corresponds to indices of observations corresponding to $\varepsilon_{(\delta+1)}, \ldots, \varepsilon_{(\delta+h)}$.

4.2 Maximum likelihood for the LTS location-scale model

We find the ϵ-likelihood for the LTS location-scale Model 1.

We start by finding the probability that the random n-vector y belongs to an ϵ-cube C_x^ϵ around an n-vector x of outcomes. Since the 'outliers' depend on the 'good' observations, we write

$$\mathsf{P}(y \in C_x^\epsilon) = \prod_{i\in\zeta} \mathsf{P}(x_i - \epsilon < y_i \leq x_i) \prod_{j\notin\zeta} \mathsf{P}(x_j - \epsilon < y_j \leq x_j \mid y_i \text{ for } i \in \zeta). \tag{4.3}$$

For the 'good' observations $\mathsf{P}(x_i - \epsilon < y_i \leq x_i) = \Phi\{(x_i - \mu)/\sigma\} - \Phi\{(x_i - \epsilon - \mu)/\sigma\}$, which we denote $\Delta^\epsilon\Phi\{(x_i - \mu)/\sigma\}$. For the 'outliers', combine the model equation $y_i = \mu + \sigma\varepsilon_i$ and the error definition (4.1) to get

$$y_j = (\max_{i\in\zeta} y_i + \sigma\xi_j)1_{(\xi_j\geq 0)} + (\min_{i\in\zeta} y_i + \sigma\xi_j)1_{(\xi_j<0)}.$$

It follows that for $y_j > \max_{i\in\zeta} y_i$ we have $\xi_j = (y_j - \max_{i\in\zeta} y_i)/\sigma$, which is invariant to μ. Thus, the conditional probability that an 'outlier' belongs to an ϵ-interval given the 'good' observations is, for $x_j > \max_{i\in\zeta} y_i$,

$$\mathsf{P}(x_j - \epsilon < y_j \le x_j \mid y_i \text{ for } i \in \zeta) = \mathsf{G}\{(x_j - \max_{i\in\zeta} y_i)/\sigma\} - \mathsf{G}\{(x_j - \epsilon - \max_{i\in\zeta} y_i)/\sigma\},$$

which we denote $\Delta^\epsilon \mathsf{G}\{(x_j - \max_{i\in\zeta} y_i)/\sigma\}$. Correspondingly, for $x_j < \min_{i\in\zeta} y_i$, we have $\mathsf{P}(x_j - \epsilon < y_j \le x_j \mid y_i \text{ for } i \in \zeta) = \Delta^\epsilon \mathsf{G}\{(x_j - \min_{i\in\zeta} y_i)/\sigma\}$.

The ϵ-likelihood arises from the probability (4.3) by inserting the above expressions for the individual probabilities and replacing the outcome vector x with observations y to get

$$\mathsf{L}^\epsilon(\mu,\sigma,\zeta,\mathsf{G}) = 1_{(y_i\neq y_{i^\dagger} \text{ for } i,i^\dagger\in\zeta)} \prod_{i\in\zeta} \Delta^\epsilon \Phi(\frac{y_i-\mu}{\sigma}) \\ \times \prod_{j\notin\zeta} \Delta^\epsilon \mathsf{G}\{\frac{y_j - \min_{i\in\zeta} y_i}{\sigma} 1_{(y_j < \min_{i\in\zeta} y_i)} + \frac{y_j - \max_{i\in\zeta} y_i}{\sigma} 1_{(y_j > \max_{i\in\zeta} y_i)}\}, \quad (4.4)$$

where the likelihood is set to zero if any two 'good' observations are equal, due to the continuity of the 'good' observations. The factor for the 'outliers' allow repetitions. Thus, any repetitions in the sample has to be found among the 'outliers'. As an example, suppose we have $n = 9$ observations with ordered values

$$1, 1, 2, 3, 6, 6, 7, 8, 9.$$

The values 1 and 6 are repetitions and cannot be 'good'. Thus, for $h = 2$, we can select ζ as the index pairs corresponding to the ordered pairs with values (2,3), (7,8), (8,9), all other choices would have a zero likelihood.

The two products in (4.4) resemble a standard normal likelihood and a likelihood for the problem of estimating a distribution, respectively, see Examples 3.1, 3.2. We will exploit those examples using profile likelihood arguments.

First, suppose μ, σ, ζ are given. Then the first product in the LTS ϵ-likelihood (4.4) is constant. The second product depends on G and corresponds to the ϵ-likelihood in Example 3.2 for the model with unknown distribution function. The likelihood is maximized in the same way. If $y_j > \max_{i\in\zeta} y_i$, the observation y_j is shifted to $y_j - \max_{i\in\zeta} y_i$, and if $y_j < \min_{i\in\zeta} y_i$, the observation y_j is shifted to $y_j - \min_{i\in\zeta} y_i$. Let $x_1 < \cdots < x_k$ be the distinct values of the shifted values of y_j for $j \notin \zeta$. The values x_ℓ have counts n_ℓ, so that $\sum_{\ell=1}^k n_\ell = n-h$. The maximum value of the second factor is $\prod_{\ell=1}^k \{n_\ell/(n-h)\}^{n_\ell} = (\prod_{\ell=1}^k n_\ell^{n_\ell})/(n-h)^{n-h}$ for any ϵ less than the smallest spacing $x_j - x_{j-1}$. The maximum value is constant in ϵ and in ζ, where all selected 'good' observations are singletons. For any ζ with repetitions among y_i for $i \in \zeta$, the likelihood is zero. Thus, maximizing over G gives

$$\mathsf{L}^\epsilon_{\mathsf{G}}(\mu,\sigma,\zeta) = \mathsf{L}^\epsilon(\mu,\sigma,\zeta,\hat{\mathsf{G}}) = 1_{(y_i\neq y_{i^\dagger} \text{ for } i,i^\dagger\in\zeta)} \prod_{i\in\zeta} \Delta^\epsilon \Phi(\frac{y_i-\mu}{\sigma}) \prod_{\ell=1}^k (\frac{n_\ell}{n-h})^{n_\ell}.$$

Second, suppose ζ is given. Apart from a constant, we find that $\mathsf{L}_{\mathsf{G}} = \lim_{\epsilon\to 0} \epsilon^{-h} \mathsf{L}^\epsilon_{\mathsf{G}}$ is a standard normal likelihood, see Example 3.1, which is maximized by

$$\hat\mu_\zeta = h^{-1} \sum_{i\in\zeta} y_i \qquad \text{and} \qquad \hat\sigma^2_\zeta = h^{-1} \sum_{i\in\zeta} (y_i - \hat\mu_\zeta)^2. \quad (4.5)$$

Thus, the profile likelihood for ζ is

$$\mathsf{L}_{\mu,\sigma,\mathsf{G}}(\zeta) = \lim_{\epsilon \to 0} \epsilon^{-h} \mathsf{L}^{\epsilon}_{\mu,\sigma,\mathsf{G}}(\zeta) = (2\pi e \hat{\sigma}_\zeta^2)^{-h/2} 1_{(y_i \neq y_{i^\dagger} \text{ for } i, i^\dagger \in \zeta)} \prod_{\ell=1}^{k} (\frac{n_\ell}{n-h})^{n_\ell}.$$

Third, this profile likelihood is maximized by choosing ζ so that $\hat{\sigma}_\zeta$ is as small as possible. In §2.2, it was argued that the minimizer for ζ must select observations corresponding to order statistics $y_{(\delta+1)} < \cdots < y_{(\delta+h)}$ for $0 \leq \delta \leq n-h$. Thus, instead of $\hat{\mu}_\zeta, \hat{\sigma}_\zeta$ in (4.5), it suffices to consider

$$\hat{\mu}_\delta = h^{-1} \sum_{i=\delta+1}^{\delta+h} y_{(i)} \qquad \text{and} \qquad \hat{\sigma}_\delta^2 = h^{-1} \sum_{i=\delta+1}^{\delta+h} \{y_{(i)} - \hat{\mu}_\delta\}^2, \tag{4.6}$$

where $\hat{\delta}_{LTS}$ minimizes $\hat{\sigma}_\delta^2$ subject to $y_{(i)} \neq y_{(i^\dagger)}$ for $\delta + 1 \leq i, i^\dagger \leq \delta + h$. We summarize.

Theorem 4.1 *The LTS location-scale Model 1 has ϵ-likelihood $\mathsf{L}^{\epsilon}(\mu, \sigma, \zeta, \mathsf{G})$ defined in (4.4), which, for $\epsilon \to 0$, is maximized as follows. Recall the definition of $\hat{\mu}_\delta, \hat{\sigma}_\delta^2$ in (4.6). Let $\hat{\delta}_{LTS} = \arg\min_{0 \leq \delta \leq n-h} \hat{\sigma}_\delta^2$ subject to $y_{(i)} \neq y_{(i^\dagger)}$ for $\delta+1 \leq i, i^\dagger \leq \delta+h$. Then $\hat{\zeta}_{LTS}$ is given by the indices corresponding to $y_{(\hat{\delta}_{LTS}+1)}, \ldots, y_{(\hat{\delta}_{LTS}+h)}$ so that $\hat{\mu}_{LTS} = \hat{\mu}_{\hat{\delta}_{LTS}}$ and $\hat{\sigma}_{LTS} = \hat{\sigma}_{\hat{\delta}_{LTS}}$.*

We note that the LTS estimator (2.1) is the maximum likelihood estimator for the location μ in the LTS Model 1 subject to the assumption that the 'good' observations are normal. This constraint is irrelevant if all observations are distinct. We show in Theorem 6.3 that the maximum likelihood estimator for scale, $\hat{\sigma}_{LTS}$, is consistent without any need for a consistency factor. This contrasts with the estimators for scale in i.i.d. models reviewed in §2.4. The intuition is that in the LTS Model 1, the 'good' observations follow a normal distribution without truncation.

4.3 The LMS location-scale model

The LMS model has the same setup as the LTS model with the exception that the 'good' observations follow a uniform distribution.

Model 2 (LMS location-scale model) *Consider data $y_1, \ldots, y_n$. Follow the setup of the LTS location-scale Model 1 apart from the following:*
For $i \in \zeta$ let ε_i be i.i.d. uniformly distributed on $[-1, 1]$.

The LMS model permits that the 'outlier' distribution G can be chosen so that the 'good' observations and the 'outliers' are independent. This is because the 'good' observations have finite support, in contrast to the case of the LTS model. The LMS model permits that for $j \notin \zeta$, we can choose ε_j as i.i.d. with a distribution function G, which is constant and continuous on $[-1, 1]$, and which does not depend on ε_i for $i \in \zeta$. Imposing this constraint on the maximum likelihood problem would change the solution along the lines of Harter (1953) and Schechtman and Schechtman (1986). This possibility is not pursued here.

4.4 Maximum likelihood for the LMS location-scale model

The likelihood is defined in a similar fashion as in (4.4), only replacing the normal distribution function with a uniform distribution function, $\mathsf{U}(x)$, taking the value $(x+1)/2$ for $|x| \le 1$. Let $\Delta^\epsilon \mathsf{U}(y) = \mathsf{U}(y) - \mathsf{U}(y-\epsilon)$. The ϵ-likelihood is therefore

$$\begin{aligned}\mathsf{L}^\epsilon(\mu,\sigma,\zeta,\mathsf{G}) =& 1_{(y_i \neq y_{i^\dagger} \text{ for } i,i^\dagger \in \zeta)} \prod_{i\in\zeta} \Delta^\epsilon \mathsf{U}(\frac{y_i-\mu}{\sigma}) \\ & \times \prod_{j\notin\zeta} \Delta^\epsilon \mathsf{G}[\{(y_j - \min_{i\in\zeta} y_i) 1_{(y_j < \min_{i\in\zeta} y_i)} + (y_j - \max_{i\in\zeta} y_i) 1_{(y_j > \max_{i\in\zeta} y_i)}\}/\sigma]. \quad (4.7)\end{aligned}$$

We find the maximum likelihood estimator as before. First, maximize over G as before, to get the profile ϵ-likelihood for μ, σ^2, ζ:

$$\mathsf{L}^\epsilon_{\mathsf{G}}(\mu,\sigma,\zeta) = \mathsf{L}^\epsilon(\mu,\sigma^2,\zeta,\hat{\mathsf{G}}) = 1_{(y_i \neq y_{i^\dagger} \text{ for } i,i^\dagger\in\zeta)} \prod_{i\in\zeta} \Delta^\epsilon \mathsf{U}(\frac{y_i-\mu}{\sigma}) \prod_{j=1}^k (\frac{n_j}{n-h})^{n_j}.$$

Second, suppose ζ is given. We find that $\mathsf{L}_{\mathsf{G}} = \lim_{\epsilon\to 0} \epsilon^{-h} \mathsf{L}^\epsilon_{\mathsf{G}}$ is a standard uniform likelihood, see Example 3.1. The essential part is

$$\lim_{\epsilon\to 0} \prod_{i\in\zeta} \epsilon^{-1} \Delta^\epsilon \mathsf{U}\{(y_i-\mu)/\sigma\} = \prod_{i\in\zeta} (2\sigma)^{-1} 1_{(|y_i-\mu|\le\sigma)} = (2\sigma)^{-h} 1_{(\max_{i\in\zeta}|y_i-\mu|\le\sigma)}.$$

For given μ, ζ, this is maximized by choosing σ as small as possible, subject to the constraint $\sigma \ge \max_{i\in\zeta} |y_i - \mu|$. The lower bound is smallest when μ is taken as the mid-point between the largest and the smallest y_i for $i \in \zeta$. That is, for $\hat{\mu}_\zeta = (\max_{i\in\zeta} y_i + \min_{i\in\zeta} y_i)/2$ so that $\hat{\sigma}_\zeta = (\max_{i\in\zeta} y_i - \min_{i\in\zeta} y_i)/2$. Thus, the profile likelihood for ζ is

$$\mathsf{L}_{\mu,\sigma,\mathsf{G}}(\zeta) = \lim_{\epsilon\to 0} \epsilon^{-h} \mathsf{L}^\epsilon_{\mu,\sigma,\mathsf{G}}(\zeta) = 1_{(y_i \neq y_{i^\dagger} \text{ for } i,i^\dagger\in\zeta)} (2\hat{\sigma}_\zeta)^{-h}.$$

Third, this profile likelihood is maximized by minimizing $\hat{\sigma}_\zeta$. In §2.2, see also Rousseeuw (1984, Theorem 2), it was argued that the minimizer ζ must select observations corresponding to order statistics $y_{(\delta+1)} < \cdots < y_{(\delta+h)}$ for $0 \le \delta \le n-h$. Thus, instead of $\hat{\mu}_\zeta, \hat{\sigma}_\zeta$ it suffices to consider $\hat{\mu}_\delta = \{y_{(\delta+h)} + y_{(\delta+1)}\}/2$ and $\hat{\sigma}_\delta = \{y_{(\delta+h)} - y_{(\delta+1)}\}/2$. We choose $\hat{\delta}_{LMS}$ as the minimizer of $\hat{\sigma}_\delta$ subject to $y_{(i)} \neq y_{(i^\dagger)}$ for $\delta+1 \le i, i^\dagger \le \delta+h$. We summarize.

Theorem 4.2 *The LMS location-scale Model 2 has ϵ-likelihood $\mathsf{L}^\epsilon(\mu,\sigma,\zeta,\mathsf{G})$ defined in (4.7). The maximum likelihood estimator is defined as follows. For $\delta = 0,\ldots,n-h$ define $\hat{\sigma}_\delta = \{y_{(\delta+h)} - y_{(\delta+1)}\}/2$ and $\hat{\mu}_d = \{y_{(\delta+h)} + y_{(\delta+1)}\}/2$. Let $\hat{\delta}_{LMS} = \arg\min_\delta \hat{\sigma}_\delta$ subject to $y_{(i)} \neq y_{(i^\dagger)}$ for $\delta+1 \le i, i^\dagger \le \delta+h$. Then $\hat{\zeta}_{LMS}$ is given by the indices corresponding to $y_{(\hat{\delta}_{LMS}+1)},\ldots,y_{(\hat{\delta}_{LMS}+h)}$ while $\hat{\mu}_{LMS} = \hat{\mu}_{\hat{\delta}_{LMS}}$ and $\hat{\sigma}_{LMS} = \hat{\sigma}_{\hat{\delta}_{LMS}}$.*

We note that the LMS estimator is the maximum likelihood estimator for the location μ in the LMS Model 2 subject to the assumption that the 'good' observations are uniform. The maximum likelihood estimator for the scale differs from the residual sum of squares suggestion reviewed in §2.4, since the model assumes uniformity rather than normality. We show in Theorem 6.5, that the maximum likelihood estimator for scale $\hat{\sigma}_{LMS}$ does not require a consistency correction factor.

5 Regression models

5.1 The LTS regression model

The argument for the regression case is essentially the same as for the location-scale case. We do, however, need the restriction that the distribution G for the 'outliers' is continuous.

Model 3 (LTS regression model) *Consider the regression model $y_i = \beta' x_i + \sigma\varepsilon_i$ for data y_i, x_i with $i = 1, \ldots, n$. Let $h \leq n$ be given. Let ζ be a set with h elements from $1, \ldots, n$.*
For $i \in \zeta$, let ε_i be i.i.d. $\mathsf{N}(0,1)$ distributed.
For $j \notin \zeta$, let ξ_j be i.i.d with continuous distribution function $\mathsf{G}(x)$ for $x \in \mathbb{R}$. Then define the 'outlier' errors ε_j from ξ_j as in (4.1).
The parameters are $\beta \in \mathbb{R}^{\dim x}$, $\sigma > 0$, ζ which is any h-subset of $1, \ldots, n$ and G which is an arbitrary continuous distribution on $\mathbb{R}$.

The reason that we now require continuity of the 'outlier' distribution G is subtle. In the location scale model, the observations y_i are a simple translation of the errors $y_i - \mu$ for some value of μ. As a consequence, the ranks of observations y_i and the errors $y_i - \mu$ are identical and repetitions in the errors match repetitions in the observations. It is then possible to construct the likelihood so that repetitions in errors only happen among the 'outliers'.

For the regression model, where $y_i = \beta' x_i + \sigma\varepsilon_i$, there is no simple relationship between the ordering of the observations and the errors. As for the location-scale problem, repetitions of pairs y_i, x_i is not a problem in itself. However, for any set of observations x_i, y_i the parameter β can be chosen so that the errors $y_i - \beta' x_i$ are the same for different values of x_i. This is illustrated in Figure 1. Thus, the number of repetitions of the errors depends on the choice of the parameter β. This will complicate the maximization of the likelihood. This is avoided by requiring that ε_i has a continuous distribution for all i. A zero likelihood is then assigned to parameter values β with repetitions of the residual $y_i - \beta' x_i$.

The LTS location-scale ϵ-likelihood (4.4) is modified as follows, with $y_i^{\beta x} = y_i - \beta' x_i$,

$$\mathsf{L}^\epsilon(\beta, \sigma, \zeta, \mathsf{G}) = 1_{(y_i^{\beta x} \neq y_{i^\dagger}^{\beta x} \text{ for } 1 \leq i, i^\dagger \leq n)} \prod_{i \in \zeta} \Delta^\epsilon \Phi(y_i^{\beta x}/\sigma)$$
$$\times \prod_{j \notin \zeta} \Delta^\epsilon \mathsf{G}[\{(y_j^{\beta x} - \min_{i \in \zeta} y_i^{\beta x}) 1_{(y_j^{\beta x} < \min_{i \in \zeta} y_i^{\beta x})} + (y_j^{\beta x} - \max_{i \in \zeta} y_i^{\beta x}) 1_{(y_j^{\beta x} > \max_{i \in \zeta} y_i^{\beta x})}\}/\sigma]. \quad (5.1)$$

It is maximized along the lines of the LTS location-scale ϵ-likelihood. The only exception is that the very last step of switching from general sub-samples ζ to consecutive order statistics is no longer possible. We then arrive at the following result.

Theorem 5.1 *The LTS regression model (3) has ϵ-likelihood $\mathsf{L}^\epsilon(\beta, \sigma, \zeta, \mathsf{G})$ defined in (5.1) and which is maximized, for $\epsilon \to 0$, as follows. For any h-sub-sample ζ, define the least squares estimator $\hat{\beta}_\zeta = (\sum_{i \in \zeta} x_i x_i')^{-1} \sum_{i \in \zeta} x_i y_i$ and the residual sum of squares estimator $\hat{\sigma}_\zeta^2 = h^{-1} \sum_{i \in \zeta} (y_i - \hat{\beta}_\zeta' x_i)^2$. Let $\hat{\zeta}_{LTS} = \arg\min_\zeta \hat{\sigma}_\zeta^2$ subject to the constraint that $\hat{\varepsilon}_i \neq \hat{\varepsilon}_{i^\dagger}$ where $\hat{\varepsilon}_i = y_i - \hat{\beta}_\zeta' x_i$ and $1 \leq i < i^\dagger \leq n$. Then $\hat{\beta}_{LTS} = \hat{\beta}_{\hat{\zeta}_{LTS}}$ and $\hat{\sigma}_{LTS} = \hat{\sigma}_{\hat{\zeta}_{LTS}}$.*

We note that the LTS estimator (2.1) is the maximum likelihood estimator in the LTS regression Model 3 subject to a continuity assumption for all observations.

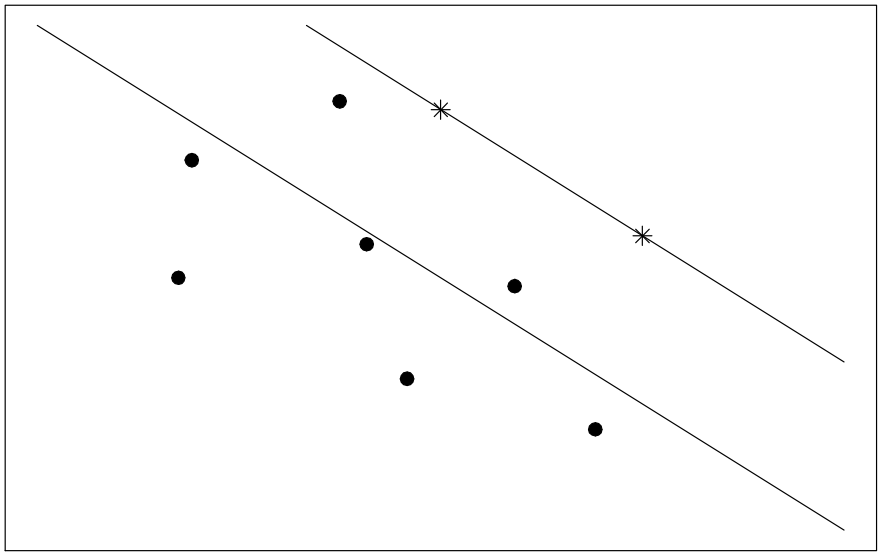

Figure 1: A set of points (x_i, y_i) and a choice of β so that the residuals $y_i - \beta' x_i$ is the same for two different points. The long line is the regression line where $y = \beta' x$. The shorter line passes through two points marked with $*$, which have the same residual.

5.2 The LMS regression model

We can move from LTS to LMS in a similar fashion.

Model 4 (LMS regression model) *Consider the regression model* $y_i = \beta' x_i + \sigma\varepsilon_i$ *for data* y_i, x_i *with* $i = 1, \ldots, n$. *Follow the setup of the LTS regression Model 3, but where* ε_i, *for* $i \in \zeta$, *are i.i.d. uniformly distributed on* $[-1, 1]$.

The likelihood is similar to that of the LTS regression, where the normal distribution function Φ is replaced with the distribution function U for a uniform distribution on $[-1, 1]$.

Theorem 5.2 *The likelihood for the LMS regression Model 4 is maximized as follows. For any h-sub-sample* ζ *define the Chebychev estimator* $\hat{\beta}_\zeta = \arg\min_\beta \max_{i\in\zeta} |y_i - \beta' x_i|$ *and the scale estimator* $\hat{\sigma}_\zeta = \max_{i\in\zeta} |y_i - \hat{\beta}'_\zeta x_i|$. *Let* $\hat{\zeta}_{LTS} = \arg\min_\zeta \hat{\sigma}_\zeta$ *subject to the constraint that* $y_i - \hat{\beta}'_\zeta x_i \neq y_{i^\dagger} - \hat{\beta}'_\zeta x_{i^\dagger}$ *for* $1 \leq i, i^\dagger \leq n$. *Then* $\hat{\beta}_{LMS} = \hat{\beta}_{\hat{\zeta}_{LMS}}$ *and* $\hat{\sigma}_{LMS} = \hat{\sigma}_{\hat{\zeta}_{LMS}}$.

6 Asymptotics for the location-scale case

We now consider asymptotic theory for the OLS, LTS, LMS estimators in the LTS, LMS location-scale models. We start by choosing a sequence of data generating processes.

6.1 Sequence of data generating processes

For each n, the LTS and LMS location-scale models involve a choice h_n, which is known to the investigator. If $h_n = n$ the LTS and LMS estimators reduce to the full sample least squares

and Chebychev estimators, respectively, with standard asymptotic theory as described in Lemmas A.7, A.12. Here, we choose h_n so that

$$h_n/n \to \gamma \quad \text{for} \quad 0 < \gamma < 1, \tag{6.1}$$

where γ is the asymptotic proportion of 'good' observations. We will not consider the case where $\gamma = 1$ and $h_n < n$. The parameters μ, σ, G are constant in n.

When choosing the sets ζ_n, it is convenient to reparametrize G in terms of

$$\rho = \mathsf{G}(0), \quad \overline{\mathsf{G}}(x) = (1-\rho)^{-1}\{\mathsf{G}(x) - \rho\}1_{(x>0)}, \quad \underline{\mathsf{G}}(x) = 1 - \rho^{-1}\lim_{\epsilon \downarrow 0}\mathsf{G}(-x-\epsilon)1_{(x>0)}, \tag{6.2}$$

so that $\underline{\varepsilon}_j = -\xi_j 1_{(\xi_j<0)}$ is $\underline{\mathsf{G}}$-distributed and $\overline{\varepsilon}_j = \xi_j 1_{(\xi_j>0)}$ is $\overline{\mathsf{G}}$-distributed. This gives the decomposition

$$\mathsf{G}(x) = \{\rho + (1-\rho)\overline{\mathsf{G}}(x)\}1_{(x>0)} + \rho\{1 - \underline{\mathsf{G}}(-x)\}1_{(x\leq 0)}.$$

The 'outliers', ε_j for $j \notin \zeta$, can be constructed through a binomial experiment. Draw $n-h$ independent $\mathsf{Bernoulli}(\rho)$ variables. If the jth variable is unity then $\varepsilon_j = \min_{i\in\zeta}\varepsilon_i - \underline{\varepsilon}_j$. If it is zero then $\varepsilon_j = \max_{i\in\zeta}\varepsilon_i + \overline{\varepsilon}_j$. In this way, the number of left 'outliers' is

$$\delta_n = \sum_{j\in\zeta_n} 1_{(\varepsilon_j < \min_{i\in\zeta_n}\varepsilon_i)}. \tag{6.3}$$

When maximizing the likelihood it suffices to consider sets ζ_n corresponding to order statistics $y_{(\delta_n+1)}, \ldots, y_{(\delta_n+h_n)}$. The Law of Large Numbers gives

$$\delta_n/(n-h_n) \xrightarrow{a.s.} \rho. \tag{6.4}$$

The number of 'outliers' to the right are $\overline{n} = n - h_n - \delta_n$, so that $\overline{n}/(n-h_n) \to 1-\rho$ $a.s.$ We note the following limits

$$\delta_n/h_n \xrightarrow{a.s.} \underline{\omega} = \rho(1-\gamma)/\gamma, \qquad \overline{n}/h_n \xrightarrow{a.s.} \overline{\omega} = (1-\rho)(1-\gamma)/\gamma. \tag{6.5}$$

In summary, the sequence of data generating processes is defined by $\mu, \sigma, \rho, \underline{\mathsf{G}}, \overline{\mathsf{G}}$ and h_n, δ_n.

In the asymptotic analysis of the LTS and LMS estimators, the main challenge is to show that the estimation error for the frequency of left 'outliers' $(\hat{\delta} - \delta_n)/h_n$ vanishes at various rates, where $\hat{\delta}$ could be $\hat{\delta}_{LTS}$ or $\hat{\delta}_{LMS}$. We will write out detailed proofs for the situation where $\hat{\delta} > \delta_n$, so that some of the small 'good' observations are considered left 'outliers' and some of the small right 'outliers' are considered 'good'. The case $\hat{\delta} < \delta_n$ is analogous due to the setup for the left and right 'outliers' in (6.2), since we can multiply all observations by -1 and relabel left and right. Moreover, when considering $\hat{\delta} > \delta_n$ we note that $\hat{\delta} - \delta_n \leq \overline{n}$. Due to the binomial construction of δ_n then $\overline{n} = 0$ $a.s.$ when $\rho = 1$, so that the event $\hat{\delta} > \delta_n$ is a null set. Thus, when analysing $(\hat{\delta} - \delta_n)/h_n$ it suffices to consider $\hat{\delta} > \delta_n$ and $\rho < 1$.

6.2 OLS estimator in the LTS and LMS location-scale models

We start by showing that the least squares estimator $\hat{\mu}_{OLS} = n^{-1}\sum_{i=1}^n y_i$ can diverge in the LTS model when $h_n/n \to \gamma$ where $0 < \gamma < 1$. This implies that the least squares estimator is not robust within the LTS model in the sense of Hampel (1971). For simplicity, we consider the case where all 'outliers' are to the right so that $\rho = 0$.

Theorem 6.1 *Consider the LTS location-scale Model 1 and the sequence of data generating processes outlined in §6.1, so that $h_n/n \to \gamma$ where $0 < \gamma < 1$ and $\rho = 0$. Suppose $\overline{\mathsf{G}}$ has finite expectation $\mu_{\mathsf{G}} = \int_0^\infty \{1 - \overline{\mathsf{G}}(x)\}dx$. Then, $\hat{\mu}_{OLS}$ diverges at a $(2\log n)^{1/2}$ rate,*

$$(2\log n)^{-1/2}(\hat{\mu}_{OLS} - \mu)/\sigma \xrightarrow{\mathsf{P}} 1 - \gamma > 0.$$

We now consider an LMS model with a similar type of outlier distribution. In this case, $\hat{\mu}_{OLS}$ is inconsistent but the bias is bounded. This indicates that the least squares estimator is robust in the sense of Hampel (1971) within a wider class of contamination in the LMS model than in the LTS model.

Theorem 6.2 *Consider the LMS location-scale Model 2 and the sequence of data generating processes outlined in §6.1, so that $h_n/n \to \gamma$ where $0 < \gamma < 1$ and $\rho = 0$. Suppose $\overline{\mathsf{G}}$ has finite expectation $\mu_{\mathsf{G}} = \int_0^\infty \{1 - \overline{\mathsf{G}}(x)\}dx$. Then, $\hat{\mu}_{OLS}$ has an asymptotic bias,*

$$(\hat{\mu}_{OLS} - \mu)/\sigma \xrightarrow{\mathsf{P}} (1 - \gamma)(1 + \mu_{\mathsf{G}}) > 0.$$

6.3 LTS estimator in LTS location-scale model

We show that the asymptotic distribution of the LTS estimators is the same as it would have been, if we knew which observations were 'good'. In the analysis, we distinguish between the cases where less than and where more than half of the observations are 'outliers' as the latter case requires further regularity conditions.

Theorem 6.3 *Consider the LTS location-scale Model 1 and the sequence of data generating processes outlined in §6.1, so that $h_n/n \to \gamma$ where $1/2 < \gamma < 1$. Then, for any $\alpha > 0$,*

$$\hat{\delta}_{LTS} = \delta_n + \mathsf{o}_{\mathsf{P}}(h_n^\alpha), \quad h_n^{1/2}(\hat{\mu}_{LTS} - \mu)/\sigma \xrightarrow{\mathsf{D}} \mathsf{N}(0,1), \quad h_n^{1/2}(\hat{\sigma}^2_{LTS} - \sigma^2)/\sigma^2 \xrightarrow{\mathsf{D}} \mathsf{N}(0,2),$$

where $n^{1/2}(\hat{\mu}_{LTS} - \mu)$ and $n^{1/2}(\hat{\sigma}^2_{LTS} - \sigma^2)$ are asymptotically independent.

The main difficulty in the proofs is to analyze the estimation error $\hat{\delta}_{LTS} - \delta_n$ for the number of 'outliers' to the left, see (6.3). By construction, we have that $-\delta_n \leq \hat{\delta}_{LTS} - \delta_n \leq \overline{n}$, where δ_n and $\overline{n} = n - h_n - \delta_n$ are the number of 'outliers' to the left and to the right, respectively. We want to show that $\hat{\delta}_{LTS}$ is consistent in the sense that $(\hat{\delta}_{LTS} - \delta_n)/h_n = \mathsf{o}_{\mathsf{P}}(h_n^{\alpha-1})$, or equivalently $(\hat{\delta}_{LTS} - \delta_n)/n = \mathsf{o}_{\mathsf{P}}(n^{\alpha-1})$, for any $\alpha > 1$. In the proof of Theorem 6.3, we analyze the criterion function $\hat{\sigma}_\delta - \hat{\sigma}_{\delta_n}$ for varying δ and require that there are less than h_n 'outliers' to the left and to the right, so that $\delta_n/h_n \to \underline{\omega} = \rho(1-\gamma)/\gamma < 1$ and $\overline{n}/h_n \to \overline{\omega} = (1-\rho)(1-\gamma)/\gamma < 1$, see (6.5). This is satisfies when $\gamma > 1/2$ as in Theorem 6.3, but also in some cases where $\gamma \leq 1/2$. At the extreme, this covers the situation with $\gamma > 1/3$ and symmetric 'outliers'.

We now turn to the case allowing more than half of the observations to be outliers. When $\underline{\omega} \geq 1$ or $\overline{\omega} \geq 1$, additional regularity conditions are needed. Those regularity conditions require the following definition from empirical process and quantile process theory.

Definition 6.1 *A distribution function* H *is said to be regular, if it is twice differentiable on an open interval* $\mathcal{S} =]\underline{s}, \overline{s}[$ *with* $-\infty \leq \underline{s} < \overline{s} \leq \infty$ *so that* $\mathsf{H}(\underline{s}) = 0$ *and* $\mathsf{H}(\overline{s}) = 1$ *and the density* h *and its derivative* $\dot{\mathsf{h}}$ *satisfy*
(*a*) $\sup_{x\in\mathcal{S}} \mathsf{h}(x) < \infty$ *and* $\sup_{x\in\mathcal{S}} \mathsf{H}(x)\{1-\mathsf{H}(x)\}|\dot{\mathsf{h}}(x)|/\{\mathsf{h}(x)\}^2 < \infty$*;*
(*b*) *If* $\lim_{x\downarrow\underline{s}} \mathsf{h}(x) = 0$ *(resp.* $\lim_{x\uparrow\overline{s}} \mathsf{h}(x) = 0$*) then* h *is non-decreasing (resp. non-increasing) on a interval to the right of* $\underline{s}$ *(left of* $\overline{s}$*).*

Example 6.1 *The exponential distribution with* $\mathsf{H}(x) = 1-\exp(-\lambda x)$ *is regular with* $\mathcal{S} = \mathbb{R}_+$ *and* $\mathsf{H}(x)\{1-\mathsf{H}(x)\}|\dot{\mathsf{h}}(x)|/\{\mathsf{h}(x)\}^2 = \mathsf{H}(x) \leq 1$ *while* $\mathsf{h}(x)$ *is decreasing as* $x \to \infty$*.*

Assumption 6.1 *Recall the definitions of* $\underline{\omega}, \overline{\omega}$ *from (6.5). Let* $q > 4$*.*
(*i*) *If* $\overline{\omega} \geq 1$ *suppose* $\overline{\mathsf{G}}$ *is regular with* $\int_0^\infty x^q d\overline{\mathsf{G}}(x) < \infty$ *and*
$\overline{v} = \min_{\overline{\omega}^{-1}\leq\varsigma\leq 1} \mathsf{Var}\{\overline{\varepsilon}_1 | \varsigma - \overline{\omega}^{-1} \leq \overline{\mathsf{G}}(\overline{\varepsilon}_1) \leq \varsigma\} > 1$*.*
(*ii*) *If* $\underline{\omega} \geq 1$ *suppose* $\underline{\mathsf{G}}$ *is regular with* $\int_0^\infty x^q d\underline{\mathsf{G}}(x) < \infty$ *and*
$\underline{v} = \min_{\underline{\omega}^{-1}\leq\varsigma\leq 1} \mathsf{Var}\{\underline{\varepsilon}_1 | \varsigma - \underline{\omega}^{-1} \leq \underline{\mathsf{G}}(\underline{\varepsilon}_1) \leq \varsigma\} > 1$*.*

Theorem 6.4 *Consider the LTS location-scale Model 1 and the sequence of data generating processes outlined in §6.1 satisfying Assumptions 6.1, so that* $h_n/n \to \gamma$ *where* $0 < \gamma < 1$*. Then, the limit distributions established in Theorem 6.3 apply.*

6.4 LMS estimator in LMS location-scale model

For the LMS estimator, we also distinguish between the cases where less than half and more than half of the observations are 'outliers'. Since the 'good' observations are uniform, the 'outliers' have to be constrained even in the case with less 'outliers' than good observations.

Assumption 6.2 *Let* $G(x)$ *for* $x \geq 0$ *represent* $\overline{\mathsf{G}}(x)$ *or* $\underline{\mathsf{G}}(x)$*. Suppose*
(*i*) $\exists \epsilon > 0$ *so that* $\forall 0 < \psi < 1$ *then* $G^{-1}(\psi) \geq 2\psi\varrho$ *where* $\varrho = (1-\rho+\epsilon)(1-\gamma)/\gamma$*;*
(*ii*) $\exists \psi_0 > 0, \tau < 1$ *so that* $\forall 0 < \psi < \psi_0$ *then* $G^{-1}(\psi) \geq \psi^\tau$*.*

Figure 2 illustrates Assumption 6.2. The assumption constrains the 'outliers', so that the quantile function G^{-1} for the 'outliers' grows faster than the uniform quantile function close to the 'good' observations. This way, the 'outliers' close to the 'good' observations are more dispersed than the 'good' observations. The first condition is needed to establish that $\hat{\delta} = \delta_n + \mathsf{o}_\mathsf{P}(h_n)$. It gives a global bound on the entire distribution function, while allowing bursts of very concentrated 'outliers' as long as they are not in the vicinity of the 'good' observations. The second condition is needed to show the stronger result that $\mathsf{P}(\hat{\delta} = \delta_n) \to 1$. This requires a strengthened bound to the 'outliers' in the vicinity of the 'good' observations to ensure separation of 'outliers' and 'good' observations.

Theorem 6.5 *Consider the LMS location-scale Model 2 and the sequence of data generating processes outlined in §6.1, so that* $h_n/n \to \gamma$ *where* $1/2 < \gamma < 1$*, and suppose Assumption 6.2 holds. Let* $\overline{e}, \underline{e}$ *be independent, standard exponential variables. Then*

$$\mathsf{P}(\hat{\delta} = \delta_n) \to 1, \qquad h_n(\hat{\mu} - \mu)/\sigma \xrightarrow{\mathsf{D}} \underline{e} - \overline{e}, \qquad h_n(\hat{\sigma} - \sigma)/\sigma \xrightarrow{\mathsf{D}} -(\underline{e} + \overline{e}),$$

where $\underline{e} - \overline{e}$ *and* $\underline{e} + \overline{e}$ *are dependent* $\mathsf{Laplace}(0, \gamma)$ *and* $-\Gamma(2, \gamma)$ *variables.*

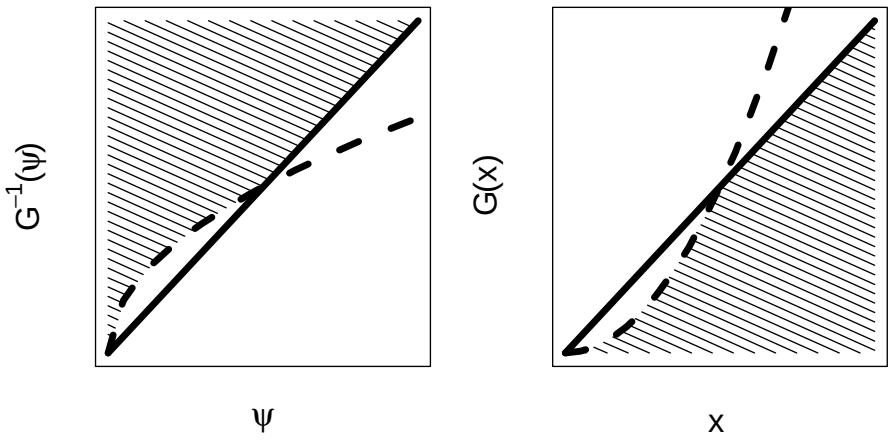

Figure 2: Illustration of Assumption 6.2. Left panel: any permitted quantile $G^{-1}(\psi)$ falls in the shaded area formed by the linear (—) lower bound $2\psi(1-\rho+\epsilon_G)(1-\gamma)/\gamma$ and the curved (- - -) lower bound ψ^τ. Right panel: any permitted distribution function $G(x)$ must be in the shaded area.

In the regression case, the asymptotic distribution is likely to be considerably more complicated and similar to that found by Knight (2017).

For a general proportion of 'outliers', we need to strengthen part (i) of Assumption 6.2 to uniformly bound the dispersion of the 'outliers' from below.

Assumption 6.3 *Recall the definitions of $\underline{\omega}$, $\overline{\omega}$ from (6.5). Let $\varrho = (1-\rho+\epsilon)(1-\gamma)/\gamma$ for some $\epsilon > 0$.*
(i) If $\overline{\omega} \geq 1$ suppose that $\forall 0 < \xi < \xi + \psi \leq 1$ then $\overline{\mathsf{G}}^{-1}(\xi+\psi) - \overline{\mathsf{G}}^{-1}(\xi) \geq 2\psi\varrho$;
(ii) If $\underline{\omega} \geq 1$ suppose that $\forall 0 < \xi < \xi + \psi \leq 1$ then $\underline{\mathsf{G}}^{-1}(\xi+\psi) - \underline{\mathsf{G}}^{-1}(\xi) \geq 2\psi\varrho$.

Theorem 6.6 *Consider the LMS location-scale Model 2 and the sequence of data generating processes outlined in §6.1, so that $h_n/n \to \gamma$ where $0 < \gamma < 1$, and suppose Assumptions 6.2,6.3 hold. Then, the limit distributions established in Theorem 6.5 apply.*

7 Discussion

Other models of the LTS/LMS type. New models and estimators can be generated by replacing the normal/uniform assumption in the LTS/LMS models with some other distribution. An example is the Laplace distribution. This yields what is known as the Least Trimmed sum of Absolute deviations (LTA) estimator. That estimator has been studied by Hawkins and Olive (1999) and it is a special case of the rank-based estimators studied by Hössjer (1994), see also Dodge and Jurečková (2000, §2.7).

Applying LTS with less than $n/2$ 'good' observations. This is commonly done, despite the original recommendation by Rousseeuw (1984) of having more than $n/2$ 'good' observations. A particular example is when starting the Forward Search with an LTS estimator with a small selection h, see for instance Atkinson, Riani and Cerioli (2010). Theorem 6.4 supports

the idea that this could give a consistent starting point. Johansen and Nielsen (2016b) give an asymptotic theory for the Forward Search conditional on a consistent start and under the assumption of i.i.d. normal errors.

Inference requires a model for the outliers. In the presented theory, the 'good' and the 'outlying' observations are separated, which gives a nice distribution theory for inference. The traditional approach, as advocated by Huber (1964), is to consider mixture distributions formed by mixing a reference distribution with a contamination distribution. This gives a different distribution theory for inference. In practice, an investigator using LTS/LMS estimation should seek to test whether one of these models is appropriate and conduct inference accordingly.

Misspecification tests should be developed for the present model. An investigator will have to choose the number h of 'good' observations. This choice may seem daunting, but in principle it is not different from any other model choice in statistics. Given that choice, the set of 'good' observations is estimated and misspecification tests can be applied. The asymptotic theory developed here indicates that standard tests for normality or uniformity can be applied to the set of estimated 'good' observations.

Recently, misspecification tests have been developed for i.i.d. models, where the 'good' observations are truncated normal. Berenguer-Rico and Nielsen (2017) show that the cumulant based normality tests needs consistency correction factors, while Bererenguer-Rico and Wilms (2018) show that the White heteroskedasticity test applies without consistency correction under symmetry.

Applying LTS/LMS with unknown h. It would be interesting to study the properties of the LTS/LMS estimators when using the present model framework, but with a wrong choice of h. It is possible that it could result in a framework for consistently estimating h along the lines of the above mentioned Forward Search.

References

Alfons, A., Croux, C. and Gelper, S. (2013) Sparse least trimmed squares regression for analyzing high-dimensional large data sets. *Annals of Applied Statistics* 7, 226–248.

Atkinson, A.C., Riani, M. and Cerioli, A. (2010) The forward search: Theory and data analysis (with discussion). *Journal of the Korean Statistical Society* 39, 117–163.

Berenguer-Rico, V., Johansen, S. and Nielsen, B. (2019) Uniform consistency of the marked and weighted empirical distribution of residuals. Department of Economics Discussion Paper 19-09, University of Copenhagen.

Berenguer-Rico, V. and Nielsen, B. (2017) Marked and weighted empirical processes of residuals with applications to robust regressions. Department of Economics Discussion Paper 841, University of Oxford.

Berenguer-Rico, V. and Wilms, I. (2018) Heteroscedasticty testing after outlier removal. Department of Economics Discussion Paper 853, University of Oxford

Billingsley, P. (1968) *Convergence of Probability Measures.* New York: Wiley.

Butler, R.W. (1982) Nonparametric interval and point prediction using data trimmed by a Grubbs-type outlier rule. *Annals of Statistics* 10, 197–204.

Čižek, P. (2005) Least trimmed squares in nonlinear regression under dependence. *Journal of Statistical Planning and Inference* 136, 3967–3988.

Croux, C. and Rousseeuw, P.J. (1992) A class of high-breakdown scale estimators based on subranges. *Communications in Statistics. Theory and Methods* 21, 1935–1951.

Csörgő, M. (1983) *Quantile Processes with Statistical Applications.* Philadelphia, PA: Society for Industrial and Applied Mathematics.

Davidson, J. (1994) *Stochastic Limit Theory.* Oxford: Oxford University Press.

Dodge, Y. and Jurečková, J. (2000) *Adaptive Regression.* New York: Springer.

Doornik, J.A. (2016) An example of instability: Discussion of the paper by Søren Johansen and Bent Nielsen. *Scandinavian Journal of Statistics* 43, 357–359.

Fisher, R.A. (1922) On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London* A222, 309–368.

Gissibl, N., Klüppelberg, C. and Lauritzen, S. (2019) Identifiability and estimation of recursive max-linear models. arXiv:1901.03556v1.

Hald, A. (2007) *A history of parametric statistical inference from Bernoulli to Fisher, 1713-1935.* New York: Springer.

Hampel, F.R. (1971) A general qualitative definition of robustness. *Annals of Mathematical Statistics* 42, 1887–1896.

Harter, H.L. (1953) Maximum likelihood regression equations (abstract) *Annals of Mathematical Statistics* 24, 687.

Hawkins, D.M. and Olive, D.J. (1999) Improved feasible solution algorithms for high breakdown estimation. *Computational Statistics & Data Analysis* 30, 1–11.

Hössjer, O. (1994) Rank-based estimates in the linear model with high breakdown point. *Journal of the American Statistical Association* 89, 149–158.

Huber, P.J. (1964) Robust estimation of a location parameter. *Annals of Mathematical Statistics* 35, 73–101.

Huber, P.J. and Ronchetti, E. (2009) *Robust Statistics*, 2nd edition. Hoboken, NJ: Wiley.

Johansen, S. (1978) The product limit estimator as maximum likelihood estimator. *Scandinavian Journal of Statistics* 5, 195–199.

Johansen, S. and Nielsen, B. (2016a) Analysis of the Forward Search using some new results for martingales and empirical processes. *Bernoulli* 22, 1131–1183

Johansen, S. and Nielsen, B. (2016b) Asymptotic theory of outlier detection algorithms for linear time series regression models (with discussion). *Scandinavian Journal of Statistics* 43, 321–381.

Johnson, N.L., Kotz, S. and Balakrishnan, N. (1994a) *Continuous Univariate Distributions, volume 1*, 2nd edition. New York: Wiley.

Johnson, N.L., Kotz, S. and Balakrishnan, N. (1994b) *Continuous Univariate Distributions, volume 2*, 2nd edition. New York: Wiley.

Kiefer, J. and Wolfowitz, J. (1956) Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Annals of Mathematical Statistics* 27, 887–906.

Kim, J. and Pollard, D. (1990) Cube root asymptotics. *Annals of Statistics* 18, 191–219.

Knight, K. (2017) On the asymptotic distribution of the L_∞ estimator in linear regression. Mimeo, http://www.utstat.utoronto.ca/keith/home.html.

Leadbetter, M.R., Lindgren, G. and Rootzén, H. (1982) *Extremes and Related Properties of Random Sequences and Processes.* New York: Springer.

Pyke, R. (1965) Spacings (with discussion). *Journal of the Royal Statistical Society* B27, 395–449.

Riani, M., Atkinson, A.C. and Perrotta, D. (2014) A parametric framework for the comparison of methods of very robust regression. *Statistical Science* 29, 128–143.

Rousseeuw, P.J. (1984) Least median of squares regression. *Journal of the American Statistical Association* 79, 871–880.

Rousseeuw, P., Perrotta, D., Riani, M. and Hubert, M. (2019) Robust monitoring of time series with application to fraud detection. *Econometrics and Statistics* 9, 108–121.

Rousseeuw P.J. and van Driessen K. (2000) An algorithm for positive-breakdown regression based on concentration steps. In: Gaul W., Opitz O., Schader M. (Eds.), *Data Analysis: Scientific Modeling and Practical Application* (335-346). Springer Verlag.

Schechtman, E. and Schechtman, G. (1986) Estimating the parameters in regression with uniformly distributed errors. *Journal of Statistical Computation and Simulation* 26, 269–281.

Scholz, F.W. (1980) Towards a unified definition of maximum likelihood. *Canadian Journal of Statistics* 8, 193–203.

Víšek, J.Á. (2006) The least trimmed squares. Part III: asymptotic normality. *Kybernetika* 42, 203–224.

Wagner, H.M. (1959) Linear programming techniques for regression analysis. *Journal of the American Statistical Association* 54, 206–212.

Watts, V. (1980) The almost sure representation of intermediate order statistics. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* 54, 281–285.

A Appendix

A.1 The OLS estimator in the LTS and LMS models

We start with some preliminary results on normal extreme values and uniform spacings used throughout for the analysis of the LTS and LMS models, respectively.

A.1.1 Normal extreme values and uniform spacings

Lemma A.1 *Let $\varepsilon_1, \ldots, \varepsilon_n$ be independent standard normal.*
(i) $(2\log n)^{-1/2} \max_{1\leq i\leq n} \varepsilon_i \to 1$ *in probability;*
(ii) $k_n^{1/2}(2\log n)^{-1/2}\varepsilon_{(k_n)} \to -1$ *a.s. for sequences so that* $k_n/n \to 0$ *and* $k_n/\log^3 n \to \infty$.

Proof. *(i)* We have the extreme value result $\mathsf{P}(\max_{1\leq i\leq n} \varepsilon_i \leq a_n x + b_n) \to \exp\{-\exp(-x)\}$ where $a_n^{-1} = \sqrt{2\log n}$ and $b_n = \sqrt{2\log n} - (\log\log n + \log 4\pi)/(2\sqrt{2\log n})$, see Leadbetter, Lindgren and Rootzén (1982, Theorem 1.5.3). Write

$$a_n \max_{1\leq i\leq n} \varepsilon_i = a_n^2 \frac{\max_{1\leq i\leq n} \varepsilon_i - b_n}{a_n} + a_n b_n.$$

Here, a_n^2 vanishes, the ratio converges in distribution, while $a_n b_n \to 1$.

(ii) Watts (1980) gives this as an example of his intermediate extreme value result. ■

The following lemma is well-known, see for instance Pyke (1965, §4.1).

Lemma A.2 *Let $u_1, \ldots, u_n$ be independent standard uniform with order statistics $u_{(1)} < \cdots < u_{(n)}$. The spacings $s_1, \ldots, s_n$ are defined as $s_i = u_{(i)} - u_{(i-1)}$ for $i = 2, \ldots, n$ while $s_1 = u_{(1)}$ and $s_{n+1} = 1 - u_{(n)}$. Further, let $e_1, \ldots, e_{n+1}$ be standard exponential variables $e_1, \ldots, e_{n+1}$. Then $(s_1, \ldots, s_{n+1})$ have the same distribution as $(e_1, \ldots, e_{n+1})/(e_1 + \cdots + e_{n+1})$.*

A.1.2 Proofs for OLS

Proof of Theorem 6.1. The least squares estimator satisfies $(\hat{\mu}_{OLS} - \mu)/\sigma = n^{-1}\sum_{i=1}^n \varepsilon_i$. Since $\rho = 0$ there are only right 'outliers'. Separate the 'good' observations ε_i for $i = 1, \ldots, h_n$ with maximum $\varepsilon_{(h_n)}$ and 'outliers' $\varepsilon_{h_n+j} = \varepsilon_{(h_n)} + \overline{\varepsilon}_j$ for $j = 1, \ldots, n - h_n$, to get

$$\frac{\hat{\mu}_{OLS} - \mu}{\sigma} = \frac{1}{n}\sum_{i=1}^{h_n} \varepsilon_i + (\frac{n - h_n}{n})\varepsilon_{(h_n)} + \frac{1}{n}\sum_{j=1}^{n-h_n} \overline{\varepsilon}_j. \tag{A.1}$$

For the extreme value $\varepsilon_{(h_n)}$, Lemma A.1(i) gives $\varepsilon_{(h_n)}/(2\log h_n)^{1/2} = 1 + \mathsf{o_P}(1)$, while $(n - h_n)/n \to 1 - \gamma$. The average $n^{-1}\sum_{i=1}^{h_n} \varepsilon_i$ has mean zero and variance proportional to h_n/n^2, which vanishes for any γ. In particular, $n^{-1}\sum_{i=1}^{h_n} \varepsilon_i/(2\log h_n)^{1/2} = \mathsf{o_P}(1)$. The average $(n-h_n)^{-1}\sum_{j=1}^{n-h_n} \overline{\varepsilon}_j$ converges to μ_{G} by the Law of Large Numbers since $(n-h_n)/n \to 1-\gamma > 0$. Thus, we have $n^{-1}\sum_{j=1}^{n-h_n} \overline{\varepsilon}_j \to (1-\gamma)\mu_{\mathsf{G}}$ so that $n^{-1}\sum_{j=1}^{n-h_n} \overline{\varepsilon}_j/(2\log h_n)^{1/2} = \mathsf{o_P}(1)$. Combine to see that $(2\log h_n)^{-1/2}(\hat{\mu}_{OLS} - \mu)/\sigma = 1 - \gamma + \mathsf{o_P}(1)$. ■

Proof of Theorem 6.2. Proceed as in the proof of Theorem 6.1, apart from a different analysis of the order statistic $\varepsilon_{(h_n)}$. Specifically, for $1 \leq i \leq h_n$ then ε_i is uniform on $[-1,1]$, so that $u_i = (\varepsilon_i + 1)/2$ is standard uniform. The uniform spacings Lemma A.2 shows that $1 - u_{(h_n)}$ is distributed as $e_{h_n+1}/\sum_{i=1}^{h_n+1} e_i$, which vanishes by the Law of Large Numbers. Thus, $\varepsilon_{(h_n)} = 2u_{(h_n)} - 1 = 1 + \mathsf{o_P}(1)$. This now has the same order as the last sum in (A.1), so that $(\hat{\mu}_{OLS} - \mu)/\sigma = (1-\gamma)(1+\mu_{\mathsf{G}}) + \mathsf{o_P}(1)$. ■

A.2 The LTS estimator in the LTS model

We start with some preliminary results on marked empirical processes evaluated at quantiles.

A.2.1 Marked empirical processes evaluated at quantiles

Let ε_i for $i = 1, \ldots, n$ be i.i.d. continuous random variables. Let $p \in \mathbb{N}$. Consider the marked empirical distribution, compensator and process defined for $c > 0$ by

$$\mathsf{F}_n^p(c) = n^{-1}\sum_{i=1}^{n} \varepsilon_i^p 1_{(\varepsilon_i \leq c)}, \qquad \overline{\mathsf{F}}^p(c) = \mathsf{E}\varepsilon_1^p 1_{(\varepsilon_1 \leq c)}, \qquad \mathbb{F}_n^p(c) = n^{1/2}\{\mathsf{F}_n^p(c) - \overline{\mathsf{F}}^p(c)\}. \quad \text{(A.2)}$$

For $p = 0$, let $\mathsf{F}_n^0 = \mathsf{F}_n$. We also define the quantile function $\mathsf{Q}(\psi) = \inf\{c : \mathsf{F}(c) \geq \psi\}$ and the empirical quantiles $\mathsf{Q}_n(\psi) = \inf\{c : \mathsf{F}_n(c) \geq \psi\}$. The first result concerns the lower tail of quantiles of a non-negative random variable.

Lemma A.3 *Suppose* $\mathsf{F}(c) = 0$ *for* $c < 0$. *Let* $\psi_n = \mathsf{o_P}(1)$. *Then* $\mathsf{Q}_n(\psi_n) = \mathsf{O_P}(1)$.

Proof. Let a small $\epsilon > 0$ be given. Then a finite $x \geq 0$ exists to that $f = \mathsf{F}(x) \geq 1 - \epsilon$. We show $\mathcal{P}_n = \mathsf{P}(A_n) \leq 2\epsilon$ where $A_n = \{\mathsf{Q}_n(\psi_n) \geq x\}$. Applying F_n, we get $A_n = \{\psi_n \geq \mathsf{F}_n(x)\}$. By the Law of Large Numbers, $\mathsf{F}_n(x) = f + \mathsf{o_P}(1)$. Hence, if $B_n = \{\mathsf{F}_n(x) > f - \epsilon\}$ then $\mathsf{P}_n(B_n) > 1 - \epsilon$ for large n. Since $A_n = (A_n \cap B_n) \cup (A_n \cap B_n^c)$, we have $A_n \subset (A_n \cap B_n) \cup B_n^c$. Here, $\mathsf{P}(B_n^c) \leq \epsilon$ by construction. Moreover, $A_n \cap B_n \subset (\psi_n > f - \epsilon)$ where $\mathsf{P}(\psi_n > f - \epsilon) \leq \epsilon$ for large n since $\psi_n = \mathsf{o_P}(1)$ by assumption. Thus, $\mathcal{P}_n \leq 2\epsilon$. ■

The following lemma follows from the theory of empirical quantile processes.

Lemma A.4 *Suppose* F *is regular (Definition 6.1). Then, for all* $\zeta > 0$,
(*a*) $n^{1/2}[\mathsf{F}_n\{\mathsf{Q}(\psi)\} - \psi]$ *converges in distribution on* $D[0,1]$ *to a Brownian bridge;*
(*b*) $\sup_{0 \leq \psi \leq 1} |n^{1/2}[\mathsf{F}\{\mathsf{Q}_n(\psi)\} - \psi] + n^{1/2}[\mathsf{F}_n\{\mathsf{Q}(\psi)\} - \psi]| \overset{a.s.}{=} \mathsf{o}(n^{\zeta - 1/4})$.

Proof. (*a*) This is Theorem 16.4 of Billingsley (1968).
(*b*) Let $D(\psi) = \mathsf{f}\{\mathsf{Q}(\psi)\} n^{1/2}\{\mathsf{Q}_n(\psi) - \mathsf{Q}(\psi)\}$ and write the object of interest as the sum of $n^{1/2}[\mathsf{F}\{\mathsf{Q}_n(\psi)\} - \psi] - D(\psi)$ and $n^{1/2}[\mathsf{F}_n\{\mathsf{Q}(\psi)\} - \psi] + D(\psi)$. These two terms are $\mathrm{o}(n^{\zeta - 1/4})$ *a.s.* by Corollaries 6.2.1 and 6.2.2 of Csörgő (1983), uniformly in ψ. ■

The following result is a special case of the Glivenko-Cantelli result of Berenguer-Rico, Johansen and Nielsen (2019, Theorem 3.2).

Lemma A.5 *Let* $p \in \mathbb{N}_0$. *Suppose* $\int_{\mathbb{R}} |x|^{2p} d\mathsf{F}(x) < \infty$. *Then* $\sup_{c \in \mathbb{R}} |\mathbb{F}_n^p(c)| = \mathsf{o_P}(n^{1/2})$.

The next result is inspired by Johansen and Nielsen (2016a, Lemma D.11).

Lemma A.6 *Let $p \in \mathbb{N}_0$. Suppose F is regular and $\int_{\mathbb{R}} |x|^q d\mathsf{F}(x) < \infty$ for some $q > 2p$. Then,*

$$\sup_{1/(n+1) \leq \psi \leq n/(n+1)} |\mathsf{F}_n^p\{\mathsf{Q}_n(\psi)\} - \overline{\mathsf{F}}^p\{\mathsf{Q}(\psi)\}| = \mathrm{o}_{\mathsf{P}}(1).$$

Proof. Let $R_n(\psi) = \mathsf{F}_n^p\{\mathsf{Q}_n(\psi)\} - \overline{\mathsf{F}}^p\{\mathsf{Q}(\psi)\}$. Add and subtract $\overline{\mathsf{F}}_n^p\{\mathsf{Q}_n(\psi)\}$ to get $R_n(\psi) = \sum_{\ell=1}^2 R_{n\ell}(\psi)$ where

$$\begin{aligned} R_{n1}(\psi) &= \mathsf{F}_n^p\{\mathsf{Q}_n(\psi)\} - \overline{\mathsf{F}}_n^p\{\mathsf{Q}_n(\psi)\} = n^{-1/2}\mathbb{F}_n^p\{\mathsf{Q}_n(\psi)\}, \\ R_{n2}(\psi) &= \overline{\mathsf{F}}_n^p\{\mathsf{Q}_n(\psi)\} - \overline{\mathsf{F}}^p\{\mathsf{Q}(\psi)\}. \end{aligned}$$

We show that each of $R_{n\ell}(\psi)$ vanishes uniformly for ψ in $\Psi_n = [1/(n+1), n/(n+1)]$.

1. *The term $R_{n1}(\psi)$.* Lemma A.5 shows that $\mathbb{F}_n^p(c) = \mathrm{o}_{\mathsf{P}}(n^{1/2})$ uniformly in $c \in \mathbb{R}$. In turn, $\mathbb{F}_n^p\{\mathsf{Q}_n(\psi)\} = \mathrm{o}_{\mathsf{P}}(n^{1/2})$ uniformly in $0 < \psi < 1$.

2. *The term $R_{n2}(\psi)$.* We write Q_n in terms of Q. Note $\mathsf{Q}_n(\psi) = \mathsf{Q}[\mathsf{F}\{\mathsf{Q}_n(\psi)\}]$. Expand $\mathsf{F}\{\mathsf{Q}_n(\psi)\} = \psi + n^{-1/2}\phi_n$ where $\phi_n = n^{1/2}[\mathsf{F}\{\mathsf{Q}_n(\psi)\} - \psi]$. Thus,

$$\mathsf{Q}_n(\psi) = \mathsf{Q}(\psi + n^{-1/2}\phi_n). \tag{A.3}$$

Let $R_{n2}(\psi) = S_n(\psi, \phi_n)$, where $S_n(\psi, \phi) = \overline{\mathsf{F}}_n^p\{\mathsf{Q}(\psi + n^{-1/2}\phi)\} - \overline{\mathsf{F}}^p\{\mathsf{Q}(\psi)\}$. Write $S_n(\psi, \phi)$ as an integral, change variable $u = \mathsf{F}(c)$, $du = \mathsf{f}(c)dc$, so that $c = \mathsf{Q}(u)$, and use the mean value theorem to get, $\forall \phi \in \mathbb{R}, \exists \psi^*$ so that $|\psi^* - \psi| \leq n^{-1/2}\phi$, so that

$$S_n(\psi, \phi) = \int_{\mathsf{Q}(\psi)}^{\mathsf{Q}(\psi + n^{-1/2}\phi)} c^p \mathsf{f}(c) dc = \int_{\psi}^{\psi + n^{-1/2}\phi} \{\mathsf{Q}(u)\}^p du = \{\mathsf{Q}(\psi^*)\}^p n^{-1/2}\phi.$$

Note that $\mathsf{Q}(\psi^*)$ belongs to the interval from $\mathsf{Q}(\psi)$ to $\mathsf{Q}(\psi + n^{-1/2}\phi)$, see (A.3). Thus, when inserting ϕ_n we get $R_{n2}(\psi) = \{\mathsf{Q}(\psi_n^*)\}^p n^{-1/2}\phi_n$, where $\mathsf{Q}(\psi_n^*)$ belongs to the interval from $\mathsf{Q}(\psi)$ to $\mathsf{Q}(\psi + n^{-1/2}\phi_n) = \mathsf{Q}_n(\psi)$. Lemma A.4 shows that ϕ_n is bounded in probability uniformly in $0 \leq \psi \leq 1$ and in particular on Ψ_n. Thus, $R_{n2}(\psi)$ vanishes uniformly in $\psi \in \Psi_n$ if $\{\mathsf{Q}(\psi_n^*)\}^p = \mathrm{o}_{\mathsf{P}}(n^{1/2})$ uniformly in $\psi \in \Psi_n$. It suffices that $\{\mathsf{Q}(\psi)\}^p$ and $\{\mathsf{Q}_n(\psi)\}^p$ are $\mathrm{o}_{\mathsf{P}}(n^{1/2})$ uniformly in $\psi \in \Psi_n$. For $p = 0$ this is satisfied as $x^0 = 1$.

Consider $\mathsf{Q}_n(\psi)$ for $p \in \mathbb{N}$, $\psi \in \Psi_n$. Bound $|\mathsf{Q}_n(\psi)| \leq \max_{1 \leq i \leq n} |\varepsilon_i|$. For any $\epsilon > 0$ we have $\mathcal{P}_n = \mathsf{P}\{\max_{1 \leq i \leq n} |\varepsilon_i| \geq \epsilon n^{1/(2p)}\} = \mathsf{P}\bigcup_{i=1}^n (|\varepsilon_i|^p \geq \epsilon^p n^{1/2})$. Boole's and Markov's inequalities give $\mathcal{P}_n \leq \sum_{i=1}^n \mathsf{P}(|\varepsilon_i|^p \geq \epsilon^p n^{1/2}) \leq \epsilon^{-q} n^{1-q/(2p)} \mathsf{E}|\varepsilon_i|^q$, which vanishes, since $q > 2p$, so that n has a negative power. Hence, $|\mathsf{Q}_n(\psi)|^p = \mathrm{o}_{\mathsf{P}}(n^{1/2})$.

Consider $\mathsf{Q}_n(\psi)$ for $p \in \mathbb{N}$, $\psi \in \Psi_n$. Bound $|\mathsf{Q}(\psi)| \leq c_n$, where c_n satisfies $(n+1)^{-1} = \mathsf{P}(|\varepsilon_1| > c_n)$. We show $c_n^p = \mathrm{o}(n^{1/2})$. By the Markov inequality $\mathsf{P}(|\varepsilon_1| > c_n) \leq c_n^{-q} \mathsf{E}|\varepsilon_i|^q$, so that $c_n = \mathrm{O}(n^{1/q})$. Since $q > 2p$ we get $c_n^p = \mathrm{o}(n^{1/2})$. Hence, $|\mathsf{Q}(\psi)|^p = \mathrm{o}(n^{1/2})$. ∎

A.2.2 Proofs for LTS

We consider the LTS location-scale Model 1 and the sequence of data generating processes outlined in §6.1. The most difficult part of the proof is to analyze the minimizer $\hat{\delta}_{LTS}$ of $\hat{\sigma}_\delta^2$, defined in (4.6), which counts the 'outliers' to the left of the 'good' observations.

It has to be shown that $\hat{\delta}_{LTS}$ is close to the $\mathsf{Binomial}(n-h_n,\rho)$-distributed variable $\delta_n = \sum_{j\in\zeta_n} 1_{(\varepsilon_j<\min_{i\in\zeta_n}\varepsilon_i)}$. In the following lemmas, we condition on the sequence δ_n, so that the randomness stems from 'good' errors $\varepsilon_{(\delta_n+1)},\ldots,\varepsilon_{(\delta_n+h_n)}$ and the magnitudes of the 'outliers', $\underline{\varepsilon}_{(\delta_n+1-j)}$ for $j\le\delta_n$ and $\overline{\varepsilon}_{(j-\delta_n-h_n)}$ for $j>\delta_n+h_n$. The unconditional statements in the Theorems about $\hat{\delta}_{LTS}$ are then derived as follows. If $\mathsf{P}(\hat{\delta}_{LTS}-\delta_n\in I_n|\delta_n)\to 0$ for an interval I_n and a sequence δ_n then by the law of iterated expectations

$$\mathsf{P}(\hat{\delta}_{LTS}-\delta_n\in I_n)=\mathsf{E}\{\mathsf{P}(\hat{\delta}_{LTS}-\delta_n\in I_n|\delta_n)\}\to 0, \tag{A.4}$$

due to the dominated converges theorem, because $\mathsf{P}(\hat{\delta}_{LTS}-\delta_n\in I_n|\delta_n)$ is bounded and vanishes. It is convenient to define

$$\underline{s}_n=(2\log h_n)^5, \qquad \overline{s}_n=(2\log h_n)^{-1/4}h_n. \tag{A.5}$$

For large n, we have $0<\underline{s}_n<\overline{s}_n$. Numerical calculations indicate that this require $h_n>2.042\times 10^8$, which is of course sufficient for asymptotic analysis.

The first Lemma concerns $\hat{\mu}_\delta,\hat{\sigma}^2_\delta$ defined in (4.6), when $\delta=\delta_n$ is known.

Lemma A.7 *Consider the LTS Model 1 and the sequence of data generating processes in §6.1. Then, conditional on δ_n,*

$$h_n^{1/2}(\hat{\mu}_{\delta_n}-\mu)/\sigma\xrightarrow{\mathsf{D}}\mathsf{N}(0,1), \qquad h_n^{1/2}(\hat{\sigma}^2_{\delta_n}-\sigma^2)/\sigma^2\xrightarrow{\mathsf{D}}\mathsf{N}(0,2),$$

where $n^{1/2}(\hat{\mu}_{\delta_n}-\mu)$ and $n^{1/2}(\hat{\sigma}^2_{\delta_n}-\sigma^2)$ are asymptotically independent.

Proof. By construction, δ_n, the number of left 'outlier' errors, is independent of the 'good' errors. Thus, conditioning on δ_n does not change the distribution of $\hat{\mu}_{\delta_n}$, $\hat{\sigma}^2_{\delta_n}$. As the 'good' errors are normal then $h_n^{1/2}(\hat{\mu}_{\delta_n}-\mu)/\sigma$ is standard normal while $h_n\hat{\sigma}^2_{\delta_n}/\sigma^2$ is $\chi^2_{h_n-1}$, so that $h_n^{1/2}(\hat{\sigma}^2_{\delta_n}-\sigma^2)/\sigma^2$ is asymptotically normal by the Central Limit Theorem. ■

The next lemma concerns cases where $\hat{s}=\hat{\delta}_{LTS}-\delta_n$ is small. We write $\varepsilon^p_{(\hat{\delta}+i)}$ for $\{\varepsilon_{(\hat{\delta}+i)}\}^p$.

Lemma A.8 *Consider the LTS Model 1 and the sequence of data generating processes in §6.1. Suppose $\hat{\delta}_{LTS}=\delta_n+\mathsf{o_P}(h_n^{1/2-\omega})$ for some $\omega>0$. Then, conditional on δ_n, we get $h_n^{1/2}(\hat{\mu}_{LTS}-\hat{\mu}_{\delta_n})$ and $h_n^{1/2}(\hat{\sigma}^2_{LTS}-\hat{\sigma}^2_{\delta_n})$ are $\mathsf{o_P}(1)$.*

Proof. We suppress the index LTS. The estimators $\hat{\mu}$ and $\hat{\sigma}^2$ are formed from the sample moments $h_n^{-1}\sum_{i=1}^{h_n}\varepsilon^p_{(\hat{\delta}+i)}$ for $p=1,2$. Thus, we must show that $\mathcal{E}_{np}=\sum_{i=1}^{h_n}\varepsilon^p_{(\hat{\delta}+i)}-\sum_{i=1}^{h_n}\varepsilon^p_{(\delta_n+i)}=\mathsf{o_P}(h_n^{1/2})$.

It suffices to consider $\hat{\delta}>\delta_n$ and assume $\rho<1$ as remarked in §6.1. Then

$$\mathcal{E}_{np}=\sum_{i=1}^{h_n+\delta_n-\hat{\delta}}\varepsilon^p_{(\hat{\delta}+i)}+\sum_{i=h_n+\delta_n-\hat{\delta}+1}^{h_n}\varepsilon^p_{(\hat{\delta}+i)}-\sum_{i=1}^{\hat{\delta}-\delta_n}\varepsilon^p_{(\delta_n+i)}-\sum_{i=\hat{\delta}-\delta_n+1}^{h_n}\varepsilon^p_{(\delta_n+i)}.$$

The first and the fourth sum cancel, so that

$$\mathcal{E}_{np} = \sum_{i=h_n+\delta_n-\hat{\delta}+1}^{h_n} \varepsilon^p_{(\hat{\delta}+i)} - \sum_{i=1}^{\hat{\delta}-\delta_n} \varepsilon^p_{(\delta_n+i)} = \sum_{i=1}^{\hat{\delta}-\delta_n} \{\varepsilon^p_{(\delta_n+h_n+i)} - \varepsilon^p_{(\delta_n+i)}\}.$$

Since $\varepsilon_{(\delta_n+h_n+j)} = \varepsilon_{(\delta_n+h_n)} + \overline{\varepsilon}_{(j)}$ we get

$$\mathcal{E}_{np} = \sum_{i=1}^{\hat{\delta}-\delta_n} [\{\varepsilon_{(\delta_n+h_n)} + \overline{\varepsilon}_{(i)}\}^p - \varepsilon^p_{(\delta_n+i)}],$$

where $\varepsilon_{(\delta_n+h_n)}$ is the maximum and $\varepsilon_{(\delta_n+i)}$ is the ith order statistic of the 'good' errors. We want to prove that if $\hat{\delta} - \delta_n = \mathsf{o_P}(h_n^{1/2-\omega})$ for some $\omega > 0$, then $\mathcal{E}_{np} = \mathsf{o_P}(h_n^{1/2})$.

For $p = 1$ we find $\varepsilon_{(\delta_n+i)} \geq \varepsilon_{(\delta_n+1)}$ and $\overline{\varepsilon}_{(i)} \leq \overline{\varepsilon}_{(\hat{\delta}-\delta_n)}$, so that

$$\mathcal{E}_{n1} \leq (\hat{\delta} - \delta_n)\{\varepsilon_{(\delta_n+h_n)} - \varepsilon_{(\delta_n+1)} + \overline{\varepsilon}_{(\hat{\delta}-\delta_n)}\}.$$

The normal extreme value Lemma A.1(i) shows that $\varepsilon_{(\delta_n+1)}$ and $\varepsilon_{(\delta_n+h_n)}$ are $\mathrm{O_P}\{(2\log h_n)^{1/2}\}$ $= \mathsf{o_P}(h_n^{\omega})$. Next, $\overline{\varepsilon}^p_{(\hat{\delta}-\delta_n)}$ is the $(\hat{\delta} - \delta_n)/\overline{n}$ empirical quantile in the $\overline{\mathsf{G}}$ distribution. By assumption $\hat{\delta} - \delta_n = \mathsf{o_P}(h_n^{1/2-\omega})$ and $\overline{n}/h_n \to \overline{\omega} = (1-\gamma)(1-\rho)/\gamma > 0$ so that $(\hat{\delta} - \delta_n)/\overline{n} = \mathsf{o_P}(h_n^{-1/2-\omega}) = \mathsf{o_P}(1)$. Lemma A.3 then shows $\overline{\varepsilon}_{(\hat{\delta}-\delta_n)} = \mathrm{O_P}(1)$. In combination, $\mathcal{E}_{n1} = \mathsf{o_P}(h_n^{1/2-\omega})\{\mathsf{o_P}(h_n^{\omega}) + \mathrm{O_P}(1)\} = \mathsf{o_P}(n^{1/2})$.

For $p = 2$ we find similarly, using the inequality $(x+y)^2 \leq 2(x^2+y^2)$,

$$\mathcal{E}_{n2} \leq \sum_{i=1}^{\hat{\delta}-\delta_n} \{2\varepsilon^2_{(\delta_n+h_n)} + 2\overline{\varepsilon}^2_{(i)} - \varepsilon^2_{(\delta_n+i)}\} \leq 2(\hat{\delta} - \delta_n)\{\varepsilon^2_{(\delta_n+h_n)} + \overline{\varepsilon}^2_{(\hat{\delta}-\delta_n)}\},$$

noting that $\overline{\varepsilon}^2_{(i)} \leq \overline{\varepsilon}^2_{(\hat{\delta}-\delta_n)}$ and $\varepsilon^2_{(\delta_n+i)} \geq 0$. Apply the above bounds $\varepsilon^2_{(\delta_n+h_n)} = \mathrm{O_P}(2\log h_n) = \mathsf{o_P}(h_n^{\omega})$ and $\overline{\varepsilon}_{(\hat{\delta}-\delta_n)} = \mathrm{O_P}(1)$ to get $\mathcal{E}_{n2} = \mathsf{o_P}(n^{1/2})$. ■

The next Lemma is the main ingredient to showing consistency of $\hat{\delta}_{LTS}$, when less than half of the observations are 'outliers'.

Lemma A.9 *Consider the LTS Model 1 and the sequence of data generating processes in §6.1 where $\rho < 1$. Recall $\underline{s}_n = (2\log h_n)^5$ and $\overline{s}_n = (2\log h_n)^{-1/4} h_n$ from (A.5). Then, conditional on δ_n, we have $\min_{\underline{s}_n \leq s \leq h_n - \overline{s}_n} h_n(\hat{\sigma}^2_{\delta_n+s} - \hat{\sigma}^2_{\delta_n}) \to \infty$ in probability.*

Proof. Recall that δ_n is the number of left 'outliers' and $\varepsilon_{(\delta_n+1)}, \ldots, \varepsilon_{(\delta_n+h_n)}$ are the ordered 'good' observations. If $\rho < 1$ then $\overline{n}/n \to \overline{\omega} = (1-\rho)(1-\gamma) > 0$. Let $\varepsilon_{(\delta_n+h_n+j)} = \varepsilon_{(\delta_n+h_n)} + \overline{\varepsilon}_{(j)}$ for $1 \leq j \leq \overline{n}$.

Expand, see §B.3 in Supplementary material,

$$S_s = (\hat{\sigma}^2_{\delta_n+s} - \hat{\sigma}^2_{\delta_n})/\sigma = (s/h_n)\{1 - (s/h_n)\}\varepsilon^2_{(\delta_n+h_n)} + A_n, \tag{A.6}$$

with coefficients $A_n = A_{n1} - A_{n2} + 2A_{n3} - 2A_{n4}$ where

$$A_{n1} = \frac{1}{h_n}\sum_{j=1}^{s}\overline{\varepsilon}^2_{(j)} - \{\frac{1}{h_n}\sum_{j=1}^{s}\overline{\varepsilon}_{(j)}\}^2, \quad A_{n3} = \frac{1}{h_n}\sum_{j=1}^{s}\overline{\varepsilon}_{(j)}\frac{1}{h_n}\sum_{i=s+1}^{h_n}\{\varepsilon_{(\delta_n+h_n)} - \varepsilon_{(\delta_n+i)}\},$$

$$A_{n2} = \frac{1}{h_n}\sum_{i=1}^{s}\varepsilon^2_{(\delta_n+i)} - \{\frac{1}{h_n}\sum_{i=1}^{s}\varepsilon_{(\delta_n+i)}\}^2, \quad A_{n4} = \frac{1}{h_n}\sum_{i=s+1}^{h_n}\varepsilon_{(\delta_n+i)}\frac{1}{h_n}\sum_{i=1}^{s}\{\varepsilon_{(\delta_n+h_n)} - \varepsilon_{(\delta_n+i)}\}.$$

We find a lower bound for A_n. Notice $A_{n1}, A_{n3} \geq 0$, which only requires $0 \leq s \leq h_n$ and $0 < h_n$. Thus, $A_n \geq -A_{n2} - 2A_{n4}$, which does not involve the 'outliers'. For A_{n2}, bound the sample variance by a sample second moments to get $A_{n2} \leq h_n^{-1}\sum_{i=1}^{s}\varepsilon^2_{(\delta_n+i)} = B_{n2}$ say. For A_{n4}, use Jensen's inequality, add further summand and use the Law of Large Numbers for the unordered normal 'good' ε_{δ_n+i} to get

$$|\frac{1}{h_n}\sum_{i=s+1}^{h_n}\varepsilon_{(\delta_n+i)}|^2 \leq \frac{1}{h_n}\sum_{i=s+1}^{h_n}\varepsilon^2_{(\delta_n+i)} \leq \frac{1}{h_n}\sum_{i=1}^{h_n}\varepsilon^2_{(\delta_n+i)} = \frac{1}{h_n}\sum_{i=1}^{h_n}\varepsilon^2_{\delta_n+i} \xrightarrow{\mathsf{P}} 1. \tag{A.7}$$

Further, we have $h_n^{-1}\sum_{i=1}^{s}\{\varepsilon_{(\delta_n+h_n)} - \varepsilon_{(\delta_n+i)}\} \leq (s/h_n)\{\varepsilon_{(\delta_n+h_n)} - \varepsilon_{(\delta_n+1)}\} = B_{n4}$ say, so that $|A_{n4}| \leq B_{n4}\{1 + \mathrm{o}_{\mathsf{P}}(1)\}$, where the remainder term coming from (A.7) is uniform in s. In combination,

$$S_s \geq (s/h_n)(1 - s/h_n)\varepsilon^2_{(\delta_n+h_n)} - B_{n2} - 2B_{n4}\{1 + \mathrm{o}_{\mathsf{P}}(1)\}, \tag{A.8}$$

where the $\mathrm{o}_{\mathsf{P}}(1)$ term is uniform in s. This, we analyze separately for $\underline{s}_n \leq s \leq \overline{s}_n$ and $\overline{s}_n \leq s \leq h_n - \overline{s}_n$.

1. *Consider $\overline{s}_n \leq s \leq h_n - \overline{s}_n$ where $\overline{s}_n/h_n = (2\log h_n)^{-1/4}$*, see (A.5). Since the function $x(1-x)$ is concave with roots at $x = 0$ and $x = 1$, so that, for $x_0 < x < 1 - x_0$ with $0 < x_0 < 1/2$, we get $x(1-x) \geq x_0/2$. Thus, $2(s/h_n)(1 - s/h_n) \geq \overline{s}_n/h_n = (2\log h_n)^{-1/4}$ on the considered range. We bound $B_{n2} \leq 1 + \mathrm{o}_{\mathsf{P}}(1)$ uniformly in s as in (A.7). In B_{n4}, we use $s/h_n \leq 1$. Thus, (A.8) reduces to

$$2S_s \geq (2\log h_n)^{-1/4}\varepsilon^2_{(\delta_n+h_n)} - 2\{1 + 2\varepsilon_{(\delta_n+h_n)} - 2\varepsilon_{(\delta_n+1)}\}\{1 + \mathrm{o}_{\mathsf{P}}(1)\}.$$

Lemma A.1(i) shows that

$$\varepsilon_{(\delta_n+1)}/(2\log h_n)^{1/2} \xrightarrow{\mathsf{P}} -1, \qquad \varepsilon_{(\delta_n+h_n)}/(2\log h_n)^{1/2} \xrightarrow{\mathsf{P}} 1. \tag{A.9}$$

Thus, $\min_{\overline{s}_n \leq s \leq h_n - \overline{s}_n} 2S_s/(2\log h_n)^{3/4} \geq 1 + \mathrm{o}_{\mathsf{P}}(1)$.

2. Consider $\underline{s}_n \leq s \leq \overline{s}_n$ *where* $\underline{s}_n = (2\log h_n)^5$ *and* $\overline{s}_n = (2\log h_n)^{-1/4}h_n$, see (A.5). We find lower bounds for the $B_{n\ell}$ terms in (A.8). For B_{n2}, let $r_n = (2\log h_n)^4$. We apply Lemma A.1(ii) for $k_n = r_n$ and $k_n = \overline{s}_n$, noting that $r_n \leq \overline{s}_n$ and $\overline{s}_n/h_n \to 0$ while $r_n/(2\log h_n)^3 \to \infty$, and get

$$r_n^{1/2}\varepsilon_{(\delta_n+r_n)}/(2\log h_n)^{1/2} \xrightarrow{a.s.} -1, \qquad \overline{s}_n^{1/2}\varepsilon_{(\delta_n+\overline{s}_n)}/(2\log h_n)^{1/2} \xrightarrow{a.s.} -1. \tag{A.10}$$

Since $\varepsilon_{(\delta_n+\overline{s}_n)}$, when normalized, has a negative limit $a.s.$, then Egoroff's Theorem, see Davidson (1994, Theorem 18.4), shows that a set Ω_ϵ with large probability exists so that on Ω_ϵ

and for large n then $\varepsilon_{(\delta_n+\overline{s}_n)} < 0$. In particular, $\varepsilon_{(\delta_n+1)} < \varepsilon_{(\delta_n+r_n)} < \varepsilon_{(\delta_n+\overline{s}_n)} < 0$. On the set Ω_ϵ, noting $r_n \le s \le \overline{s}_n$, we can bound

$$B_{n2} = \frac{1}{h_n}\sum_{i=1}^{r_n} \varepsilon^2_{(\delta_n+i)} + \frac{1}{h_n}\sum_{i=r_n+1}^{s} \varepsilon^2_{(\delta_n+i)} \le \frac{1}{h_n}\{r_n\varepsilon^2_{(\delta_n+1)} + s\varepsilon^2_{(\delta_n+r_n)}\}.$$

We have $s/h_n \le \overline{s}_n/h_n \to 0$ so that $2(1-s/h_n) \ge 1$ and $2(s/h_n)(1-s/h_n) \ge s/h_n$ for large n. Therefore, the expansion (A.8) is bounded from below by

$$2S_s \ge \frac{s}{h_n}[\varepsilon^2_{(\delta_n+h_n)} - 2\frac{r_n}{s}\varepsilon^2_{(\delta_n+1)} - 2\varepsilon^2_{(\delta_n+r_n)} - 4\{\varepsilon_{(\delta_n+h_n)} - \varepsilon_{(\delta_n+1)}\}]\{1+\mathsf{o}_\mathsf{P}(1)\}.$$

Applying the limits from (A.9), (A.10) and noting $r_n/s \le r_n/\underline{s}_n \to 0$ while $r_n \to \infty$ we get $S_s/\log h_n \ge (s/h_n)\{1+\mathsf{o}_\mathsf{P}(1)\}$ where the $\mathsf{o}_\mathsf{P}(1)$ term is uniform in $\underline{s}_n \le s \le \overline{s}_n$. The minimum is taken at the left end point, so that $\min_{\underline{s}_n \le s \le \overline{s}_n} S_s/\log h_n \ge \{(2\log h_n)^5/h_n\}\{1+\mathsf{o}_\mathsf{P}(1)\}$ and therefore $h_n S_s \to \infty$ in probability, uniformly in $\underline{s}_n \le s \le \overline{s}_n$. ■

Proof of Theorem 6.3. We will show that $\hat{s} = \hat{\delta}_{LTS} - \delta_n = \mathsf{o}_\mathsf{P}(h_n^\alpha)$ for any $\alpha > 0$. It suffices to consider the case where $\mathsf{P}(\hat{\delta}_{LTS} - \delta_n > h_n^\alpha) \to 0$ and where $\rho < 1$ as remarked in §6.1. We consider $\gamma, \rho < 1$ so that $\overline{\omega} = (1-\rho)(1-\gamma)/\gamma$ satisfies $0 < \overline{\omega} < 1$. In particular, the constraint $\overline{\omega} < 1$ applies when $1/2 < \gamma < 1$ as required in the Theorem.

Recall $\underline{s}_n, \overline{s}_n$ from (A.5). In particular, $h_n^\alpha > \underline{s}_n$ for large n, so that $\mathsf{P}(\hat{s} > h_n^\alpha) \le \mathsf{P}(\hat{s} \ge \underline{s}_n)$. We show, that the latter probability vanishes.

We have $\hat{s} \le \overline{n} \le n - h_n$, since there are $\overline{n}$ right 'outliers' and $n - h_n$ 'outliers' in total. Here, h_n satisfies $h_n/n \to \gamma$ by (6.1), while $\overline{n}/h_n \to \overline{\omega}$. Thus, for large n, we have that $\hat{s}/h_n \le 1 - \overline{s}_n/h_n$ and it suffices to show $\mathsf{P}(\underline{s}_n \le \hat{s} \le h_n - \overline{s}_n)$ vanishes.

Now, $\hat{s}$ is the minimizer to $\hat{\sigma}^2_{\delta_n+s} - \hat{\sigma}^2_{\delta_n}$, which is zero for $s = 0$. Thus, it suffices to show that $\min_{\underline{s}_n \le s \le h_n - \overline{s}_n} h_n(\hat{\sigma}^2_{\delta_n+s} - \hat{\sigma}^2_{\delta_n}) \to \infty$ in probability. This follows from Lemma A.9.

The Lemmas A.7, A.8 give the limiting results for $\hat{\mu}, \hat{\sigma}^2$. ■

The next lemma is needed when there are more than half of the observations are 'outliers'. Compared to Lemma A.9 we find that $\hat{\sigma}^2_\delta$ is not diverging and additional regularity conditions are needed to ensure that $\hat{\sigma}^2_\delta > \hat{\sigma}^2_{\delta_n}$.

Lemma A.10 *Consider the LTS Model 1 and the sequence of data generating processes in §6.1 where* $1 \le \overline{\omega} = (1-\rho)(1-\gamma)/\gamma < \infty$. *Suppose Assumption 6.1(i) holds, so that* $\overline{\mathsf{G}}$ *is regular with moments of higher order than 4. Recall* $\overline{s}_n = (2\log h_n)^{-1/4}h_n$ *from (A.5). Then, conditional on* δ_n, *an* $\epsilon > 0$ *exists so that* $\min_{h_n - \overline{s}_n \le s \le \overline{n}}(\hat{\sigma}^2_{\delta_n+s} - \hat{\sigma}^2_{\delta_n}) \ge \epsilon + \mathsf{o}_\mathsf{P}(1)$ *for large* n.

Proof. The errors $\varepsilon_{(\delta_n+i)}$ are standard normal order statistics for $1 \le i \le h_n$ and $\varepsilon_{(\delta_n+h_n+j)} = \varepsilon_{(\delta_n+h_n)} + \overline{\varepsilon}_{(j)}$ for $0 \le j \le \overline{n}$, where $\overline{\varepsilon}_j$ is $\overline{\mathsf{G}}$-distributed, with the convention that $\overline{\varepsilon}_0 = 0$.

It suffices to show that $\hat{\sigma}^2_{\delta_n+s}/\sigma^2 \ge 1 + \epsilon + \mathsf{o}_\mathsf{P}(1)$ uniformly in $h_n - \overline{s}_n \le s \le \overline{n}$, since $\hat{\sigma}^2_{\delta_n}/\sigma^2 = 1 + \mathsf{o}_\mathsf{P}(1)$ by Lemma A.7, so that $(\hat{\sigma}^2_{\delta_n+s} - \hat{\sigma}^2_{\delta_n})/\sigma^2 \ge \epsilon + \mathsf{o}_\mathsf{P}(1)$. We consider separately the cases $h_n \le s \le \overline{n}$ and $h_n - \overline{s}_n \le s < h_n$.

1. *Consider* $h_n \le s \le \overline{n}$. In this case, $\hat{\sigma}^2_{\delta_n+s}$ is the sample variance of $\varepsilon_{(\delta_n+s+j)}$ for $1 \le j \le h_n$, where $\varepsilon_{(\delta_n+s+j)}$ are 'outliers'. As sample variances are invariant to the level, $\hat{\sigma}^2_{\delta_n+s}$ is the sample variance of $\varepsilon_{(\delta_n+s+j)} - \varepsilon_{(\delta_n+h_n)} = \overline{\varepsilon}_{(s-h_n+j)}$ for $1 \le j \le h_n$.

We express $\hat{\sigma}^2_{\delta_n+s}/\sigma^2$ in terms of the empirical distribution function for the $\overline{n}$ right 'outliers' with distribution $\overline{\mathsf{G}}$. First, write $\sum_{j=1}^{h_n} \overline{\varepsilon}^p_{(s-h_n+j)} = \sum_{k=1}^{\overline{n}} \overline{\varepsilon}^p_{(k)} 1_{(s-h_n<k\leq s)}$ for $p = 0, 1, 2$. This sum over the order statistics $\overline{\varepsilon}_{(k)}$ can be written as a sum over the unordered observations $\overline{\varepsilon}_k$ through $\sum_{j=1}^{h_n} \overline{\varepsilon}^p_{(s-h_n+j)} = \sum_{k=1}^{\overline{n}} \overline{\varepsilon}^p_k 1_{\{\overline{\varepsilon}_{(s-h_n)}<\overline{\varepsilon}_k\leq\overline{\varepsilon}_{(s)}\}}$. Let $\overline{\mathsf{G}}^p_n(c) = \overline{n}^{-1}\sum_{i=1}^{\overline{n}} \overline{\varepsilon}^p_i 1_{(\varepsilon_i\leq c)}$ and $\overline{\mathsf{G}}^{-1}_n(\psi) = \inf\{c : \overline{\mathsf{G}}(c) \geq \psi\}$, so that $\overline{\varepsilon}_{(k)} = \overline{\mathsf{G}}^{-1}_n(k/\overline{n})$. Then,

$$\overline{n}^{-1}\sum_{j=1}^{h_n} \overline{\varepsilon}^p_{(s-h_n+j)} = \overline{\mathsf{G}}^p_n\{\overline{\mathsf{G}}^{-1}_n(s/\overline{n})\} - \overline{\mathsf{G}}^p_n[\overline{\mathsf{G}}^{-1}_n\{(s-h_n)/\overline{n}\}].$$

Apply Lemma A.6 with $\mathsf{F} = \overline{\mathsf{G}}$, $n = \overline{n}$, requiring the 4+ moments and regularity of Assumption 6.1(i), noting that $\overline{\mathsf{G}}^{-1}_n\{\overline{n}/(\overline{n}+1)\} = \inf\{c : \overline{\mathsf{G}}_n(c) \geq \overline{n}/(\overline{n}+1)\} = \overline{\varepsilon}^p_{(\overline{n})}$, so that, uniformly in $h_n \leq s \leq \overline{n}$,

$$\overline{n}^{-1}\sum_{j=1}^{h_n} \overline{\varepsilon}^p_{(s-h_n+j)} = \mathsf{E}\overline{\varepsilon}^p_1 1_{\{s/\overline{n}-h_n/\overline{n}<\overline{\mathsf{G}}(\overline{\varepsilon}_1)\leq s/\overline{n}\}} + \mathsf{o}_\mathsf{P}(1).$$

We have $h_n/\overline{n} \to \overline{\omega}^{-1}$, where $\overline{\omega} \geq 1$ by construction, so that, by continuity of the distribution of $\overline{\varepsilon}_1$, we get $\mathsf{E}\overline{\varepsilon}^p_1 1_{\{s/\overline{n}-h_n/\overline{n}<\overline{\mathsf{G}}(\overline{\varepsilon}_1)\leq s/\overline{n}\}} = \mathsf{E}\overline{\varepsilon}^p_1 1_{\mathcal{A}_{s/\overline{n}}} + \mathsf{o}_\mathsf{P}(1)$ uniformly in $h_n \leq s \leq \overline{n}$ where $\mathcal{A}_{s/\overline{n}} = \{s/\overline{n} - \overline{\omega}^{-1} < \overline{\mathsf{G}}(\overline{\varepsilon}_1) \leq s/\overline{n}\}$. In particular, $h_n/\overline{n} = \overline{n}^{-1}\sum_{j=1}^{h_n} \overline{\varepsilon}^0_{(s+j)} = \mathsf{E}1_{\mathcal{A}_{s/\overline{n}}} + \mathsf{o}_\mathsf{P}(1)$, where $\mathsf{E}1_{\mathcal{A}_{s/\overline{n}}} = \overline{\omega}^{-1}$. We get

$$h_n^{-1}\sum_{j=1}^{h_n} \overline{\varepsilon}^p_{(s+j)} = \frac{\mathsf{E}\overline{\varepsilon}^p_1 1_{\mathcal{A}_{s/\overline{n}}}}{\mathsf{E}1_{\mathcal{A}_{s/\overline{n}}}} + \mathsf{o}_\mathsf{P}(1) = \mathsf{E}(\overline{\varepsilon}^p_1|\mathcal{A}_{s/\overline{n}}) + \mathsf{o}_\mathsf{P}(1),$$

so that $\hat{\sigma}^2_{\delta_n+s}/\sigma^2 = \mathsf{E}(\overline{\varepsilon}^2_1|\mathcal{A}_{s/\overline{n}}) + \mathsf{o}_\mathsf{P}(1) - \{\mathsf{E}(\overline{\varepsilon}_1|\mathcal{A}_{s/\overline{n}}) + \mathsf{o}_\mathsf{P}(1)\}^2$. Since $\mathsf{E}\overline{\varepsilon}_1 1_{\mathcal{A}_{s/\overline{n}}} \leq \mathsf{E}\overline{\varepsilon}_1 < \infty$ and $\mathsf{E}1_{\mathcal{A}_{s/\overline{n}}} = \overline{\omega}^{-1} > 0$ uniformly in s, we get $\mathsf{E}(\overline{\varepsilon}_1|\mathcal{A}_s) \leq \overline{\omega}\mathsf{E}\overline{\varepsilon}_1 < \infty$. Thus,

$$\hat{\sigma}^2_{\delta_n+s}/\sigma^2 = \mathsf{E}(\overline{\varepsilon}^2_1|\mathcal{A}_{s/\overline{n}}) - \{\mathsf{E}(\overline{\varepsilon}_1|\mathcal{A}_{s/\overline{n}})\}^2 + \mathsf{o}_\mathsf{P}(1) = \mathsf{Var}(\overline{\varepsilon}_1|\mathcal{A}_{s/\overline{n}}) + \mathsf{o}_\mathsf{P}(1).$$

The sets $\mathcal{A}_{s/\overline{n}}$ are special cases of the sets $\mathcal{A}_\varsigma = \{\varsigma - \overline{\omega}^{-1} \leq \overline{\mathsf{G}}(\overline{\varepsilon}_1) \leq \varsigma\}$. Thus,

$$\min_{h_n\leq s\leq\overline{n}} \hat{\sigma}^2_{\delta_n+s}/\sigma^2 = \min_{h_n\leq s\leq\overline{n}} \mathsf{Var}(\overline{\varepsilon}_1|\mathcal{A}_{s/\overline{n}}) + \mathsf{o}_\mathsf{P}(1) \geq \min_{\overline{\omega}^{-1}\leq\varsigma\leq 1} \mathsf{Var}(\overline{\varepsilon}_1|\mathcal{A}_\varsigma) + \mathsf{o}_\mathsf{P}(1).$$

We have that $\overline{v} = \min_{\overline{\omega}^{-1}\leq\varsigma\leq 1} \mathsf{Var}(\overline{\varepsilon}_1|\mathcal{A}_\varsigma) > 1$ by Assumption 6.1(i).

2. *Consider $h_n - \overline{s}_n \leq s < h_n$ where $\overline{s}_n = (2\log h_n)^{-1/4}h_n$*, see (A.5). In this case, we have $h_n - \overline{s}_n$ 'outliers' and $\overline{s}_n$ 'good' observations. Expand,

$$\hat{\sigma}^2_{\delta_n+s}/\sigma^2 = A_n = A_{n1} + A_{n2} + A_{n3} + 2A_{n4}, \tag{A.11}$$

see §B.3 in Supplementary material, where

$$A_{n1} = h_n^{-1}\sum_{i=s+1}^{h_n} \{\varepsilon_{(\delta_n+i)} - \varepsilon_{(\delta_n+h_n)}\}^2 - [h_n^{-1}\sum_{i=s+1}^{h_n} \{\varepsilon_{(\delta_n+i)} - \varepsilon_{(\delta_n+h_n)}\}]^2,$$

$$A_{n2} = (\frac{s}{h_n})^2[\frac{1}{s}\sum_{j=1}^{s} \overline{\varepsilon}^2_{(j)} - \{\frac{1}{s}\sum_{j=1}^{s} \overline{\varepsilon}_{(j)}\}^2], \qquad A_{n3} = \frac{s}{h_n}(1-\frac{s}{h_n})\frac{1}{s}\sum_{j=1}^{s} \overline{\varepsilon}^2_{(j)}$$

$$A_{n4} = [h_n^{-1}\sum_{i=s+1}^{h_n} \{\varepsilon_{(\delta_n+h_n)} - \varepsilon_{(\delta_n+i)}\}]\{h_n^{-1}\sum_{j=1}^{s} \overline{\varepsilon}_{(j)}\}.$$

We note that $A_{n1}, A_{n2}, A_{n3}, A_{n4} \geq 0$. Therefore, $\hat{\sigma}^2_{\delta_n+s}/\sigma^2 \geq A_{n2}$. We argue, as in part (a), that $A_{n2} \geq \overline{v} + o_{\mathsf{P}}(1)$, where $\overline{v} > 1$ by Assumption 6.1(i). Indeed, we have that $1 > s/h_n \geq 1 - \overline{s}_n/h_n \to 1$. Thus, $\overline{n}^{-1}\sum_{j=1}^{h_n-\overline{s}_n} \overline{\varepsilon}^p_{(j)} \leq \overline{n}^{-1}\sum_{j=1}^{s} \overline{\varepsilon}^p_{(j)} \leq \overline{n}^{-1}\sum_{j=1}^{h_n} \overline{\varepsilon}^p_{(j)}$. Each of the bounds equal $\mathsf{E}\overline{\varepsilon}^p_1 1_{\mathcal{A}_{\overline{\omega}^{-1}}} + o_{\mathsf{P}}(1)$. Hence, we get $\overline{n}^{-1}\sum_{j=1}^{s} \overline{\varepsilon}^p_{(j)} = \mathsf{E}\overline{\varepsilon}^p_1 1_{\mathcal{A}_{\overline{\omega}^{-1}}} + o_{\mathsf{P}}(1)$, uniformly in s, where $\mathsf{P}(\mathcal{A}_{\overline{\omega}^{-1}}) = \overline{\omega}^{-1}$. In turn, $\min_{h_n-\overline{s}_n \leq s < h_n} A_{n2} \geq \mathsf{Var}(\overline{\varepsilon}_1 | \mathcal{A}_{\overline{\omega}^{-1}}) + o_{\mathsf{P}}(1) \geq \overline{v} + o_{\mathsf{P}}(1)$ ■

Proof of Theorem 6.4. We will show that $\hat{s} = \hat{\delta}_{LTS} - \delta_n = o_{\mathsf{P}}(h_n^\alpha)$ for any $\alpha > 0$. It suffices to consider the case where $\mathsf{P}(\hat{\delta}_{LTS} - \delta_n > h_n^\alpha) \to 0$ and where $\rho < 1$ as remarked in §6.1. We consider $\gamma, \rho < 1$ so that $\overline{\omega} = (1-\rho)(1-\gamma)/\gamma$ satisfies $0 < \overline{\omega}$. The case $\overline{\omega} < 1$ was covered in the proof of Theorem 6.3. Thus, suppose $\overline{\omega} \geq 1$.

Recall $\underline{s}_n, \overline{s}_n$ from (A.5). In particular, $h_n^\alpha > \underline{s}_n$ for large n, so that $\mathsf{P}(\hat{s} > h_n^\alpha) \leq \mathsf{P}(\hat{s} \geq \underline{s}_n)$. We show, that the latter probability vanishes. Note that $\hat{s} \leq \overline{n}$.

We have that $\hat{s}$ is the minimizer to $\hat{\sigma}^2_{\delta_n+s} - \hat{\sigma}^2_{\delta_n}$, which is zero for $s = \delta - \delta_n = 0$. The Lemmas A.9, A.10 using Assumption 6.1(i) show that $\hat{\sigma}^2_{\delta_n+s} - \hat{\sigma}^2_{\delta_n}$, asymptotically, has a uniform, positive lower bound on each of the intervals $\underline{s}_n \leq \hat{s} \leq h_n - \overline{s}_n$ and $h_n - \overline{s}_n \leq \hat{s} \leq \overline{n}$. Thus, $\hat{\sigma}^2_{\delta_n+s} - \hat{\sigma}^2_{\delta_n}$ is bounded away from zero on $s \geq \underline{s}_n$ so that $\mathsf{P}(\hat{s} \geq \underline{s}_n) \to 0$.

A similar argument applies for $\hat{\delta}_{LTS} - \delta_n < -h_n^\alpha$ using Assumption 6.1(ii).

The Lemmas A.7, A.8 give the limiting results for $\hat{\mu}, \hat{\sigma}^2$. ■

A.3 The LMS estimator in the LMS model

In the LMS Model 2 the 'good' observations are uniform. Uniform spacings can be written as ratios of sums of exponential variables by Lemma A.2. We start with some properties of sums of exponential variables.

A.3.1 Some results for sums of exponential variables

Lemma A.11 *Let $e_1, e_2, \ldots$ be independent standard exponentially distributed. Define $g_{jn} = \sum_{i=1}^n e_{j+i}$ for $n, j+1 \in \mathbb{N}$. Then*
(a) g_{jn} is $\Gamma(n,1)$ distributed and $\mathsf{E}|g_{jn} - n|^4 = 3n(n+2) \leq 9n^2$;
(b) $\mathsf{P}(|n^{-1}(g_{jn} - n)| \geq x) \leq 9x^{-4}n^{-2}$;
(c) $\mathsf{P}(\max_{0<j<n_1} |n_0^{-1}(g_{jn_0} - n_0)| > x) \leq 9x^{-4}n_1 n_0^{-2}$.

Proof. (a) see §17.6 and equation 17.10 of Johnson, Kotz and Balakrishnan (1994).

(b) By the Markov inequality, $\mathsf{P}(|n^{-1}(g_{jn} - n)| \geq x) \leq (nx)^{-4}\mathsf{E}|g_{jn} - n|^4$. Apply (a).

(c) Let $z_j = n_0^{-1}(g_{jn_0} - n_0)$ and $\mathcal{P}_n = \mathsf{P}(\max_{0<j<n_1} |z_j| > x)$. Boole's inequality gives $\mathcal{P}_n \leq \sum_{0<j<n_1} \mathsf{P}(|z_j| > x)$. Here, $\mathsf{P}(|z_j| > x) \leq 9x^{-4}n_0^{-2}$ by (b), so that $\mathcal{P}_n \leq 9x^{-4}n_1 n_0^{-2}$. ■

A.3.2 Proofs for LMS

We consider the LMS location-scale Model 2 and the sequence of data generating processes outlined in §6.1. As for the LTS estimator, the main difficulty is to show that the minimizer $\hat{\delta}_{LMS}$ of $\hat{\sigma}_\delta$ is close to $\delta_n = \sum_{j \in \zeta_n} 1_{(\varepsilon_j < \min_{i \in \zeta_n} \varepsilon_i)}$. Due to the argument (A.4), it suffices to analyze the asymptotic behaviour for deterministic sequences δ_n.

We start by noting that for known δ_n then $\hat{\mu}_{\delta_n}$ and $\hat{\sigma}_{\delta_n}$ are the maximum likelihood estimators for a uniform location-scale model. These estimators have been studied intensely.

The following lemma gives an overview of the asymptotic theory. Johnson, Kotz and Balakrishnan (1994b) give an overview of the finite sample theory, which we will not need here.

Lemma A.12 *Consider the LMS Model 2 and the sequence of data generating processes in §6.1 where $\rho < 1$. Let $\underline{e}, \overline{e}$ be independent standard exponential. Then, conditional on δ_n,*

$$h_n(\hat{\mu}_{\delta_n} - \mu)/\sigma \xrightarrow{\mathsf{D}} \underline{e} - \overline{e}, \qquad h_n(\hat{\sigma}_{\delta_n} - \sigma)/\sigma \xrightarrow{\mathsf{D}} -(\underline{e} + \overline{e}),$$

which are dependent $\mathsf{Laplace}(0, \gamma)$ *and* $-\Gamma(2, \gamma)$ *distributions.*

Proof. Since ε_i is uniform on $[-1, 1]$ then $u_i = (\varepsilon_i + 1)/2$ is uniform on $[0, 1]$. The spacings Lemma A.2 gives that independent standard exponential variables e_i exist, so that $u_{(\delta_n+1)} = e_1/\sum_{k=1}^{h_n+1} e_k$ and $1 - u_{(\delta_n+h_n)} = e_{h_n+1}/\sum_{k=1}^{h_n+1} e_k$. In particular,

$$(\hat{\sigma}_{\delta_n} - \sigma)/\sigma = \{\varepsilon_{(\delta_n+h_n)} - \varepsilon_{(\delta_n+1)} - 2\}/2 = u_{(\delta_n+h_n)} - u_{(\delta_n+1)} - 1 = -(e_1 + e_{h_n+1})/\textstyle\sum_{k=1}^{h_n+1} e_k,$$
$$(\hat{\mu}_{\delta_n} - \mu)/\sigma = \{\varepsilon_{(\delta_n+h_n)} + \varepsilon_{(\delta_n+1)}\}/2 = u_{(\delta_n+h_n)} + u_{(\delta_n+1)} - 1 = (e_1 - e_{h_n+1})/\textstyle\sum_{k=1}^{h_n+1} e_k.$$

By the Law of Large Numbers, $(h_n + 1)^{-1} \sum_{i=1}^{h_n+1} e_i \to 1$ in probability. ■

Lemma A.13 *Consider the LMS Model 2 and the sequence of data generating processes in §6.1 where $\gamma, \rho < 1$. Suppose Assumption 6.2 holds. Then, conditional on δ_n, an $\epsilon > 0$ exists, so that* $\min_{1 \le s < h_n} h_n(\hat{\sigma}_{\delta_n+s} - \hat{\sigma}_{\delta_n}) \ge \epsilon + \mathsf{o_P}(1)$.

Proof. Let $S_s = (\hat{\sigma}_{\delta_n+s} - \hat{\sigma}_{\delta_n})/\sigma = \{\varepsilon_{(\delta_n+s+h_n)} - \varepsilon_{(\delta_n+s+1)}\}/2 - \{\varepsilon_{(\delta_n+h_n)} - \varepsilon_{(\delta_n+1)}\}/2$. Reorganize as $S_s = \{\varepsilon_{(\delta_n+h_n+s)} - \varepsilon_{(\delta_n+h_n)}\}/2 - \{\varepsilon_{(\delta_n+s+1)} - \varepsilon_{(\delta_n+1)}\}/2$.

The 'good' errors $\varepsilon_{(\delta_n+s+1)}$ and $\varepsilon_{(\delta_n+1)}$ are order statistics of uniform errors on $[-1, 1]$. Thus, $\{\varepsilon_{(\delta_n+1+s)} - \varepsilon_{(\delta_n+1)}\}/2 = u_{(1+s)} - u_{(1)}$ is a standard uniform spacing. The uniform spacings Lemma A.2 shows that there exists independent standard exponential variables e_k where $1 \le k \le h_n + 1$ so that $u_{(1+s)} - u_{(1)} = \sum_{k=2}^{1+s} e_k/\sum_{k=1}^{h_n+1} e_k$.

The 'outliers' satisfy $\varepsilon_{(\delta_n+h_n+s)} - \varepsilon_{(\delta_n+h_n)} = \overline{\varepsilon}_{(s)}$ where $\overline{\varepsilon}_{(s)}$ is positive and an order statistic of the distribution function $\overline{\mathsf{G}}$. By the inverse probability transformation, there exist independent standard uniform variables $\overline{u}_s$, so that $\overline{\varepsilon}_{(s)} = \overline{\mathsf{G}}^{-1}\{\overline{u}_{(s)}\}$. Thus, $\varepsilon_{(\delta_n+h_n+s)} - \varepsilon_{(\delta_n+h_n)} = \overline{\mathsf{G}}^{-1}\{\overline{u}_{(s)}\}$.

We consider the cases $1 \le s < s_n$ and $s_n \le s < h_n$ separately for some sequence $s_n \to \infty$, but $s_n/n \to 0$. We choose $s_n = n^{(1-\tau)/2}$ for $\tau < 1$ defined in Assumption 6.2(ii).

The case $s_n \le s \le h_n$. For the *'good' observations* bound

$$\{\varepsilon_{(\delta_n+1+s)} - \varepsilon_{(\delta_n+1)}\}/2 = \frac{\sum_{k=2}^{1+s} e_k}{\sum_{k=1}^{h_n+1} e_k} \le \frac{\sum_{k=2}^{s+1} e_k}{\sum_{k=2}^{h_n+1} e_k} = (\frac{s}{h_n})\frac{1 + s^{-1}\sum_{k=2}^{s+1}(e_k - 1)}{1 + h_n^{-1}\sum_{k=2}^{h_n+1}(e_k - 1)}. \qquad \text{(A.12)}$$

Let $m_n^{\text{large}} = \max_{t \ge s_n} |t^{-1} \sum_{i=2}^{t+1}(e_i - 1)|$. By the strong Law of Large Numbers $x_t = t^{-1} \sum_{i=1}^{t}(e_i - 1) \to 0$ *a.s.* for $t \to \infty$. This implies $m_n^{\text{large}} \to 0$ *a.s.*, since for each outcome we have deterministic sequence $x_t \to 0$, say. But, if $x_t \to 0$, then $\limsup_{t\to\infty} |x_t| \to 0$. In particular, for $s_n \to \infty$ we get $\max_{t \ge s_n} |x_t| \to 0$. In summary, we get

$$\{\varepsilon_{(\delta_n+1+s)} - \varepsilon_{(\delta_n+1)}\}/2 \overset{a.s.}{\le} (s/h_n)\{1 + \mathrm{o}(1)\}. \qquad \text{(A.13)}$$

The 'outliers'. By Assumption 6.2(i), we have $\overline{\mathsf{G}}^{-1}(\psi) \geq 2\psi\varrho$, where $\varrho = (1-\rho+\epsilon)(1-\gamma)/\gamma$. Thus, we get $\{\varepsilon_{(\delta_n+h_n+s)} - \varepsilon_{(\delta_n+h_n)}\}/2 = \overline{\mathsf{G}}^{-1}\{\overline{u}_{(s)}\}/2 \geq \varrho\overline{u}_{(s)}$. The uniform spacings Lemma A.2 shows that there exists independent standard exponential variables $\overline{e}_k$ for $1 \leq k \leq \overline{n}+1$ so that $\overline{u}_{(s)} = \sum_{k=1}^{s}\overline{e}_k / \sum_{k=1}^{\overline{n}+1}\overline{e}_k$. Thus, we can bound

$$\{\varepsilon_{(\delta_n+h_n+s)} - \varepsilon_{(\delta_n+h_n)}\}/2 \geq \varrho\frac{\sum_{k=1}^{s}\overline{e}_k}{\sum_{k=1}^{\overline{n}+1}\overline{e}_k} = \varrho(\frac{s}{\overline{n}+1})\frac{1+s^{-1}\sum_{k=1}^{s}(\overline{e}_k-1)}{1+(\overline{n}+1)^{-1}\sum_{k=1}^{\overline{n}+1}(\overline{e}_k-1)}. \tag{A.14}$$

Let $\overline{m}_n^{\text{large}} = \max_{s\geq s_n}|s^{-1}\sum_{j=1}^{s}(\overline{e}_j-1)|$. Using the strong Law of Large Numbers as before we see that $\overline{m}_n^{\text{large}} = \mathrm{o}(1)$ $a.s.$ Thus, we get

$$\{\varepsilon_{(\delta_n+h_n+s)} - \varepsilon_{(\delta_n+h_n)}\}/2 \overset{a.s}{\geq} \varrho(\frac{s}{\overline{n}+1})\{1+\mathrm{o}(1)\}.$$

Combine with the bound (A.13) to see that $\min_{s_n\leq s\leq h_n}(\overline{n}+1)S_s \geq s\{\varrho-(\overline{n}+1)/h_n\}\{1+\mathrm{o}(1)\}$ $a.s.$ Since $(\overline{n}+1)/h_n \to \tilde{\rho} = (1-\rho)(1-\gamma)/\gamma$, so that $\varrho - \tilde{\rho} = \epsilon(1-\gamma)\gamma^{-1} > 0$, while $s > s_n$ we get $\min_{s_n\leq s\leq h_n}(\overline{n}+1)S_s \geq s_n\epsilon(1-\gamma)\gamma^{-1}\{1+\mathrm{o}(1)\}$ $a.s.$ which goes to infinity with s_n, while $\overline{n}/h_n \to \tilde{\rho} > 0$.

The case $1 \leq s < s_n = h_n^{(1-\tau)/2}$. For the 'good' observations bound

$$\{\varepsilon_{(\delta_n+1+s)} - \varepsilon_{(\delta_n+1)}\}/2 = \frac{\sum_{k=2}^{1+s}e_k}{\sum_{k=1}^{h_n+1}e_k} \leq \frac{\sum_{k=2}^{1+s_n}e_k}{\sum_{k=2}^{h_n+1}e_k} = (\frac{s_n}{h_n})\frac{1+s_n^{-1}\sum_{k=2}^{s_n+1}(e_k-1)}{1+h_n^{-1}\sum_{k=2}^{h_n+1}(e_k-1)}.$$

By the strong Law of Large Numbers we get that the averages in the numerator and denominator vanish, so that $\{\varepsilon_{(\delta_n+1+s)} - \varepsilon_{(\delta_n+1)}\}/2 \leq (s_n/h_n)\{1+\mathrm{o}(1)\}$ $a.s.$

For the 'outliers', Assumption 6.2(ii) is $\overline{\mathsf{G}}^{-1}(\psi) \geq \psi^{\tau}$ for some $\tau < 1$ and $\psi < \psi_0$. Thus, we get $\varepsilon_{(\delta_n+h_n+s)} - \varepsilon_{(\delta_n+h_n)} = \overline{\mathsf{G}}^{-1}\{\overline{u}_{(s)}\} \geq \{\overline{u}_{(s)}\}^{\tau}$. Further, $\overline{u}_{(s)} \geq \overline{u}_{(1)}$. As before, $\overline{u}_{(1)} = \overline{e}_1/\sum_{k=1}^{\overline{n}+1}\overline{e}_k$. Thus, we can replace (A.14) with

$$\varepsilon_{(\delta_n+h_n+s)} - \varepsilon_{(\delta_n+h_n)} = \overline{\mathsf{G}}^{-1}\{\overline{u}_{(s)}\} \geq \{\overline{u}_{(1)}\}^{\tau} = (\frac{\overline{e}_1}{\sum_{k=1}^{\overline{n}+1}\overline{e}_k})^{\tau} \overset{a.s.}{=} (\frac{\overline{e}_1}{\overline{n}+1})^{\tau}\{1+\mathrm{o}(1)\},$$

by the Strong Law of Large Numbers. Since $\overline{e}_1$ is exponential, then for all $\epsilon > 0$ exists a $\eta > 0$ so that $\mathsf{P}(\overline{e}_1 > 2\eta) \geq 1-\epsilon$. As before, $\overline{n}/h_n \to \tilde{\rho} > 0$. In combination, we get

$$\min_{1\leq s\leq s_n} h_nS_s \geq [(h_n/2)\{\eta/(\tilde{\rho}h_n)\}^{\tau} - h_n(s_n/h_n)]\{1+\mathrm{o}_{\mathsf{P}}(1)\}.$$

Recalling that $s_n = h_n^{(1-\tau)/2}$, this gives,

Thus, for some constant $C > 0$, we get $\min_{1\leq s\leq s_n} h_nS_s \geq (Ch_n^{1-\tau} - s_n)\{1+\mathrm{o}_{\mathsf{P}}(1)\}$. This diverges since $s_n = h_n^{(1-\tau)/2} = \mathrm{o}(h_n^{1-\tau})$ when $1-\tau > 0$. ■

Proof of Theorem 6.5. We proceed as in the proof of Theorem 6.3, conditioning on sequences δ_n satisfying $\delta_n/(n-h_n) \to \rho$, and only considering $\hat{s} = \hat{\delta}_{LMS} - \delta_n > 0$. Thus, it suffices to show that $n(\hat{\sigma}^2_{\delta_n+s} - \hat{\sigma}^2_{\delta_n}) > \epsilon + \mathrm{o}_{\mathsf{P}}(1)$ for some $\epsilon > 0$, uniformly in $1 \leq s < h_n$. This was proved in Lemma A.13 using Assumption 6.2.

Since $\mathsf{P}(\hat{\delta} = \delta_n) \to 1$, we get $\hat{\mu} = \hat{\mu}_{\hat{\delta}} = \hat{\mu}_{\delta_n}, \hat{\sigma} = \hat{\sigma}_{\hat{\delta}} = \hat{\sigma}_{\delta_n}$ with large probability. The limit distributions follow from Lemma A.12. ■

Lemma A.14 *Consider the LMS Model 2 and the sequence of data generating processes in §6.1 where* $1 \leq \overline{\omega} = (1-\rho)(1-\gamma)/\gamma < \infty$. *Suppose Assumption 6.3(i) holds. Let* $s_n \to \infty$, *but* $s_n/h_n \to 0$. *Then, conditional on* δ_n, *an* $\epsilon > 0$ *exists, so that* $\min_{h_n - s_n \leq s \leq \overline{n}}(\hat{\sigma}_{\delta_n+s} - \hat{\sigma}_{\delta_n}) \geq \epsilon + o_P(1)$.

Proof. Let $S_s = (\hat{\sigma}_{\delta_n+s} - \hat{\sigma}_{\delta_n})/\sigma = \{\varepsilon_{(\delta_n+s+h_n)} - \varepsilon_{(\delta_n+s+1)}\}/2 - \{\varepsilon_{(\delta_n+h_n)} - \varepsilon_{(\delta_n+1)}\}/2$.

The 'outliers'. For small s, then $S_{s1} = \varepsilon_{(\delta_n+s+h_n)} - \varepsilon_{(\delta_n+s+1)}$ includes both 'outliers' and 'good' errors. Since $\varepsilon_{(\delta_n+s+s_n)} \geq \varepsilon_{(\delta_n+h_n)}$, then $S_{s1} \geq \varepsilon_{(\delta_n+s+h_n)} - \varepsilon_{(\delta_n+s+s_n)}$, which only includes 'outliers'. Let $t = s + s_n - h_n$ satisfying $0 \leq t \leq \overline{n} + s_n - h_n$, so that $\varepsilon_{(\delta_n+s+h_n)} - \varepsilon_{(\delta_n+s+s_n)} = \varepsilon_{(\delta_n+h_n+t+h_n-s_n)} - \varepsilon_{(\delta_n+h_n+t)}$.

As in the proof of Lemma A.13, we have that $\varepsilon_{(\delta_n+h_n+s)} - \varepsilon_{(\delta_n+h_n)} = \overline{\varepsilon}_{(s)} = \overline{\mathsf{G}}^{-1}\{\overline{u}_{(s)}\}$. Thus, $\varepsilon_{(\delta_n+h_n+t+h_n-s_n)} - \varepsilon_{(\delta_n+h_n+s)} = \overline{\mathsf{G}}^{-1}\{\overline{u}_{(t+h_n-s_n)}\} - \overline{\mathsf{G}}^{-1}\{\overline{u}_{(t)}\}$. In combination, $S_{s1} \geq \overline{\mathsf{G}}^{-1}\{\overline{u}_{(t+h_n-s_n)}\} - \overline{\mathsf{G}}^{-1}\{\overline{u}_{(t)}\}$.

Assumption 6.3 has $\overline{\mathsf{G}}^{-1}(\xi+\psi) - \overline{\mathsf{G}}^{-1}(\xi) \geq 2\psi\varrho$ where $\varrho = (1-\rho+\epsilon)(1-\gamma)/\gamma$. Thus, $S_{s1} \geq 2\varrho\{\overline{u}_{(t+h_n-s_n)} - \overline{u}_{(t)}\}$. The uniform spacings Lemma A.2 shows that there exists independent standard exponential variables $\overline{e}_k$ for $1 \leq k \leq \overline{n}+1$ so that $\overline{u}_{(t)} = \sum_{k=1}^{t} \overline{e}_k / \sum_{k=1}^{\overline{n}+1} \overline{e}_k$. Thus, we get

$$S_{s1} \geq 2\varrho \frac{\sum_{k=t+1}^{t+h_n-s_n} \overline{e}_k}{\sum_{k=1}^{\overline{n}+1} \overline{e}_k} = 2\varrho\left(\frac{h_n - s_n}{\overline{n}+1}\right) \frac{1 + (h_n - s_n)^{-1} \sum_{k=t+1}^{t+h_n-s_n} (\overline{e}_k - 1)}{1 + (\overline{n}+1) \sum_{k=1}^{\overline{n}+1} (\overline{e}_k - 1)}.$$

In the denominator the Law of Large Numbers shows $(\overline{n}+1)\sum_{k=1}^{\overline{n}+1}(\overline{e}_k - 1) = o_P(1)$. The sum in the numerator depends on t. Thus, consider $\overline{m}_n = \max_{0 \leq s \leq \overline{n}+s_n-h_n} |(h_n - s_n)^{-1} \sum_{k=s+1}^{s+h_n-s_n} (e_k - 1)|$. Lemma A.11($c$) with $n_0 = h_n - 1$, $n_1 = \overline{n} + s_n - h_n + 1$, $x = n^{-1/5}$ shows that $\overline{m}_n = o_P(1)$. Finally, $(h_n - s_n)/(\overline{n}+1) \to \tilde{\gamma} = \gamma/\{(1-\gamma)(1-\rho)\}$ so that

$$S_{s1} \geq 2\varrho\tilde{\gamma}\{1 + o_P(1)\} = 2\{1 + \epsilon/(1-\rho)\}\{1 + o_P(1)\}. \tag{A.15}$$

The 'good' observations. Apply (A.13) with $s+1 = h_n$ to see that $\{\varepsilon_{(\delta_n+h_n)} - \varepsilon_{(\delta_n+1)}\}/2 \leq 1 + o_P(1)$. Combine with (A.15) to get $\min_{h_n \leq s \leq \overline{n}} S_s \geq \epsilon/(1-\rho) + o_P(1)$. ■

Proof of Theorem 6.6. Extend the proof of Theorem 6.5 in the same way as the proof of Theorem 6.4 extends that of Theorem 6.3 while replacing Lemmas A.9, A.10 with Lemmas A.13, A.14. ■

B Supplementary material

In the location-scale model, the LTS estimation in (2.2), that is

$$\hat{\zeta}_{LTS} = \arg\min_{\zeta} \sum_{i\in\zeta} (y_i - \hat{\beta}_\zeta' x_i)^2 \quad \text{where} \quad \hat{\beta}_\zeta = (\sum_{i\in\zeta} x_i x_i')^{-1} \sum_{i\in\zeta} x_i y_i,$$

reduces so that $\hat{\zeta}_{LTS}$ are the indices corresponding to the order statistics $y_{(\hat{\delta}_{LTS}+1)}, \ldots, y_{(\hat{\delta}_{LTS}+h)}$ where

$$\hat{\delta}_{LTS} = \arg\min_{\delta} \sum_{i=1}^{h} \{y_{(\delta+i)} - \hat{\mu}_\delta\}^2 \quad \text{where} \quad \hat{\mu}_\delta = h^{-1} \sum_{i=1}^{h} y_{(\delta+i)}.$$

Consider data $y_1, \ldots, y_n$. We want to argue that the residual sum of squares, RSS, for a h sub-sample indexed by ζ is bounded below by the RSS of h consecutive order statistics.

Suppose $n = h+1$. The smallest RSS for any h sub-sample is either that of $y_{(1)}, \ldots, y_{(h)}$ or $y_{(2)}, \ldots, y_{(h+1)}$. To see this, consider a h-sub-sample leaving out one of $y_{(1)}, \ldots, y_{(n)}$. Call that observation y. Defining $\mu_k = h^{-1} \sum_{i=1}^{h+1} y_i^k$, the RSS is

$$RSS_y = \frac{1}{h}(\sum_{i=1}^{h+1} y_i^2 - y^2) - \{\frac{1}{h}(\sum_{i=1}^{h+1} y_i - y)\}^2 = (\mu_2 - \frac{y^2}{h}) - (\mu_1 - \frac{y}{h})^2,$$

which we can expand as

$$RSS_y = -y^2 \frac{h+1}{h^2} + 2y\frac{\mu_1}{h} + \mu_2 - \mu_1^2.$$

This is a concave, quadratic function with maximum at

$$\frac{-2\mu_1/h}{-2(h+1)/h^2} = \frac{h}{h+1}\mu_1 = \frac{1}{h+1}\sum_{i=1}^{h+1} y_i = \bar{\mu}_1.$$

The extreme order statistics, $y_{(1)}$ or $y_{(n)}$, are the y values furthest from the sample average $\bar{\mu}_1$. Thus, RSS_y must attain its minimum either at $y_{(1)}$ or $y_{(n)}$. Thus, the minimum RSS of a h-sub-sample is achieved either from computing the RSS of $y_{(2)}, \ldots, y_{(h+1)}$ or $y_{(1)}, \ldots, y_{(h)}$. In particular, if $\bar{\mu}_1 \le y < y_{(h+1)}$, then RSS_y is larger than the RSS of $y_{(1)}, \ldots, y_{(h)}$. Further, if $\bar{\mu}_1 \ge y > y_{(1)}$, then RSS_y is larger than the RSS of $y_{(2)}, \ldots, y_{(h+1)}$. The condition $\bar{\mu}_1 \le y$ is equivalent to $\sum_{i=1}^{h+1} y_i \le (h+1)y$. Subtracting y from both sides, shows this is equivalent to $y \ge h^{-1}(\sum_{i=1}^{h+1} y_i - y)$, which is the average of the h sub-sample without y.

For a general n, we select an h-sub-sample ζ and compute the RSS given by $RSS_\zeta = h^{-1}\sum_{i\in\zeta} y_i^2 - (h^{-1}\sum_{i\in\zeta} y_i)^2$. Let $\delta+1$, $\delta+r$ be the ranks, in the full sample, of $\min_{i\in\zeta} y_i$ and $\max_{i\in\zeta} y_i$, so that $y_{(\delta+1)} = \min_{i\in\zeta} y_i$ and $y_{(\delta+r)} = \max_{i\in\zeta} y_i$. Since there are h indices in ζ, then $r \ge h$. If $r > h$ then y_i for $i \in \zeta$ consists of points that are not consecutive order statistics. Then there exists an order statistics $y_{(s)}$, say, that is not included among y_i for $i \in \zeta$ and so that $y_{(\delta+1)} < y_{(s)} < y_{(\delta+r)}$. Thus, we can form a new index set ζ' from ζ by replacing $y_{(\delta+r)}$ by $y_{(s)}$ if $y_{(s)} \ge h^{-1}\sum_{i\in\zeta} y_i$ or by replacing $y_{(\delta+1)}$ by $y_{(s)}$ if $y_{(s)} \le h^{-1}\sum_{i\in\zeta} y_i$. The above derivations shows that $RSS_\zeta > RSS_{\zeta'}$. Further, there exists a $\delta' \ge \delta$ and an $r' < r$ so that $y_{(\delta'+1)} = \min_{i\in\zeta'} y_i$ and $y_{(\delta'+r')} = \max_{i\in\zeta'} y_i$. This procedure can be iterated until we achieve an h-sub-sample formed from consecutive order statistics $y_{(\delta+1)}, \ldots, y_{(\delta+h)}$.

B.1 Maximum likelihood with uniform errors

Consider the location model with uniform errors with known range. That is $y_1, \ldots, y_n$ satisfy $y_i = \mu + \varepsilon_i$, where ε_i are i.i.d. $\mathsf{Uniform}[-1, 1]$. Then the likelihood is

$$\mathsf{L}(\mu) = \prod_{i=1}^{n} \frac{1}{2} 1_{(-1 \le |y_i - \mu| \le 1)} = 2^{-n} 1_{(\max_{1 \le i \le n} |y_i - \mu| \le 1)}.$$

The likelihood is maximized and constant for those μ where $\max_{1 \le i \le n} |y_i - \mu| \le 1$. Thus, the likelihood is maximized for any μ so that $y_{(n)} - 1 \le \mu \le y_{(1)} + 1$. This interval includes the Chebychev estimator $\hat{\mu}_{Cheb} = \{y_{(1)} + y_{(n)}\}/2$ as an interior point. For instance, $\hat{\mu}_{Cheb} = y_{(1)} + \{y_{(n)} - y_{(1)}\}/2 < y_{(1)} + 1$ since $|y_{(n)} - y_{(1)}| < 2$.

Consider the location-scale model with uniform errors. That is $y_1, \ldots, y_n$ satisfy $y_i = \mu + \sigma\varepsilon_i$, where ε_i are i.i.d. $\mathsf{Uniform}[-1, 1]$. Then the likelihood is

$$\mathsf{L}(\mu, \sigma) = \prod_{i=1}^{n} \frac{1}{2\sigma} 1_{(-\sigma \le |y_i - \mu| \le \sigma)} = (2\sigma)^{-n} 1_{(\max_{1 \le i \le n} |y_i - \mu| \le \sigma)}.$$

Since σ^{-n} is decreasing in σ, this is maximized for fixed μ by $\hat{\sigma}_\mu = \max_{1 \le i \le n} |y_i - \mu|$, which is a unique maximizer. This results in the profile likelihood

$$\mathsf{L}(\mu, \sigma) \le \mathsf{L}_\sigma(\mu) = (2\hat{\sigma}_\mu)^{-n} = (2 \max_{1 \le i \le n} |y_i - \mu|)^{-n},$$

which is maximized by minimizing $\hat{\sigma}_\mu$. We find, for any μ, that

$$\hat{\sigma}_\mu = \max_{1 \le i \le n} |y_i - \mu| = \max\{|y_{(n)} - \mu|, |\mu - y_{(1)}|\}.$$

The maximum of two distances is minimized by choosing μ so that they are equal. This gives the Chebychev estimator $\hat{\mu}_{Cheb} = \{y_{(1)} + y_{(n)}\}/2$, which is the unique minimizer of $\hat{\sigma}_\mu$. In other words, we have maximized $\mathsf{L}(\mu, \sigma)$ by choosing the smallest σ subject to $\sigma \ge \max_{1 \le i \le n} |y_i - \mu|$, which is the characterization of the Chebychev estimator in (2.4).

B.2 Details of examples for general maximum likelihood

Example 3.1: $\mathsf{P}, \mathsf{Q} \in \mathcal{P}$ are dominated by a σ-finite measure μ, and suppose P and Q have density versions p and q with respect to μ which are continuous at x and at least $q(x) > 0$. Then $\lim_{\epsilon \to 0} \mathsf{P}(C_{x,\epsilon})/\mathsf{Q}(C_{x,\epsilon}) \to p(x)/q(x)$. Indeed, we can expand

$$|\mathsf{P}(C_{x,\epsilon}) - p(x)\mu(C_{x,\epsilon})| = |\int 1_{C_{x,\epsilon}}(u)\{p(u) - p(x)\}d\mu(u)| \le \mu(C_{x,\epsilon}) \sup_{u \in C_{x,\epsilon}} |p(u) - p(x)|.$$

The continuity of p implies that $\forall \delta > 0$, $\exists \epsilon > 0$ so that we get the further bound $\delta\mu(C_{x,\epsilon})$. The same can be done for Q. We therefore get that

$$\frac{\mathsf{P}(C_{x,\epsilon})}{\mathsf{Q}(C_{x,\epsilon})} = \frac{p(x)\mu(C_{x,\epsilon}) + \mathrm{O}(\delta)\mu(C_{x,\epsilon})}{q(x)\mu(C_{x,\epsilon}) + \mathrm{O}(\delta)\mu(C_{x,\epsilon})}.$$

For $\delta \to 0$ the remainder terms vanish.

Example 3.2: Consider $y_1, \ldots, y_n$ that are i.i.d. with unknown distribution function F. The parameter is F, which varies among arbitrary distribution functions on $\mathbb{R}$. Let $x_1 < \cdots < x_k$ be the distinct outcomes with counts $n_1, \ldots, n_k$ so that $\sum_{j=1}^k n_j = n$. The empirical distribution function F_n has support on $x_1 < \cdots < x_k$ with jumps of size n_j/n. The probability of the hypercubes reduces to

$$\mathsf{P}_n(C_{x,\epsilon}) = \textstyle\prod_{j=1}^k (n_j/n)^{n_j}, \tag{B.1}$$

for any $\epsilon < \min_{1<j\leq k}(x_j - x_{j-1})$. We start by arguing that the estimator $\hat{\mathsf{F}}$ of F has to have discrete support on $x_1, \ldots, x_k$ to be maximum likelihood estimator. Indeed, suppose $\hat{\mathsf{F}}$ is continuous in a neighbourhood of some x_j. Then $\Delta\hat{\mathsf{F}}(x_j) = \hat{\mathsf{F}}(x_j) - \hat{\mathsf{F}}(x_j - \epsilon)$ vanishes for small ϵ and, for sufficiently small ϵ, we have $\{\Delta\hat{\mathsf{F}}(x_j)\}^{n_j} < \mathsf{P}_n(C_{x,\epsilon})/2$. For other values of x, we have the bound $\Delta\hat{\mathsf{F}}(x) \leq 1$. In combination, $\mathsf{P}_{\hat{\mathsf{F}}}(C_{x,\epsilon}) \leq \mathsf{P}_n(C_{x,\epsilon})/2$ and we get $\hat{\mathsf{F}} <_x \mathsf{F}_n$, so that this $\hat{\mathsf{F}}$ cannot be a maximizer.

Now, consider an $\hat{\mathsf{F}}$ that has discrete support on $x_1, \ldots, x_k$. It may also have support elsewhere, but that will not contribute to the likelihood. For such $\hat{\mathsf{F}}$, we let $p_j = \Delta\hat{\mathsf{F}}(x_j)$ and find that $\mathsf{P}_{\hat{\mathsf{F}}}(C_{x,\epsilon}) = \prod_{j=1}^k (p_j)^{n_j}$ where $\sum_{j=1}^k p_j \leq 1$ for all small ϵ. We apply the information inequality $\prod_{j=1}^k (p_j)^{n_j} \leq \prod_{j=1}^k (n_j/n)^{n_j}$ with equality if only if $p_j = n_j/n$ for all j, which is proved by applying Jensen's inequality to the log ratio of the two products. Thus, applying (B.1) we find that $\mathsf{P}_{\hat{\mathsf{F}}}(C_{x,\epsilon}) \leq \mathsf{P}_n(C_{x,\epsilon})$ with equality if and only if $\hat{\mathsf{F}} = \mathsf{F}_n$, so that F_n is the unique maximizer.

B.3 Details of identities in proofs of LTS asymptotics

The formula (A.6).

We expand $S_s = (\hat{\sigma}^2_{\delta_n+s} - \hat{\sigma}^2_{\delta_n})/\sigma^2$ when $0 < s < h_n$. By definition

$$S_s = h_n^{-1} \sum_{i=1}^{h_n} \varepsilon^2_{(\delta_n+s+i)} - \{h_n^{-1} \sum_{i=1}^{h_n} \varepsilon_{(\delta_n+s+i)}\}^2 - h_n^{-1} \sum_{i=1}^{h_n} \varepsilon^2_{(\delta_n+i)} + \{h_n^{-1} \sum_{i=1}^{h_n} \varepsilon_{(\delta_n+i)}\}^2.$$

We have that $s < \delta_n < s + h_n$. We can then divide the h_n errors $\{\varepsilon_{(\delta_n+s+1)}, \ldots, \varepsilon_{(\delta_n+s+h_n)}\}$ into

$$\{\varepsilon_{(\delta_n+s+1)}, \ldots, \varepsilon_{(\delta_n+h_n)}\} \qquad \text{and} \qquad \{\varepsilon_{(\delta_n+h_n+1)}, \ldots, \varepsilon_{(\delta_n+s+h_n)}\}.$$

The first group are order statistics of 'good' errors. The second group consists of 'outliers' for which $\varepsilon_{(\delta_n+h_n+j)} = \varepsilon_{(\delta_n+h_n)} + \overline{\varepsilon}_{(j)}$ for $1 \leq j \leq s$. Thus, for the second moment we have

$$\begin{aligned}
\sum_{i=1}^{h_n} \varepsilon^2_{(\delta_n+s+i)} &= \sum_{i=s+1}^{h_n} \varepsilon^2_{(\delta_n+i)} + \sum_{j=1}^{s} \{\varepsilon_{(\delta_n+h_n)} + \overline{\varepsilon}_{(j)}\}^2 \\
&= \sum_{i=s+1}^{h_n} \varepsilon^2_{(\delta_n+i)} + s\varepsilon^2_{(\delta_n+h_n)} + 2\varepsilon_{(\delta_n+h_n)} \sum_{j=1}^{s} \overline{\varepsilon}_{(j)} + \sum_{j=1}^{s} \overline{\varepsilon}^2_{(j)}.
\end{aligned}$$

For the squared first moment we have

$$\begin{aligned}\{\sum_{i=1}^{h_n}\varepsilon_{(\delta_n+s+i)}\}^2 &= [\sum_{i=s+1}^{h_n}\varepsilon_{(\delta_n+i)}+\sum_{j=1}^{s}\{\varepsilon_{(\delta_n+h_n)}+\overline{\varepsilon}_{(j)}\}]^2\\ &= \{\sum_{i=s+1}^{h_n}\varepsilon_{(\delta_n+i)}\}^2+s^2\varepsilon^2_{(\delta_n+h_n)}+\{\sum_{j=1}^{s}\overline{\varepsilon}_{(j)}\}^2\\ &\quad+2s\varepsilon_{(\delta_n+h_n)}\sum_{i=s+1}^{h_n}\varepsilon_{(\delta_n+i)}+2s\varepsilon_{(\delta_n+h_n)}\sum_{j=1}^{s}\overline{\varepsilon}_{(j)}+2\{\sum_{i=s+1}^{h_n}\varepsilon_{(\delta_n+i)}\}\{\sum_{j=1}^{s}\overline{\varepsilon}_{(j)}\}.\end{aligned}$$

Further, we can expand

$$\begin{aligned}\sum_{i=1}^{h_n}\varepsilon^2_{(\delta_n+i)} &= \sum_{i=1}^{s}\varepsilon^2_{(\delta_n+i)}+\sum_{i=s+1}^{h_n}\varepsilon^2_{(\delta_n+i)},\\ \{\sum_{i=1}^{h_n}\varepsilon_{(\delta_n+i)}\}^2 &= \{\sum_{i=1}^{s}\varepsilon_{(\delta_n+i)}\}^2+\{\sum_{i=s+1}^{h_n}\varepsilon_{(\delta_n+i)}\}^2+2\sum_{i=1}^{s}\varepsilon_{(\delta_n+i)}\sum_{i=s+1}^{h_n}\varepsilon_{(\delta_n+i)}.\end{aligned}$$

Inserting the expansions of the moments in the expression for S_s gives

$$\begin{aligned}S_s &= \frac{1}{h_n}\{\sum_{i=s+1}^{h_n}\varepsilon^2_{(\delta_n+i)}+s\varepsilon^2_{(\delta_n+h_n)}+2\varepsilon_{(\delta_n+h_n)}\sum_{j=1}^{s}\overline{\varepsilon}_{(j)}+\sum_{j=1}^{s}\overline{\varepsilon}^2_{(j)}\}\\ &\quad-\frac{1}{h_n^2}[\{\sum_{i=s+1}^{h_n}\varepsilon_{(\delta_n+i)}\}^2+s^2\varepsilon^2_{(\delta_n+h_n)}+\{\sum_{j=1}^{s}\overline{\varepsilon}_{(j)}\}^2\\ &\qquad+2s\varepsilon_{(\delta_n+h_n)}\sum_{i=s+1}^{h_n}\varepsilon_{(\delta_n+i)}+2s\varepsilon_{(\delta_n+h_n)}\sum_{j=1}^{s}\overline{\varepsilon}_{(j)}+2\{\sum_{i=s+1}^{h_n}\varepsilon_{(\delta_n+i)}\}\{\sum_{j=1}^{s}\overline{\varepsilon}_{(j)}\}]\\ &\quad-\frac{1}{h_n}\{\sum_{i=1}^{s}\varepsilon^2_{(\delta_n+i)}+\sum_{i=s+1}^{h_n}\varepsilon^2_{(\delta_n+i)}\}\\ &\quad+\frac{1}{h_n^2}[\{\sum_{i=1}^{s}\varepsilon_{(\delta_n+i)}\}^2+\{\sum_{i=s+1}^{h_n}\varepsilon_{(\delta_n+i)}\}^2+2\sum_{i=1}^{s}\varepsilon_{(\delta_n+i)}\sum_{i=s+1}^{h_n}\varepsilon_{(\delta_n+i)}]\end{aligned}$$

This reduces as

$$S_s=\frac{s}{h_n}(1-\frac{s}{h_n})\varepsilon^2_{(\delta_n+h_n)}+A_n$$

where

$$
\begin{aligned}
A_n = & \frac{1}{h_n}\{\sum_{i=s+1}^{h_n}\varepsilon^2_{(\delta_n+i)} + 2\varepsilon_{(\delta_n+h_n)}\sum_{j=1}^{s}\overline{\varepsilon}_{(j)} + \sum_{j=1}^{s}\overline{\varepsilon}^2_{(j)}\} \\
& - \frac{1}{h_n^2}[\{\sum_{i=s+1}^{h_n}\varepsilon_{(\delta_n+i)}\}^2 + \{\sum_{j=1}^{s}\overline{\varepsilon}_{(j)}\}^2 \\
& \qquad + 2s\varepsilon_{(\delta_n+h_n)}\sum_{i=s+1}^{h_n}\varepsilon_{(\delta_n+i)} + 2s\varepsilon_{(\delta_n+h_n)}\sum_{j=1}^{s}\overline{\varepsilon}_{(j)} + 2\{\sum_{i=s+1}^{h_n}\varepsilon_{(\delta_n+i)}\}\{\sum_{j=1}^{s}\overline{\varepsilon}_{(j)}\}] \\
& - \frac{1}{h_n}\{\sum_{i=1}^{s}\varepsilon^2_{(\delta_n+i)} + \sum_{i=s+1}^{h_n}\varepsilon^2_{(\delta_n+i)}\} \\
& + \frac{1}{h_n^2}[\{\sum_{i=1}^{s}\varepsilon_{(\delta_n+i)}\}^2 + \{\sum_{i=s+1}^{h_n}\varepsilon_{(\delta_n+i)}\}^2 + 2\sum_{i=1}^{s}\varepsilon_{(\delta_n+i)}\sum_{i=s+1}^{h_n}\varepsilon_{(\delta_n+i)}]
\end{aligned}
$$

There are two cancellations: term 1 in line 1 with term 2 in line 4 and term 1 in line 2 with term 2 in line 5. Thus, A_n reduces to

$$
\begin{aligned}
A_n = & 2\varepsilon_{(\delta_n+h_n)}\frac{1}{h_n}\sum_{j=1}^{s}\overline{\varepsilon}_{(j)} + \frac{1}{h_n}\sum_{j=1}^{s}\overline{\varepsilon}^2_{(j)} - \{\frac{1}{h_n}\sum_{j=1}^{s}\overline{\varepsilon}_{(j)}\}^2 \\
& - 2\frac{s}{h_n}\varepsilon_{(\delta_n+h_n)}\frac{1}{h_n}\sum_{i=s+1}^{h_n}\varepsilon_{(\delta_n+i)} - 2\frac{s}{h_n}\varepsilon_{(\delta_n+h_n)}\frac{1}{h_n}\sum_{j=1}^{s}\overline{\varepsilon}_{(j)} \\
& - 2\{\frac{1}{h_n}\sum_{i=s+1}^{h_n}\varepsilon_{(\delta_n+i)}\}\{\frac{1}{h_n}\sum_{j=1}^{s}\overline{\varepsilon}_{(j)}\} - \frac{1}{h_n}\sum_{i=1}^{s}\varepsilon^2_{(\delta_n+i)} \\
& + \{\frac{1}{h_n}\sum_{i=1}^{s}\varepsilon_{(\delta_n+i)}\}^2 + 2\frac{1}{h_n}\sum_{i=1}^{s}\varepsilon_{(\delta_n+i)}\frac{1}{h_n}\sum_{i=s+1}^{h_n}\varepsilon_{(\delta_n+i)}.
\end{aligned}
$$

Rearrange as

$$
\begin{aligned}
A_n = & [\frac{1}{h_n}\sum_{j=1}^{s}\overline{\varepsilon}^2_{(j)} - \{\frac{1}{h_n}\sum_{j=1}^{s}\overline{\varepsilon}_{(j)}\}^2] - [\frac{1}{h_n}\sum_{i=1}^{s}\varepsilon^2_{(\delta_n+i)} - \{\frac{1}{h_n}\sum_{i=1}^{s}\varepsilon_{(\delta_n+i)}\}^2] \\
& + 2\frac{1}{h_n}\sum_{j=1}^{s}\overline{\varepsilon}_{(j)}\{(1-\frac{s}{h_n})\varepsilon_{(\delta_n+h_n)} - \frac{1}{h_n}\sum_{i=s+1}^{h_n}\varepsilon_{(\delta_n+i)}\} \\
& - 2\frac{1}{h_n}\sum_{i=s+1}^{h_n}\varepsilon_{(\delta_n+i)}\{\frac{s}{h_n}\varepsilon_{(\delta_n+h_n)} - \frac{1}{h_n}\sum_{i=1}^{s}\varepsilon_{(\delta_n+i)}\}.
\end{aligned}
$$

The terms in second and in third line, respectively, can be simplified to give

$$A_n = [\frac{1}{h_n}\sum_{j=1}^{s}\overline{\varepsilon}^2_{(j)} - \{\frac{1}{h_n}\sum_{j=1}^{s}\overline{\varepsilon}_{(j)}\}^2] - [\frac{1}{h_n}\sum_{i=1}^{s}\varepsilon^2_{(\delta_n+i)} - \{\frac{1}{h_n}\sum_{i=1}^{s}\varepsilon_{(\delta_n+i)}\}^2]$$
$$+ 2\frac{1}{h_n}\sum_{j=1}^{s}\overline{\varepsilon}_{(j)}\frac{1}{h_n}\sum_{i=s+1}^{h_n}\{\varepsilon_{(\delta_n+h_n)} - \varepsilon_{(\delta_n+i)}\}$$
$$- 2\frac{1}{h_n}\sum_{i=s+1}^{h_n}\varepsilon_{(\delta_n+i)}\frac{1}{h_n}\sum_{i=1}^{s}\{\varepsilon_{(\delta_n+h_n)} - \varepsilon_{(\delta_n+i)}\}.$$

which has the desired form $A_n = A_{n1} - A_{n2} + 2A_{n3} - 2A_{n4}$ □

The formula (A.11).
We have $h_n - \underline{s}_n \le s < h_n$ where $\overline{s}_n = (2\log h_n)^{-1/4}h_n$. By definition

$$\hat{\sigma}^2_{\delta_n+s}/\sigma^2 = \frac{1}{h_n}\sum_{i=1}^{h_n}\varepsilon^2_{(\delta_n+s+i)} - \{\frac{1}{h_n}\sum_{i=1}^{h_n}\varepsilon_{(\delta_n+s+i)}\}^2.$$

A residual sums of squares is invariant to subtracting a constant from each observation. Thus, subtracting $\varepsilon_{(\delta_n+h_n)}$ from each $\varepsilon_{(\delta_n+s+i)}$ gives

$$\hat{\sigma}^2_{\delta_n+s}/\sigma^2 = \frac{1}{h_n}\sum_{i=1}^{h_n}\{\varepsilon_{(\delta_n+s+i)} - \varepsilon_{(\delta_n+h_n)}\}^2 - [\frac{1}{h_n}\sum_{i=1}^{h_n}\{\varepsilon_{(\delta_n+s+i)} - \varepsilon_{(\delta_n+h_n)}\}]^2.$$

Split into 'good' and 'outlier' errors to get

$$\hat{\sigma}^2_{\delta_n+s}/\sigma^2 = \frac{1}{h_n}\sum_{i=s+1}^{h_n}\{\varepsilon_{(\delta_n+i)} - \varepsilon_{(\delta_n+h_n)}\}^2 + \frac{1}{h_n}\sum_{j=1}^{s}\{\varepsilon_{(\delta_n+h_n+j)} - \varepsilon_{(\delta_n+h_n)}\}^2$$
$$- [\frac{1}{h_n}\sum_{i=s+1}^{h_n}\{\varepsilon_{(\delta_n+i)} - \varepsilon_{(\delta_n+h_n)}\} + \frac{1}{h_n}\sum_{j=1}^{s}\{\varepsilon_{(\delta_n+h_n+j)} - \varepsilon_{(\delta_n+h_n)}\}]^2.$$

Note that $\varepsilon_{(\delta_n+h_n+j)} - \varepsilon_{(\delta_n+h_n)} = \overline{\varepsilon}_{(j)}$ while $\varepsilon_{(\delta_n+i)} < \varepsilon_{(\delta_n+h_n)}$ so that

$$\hat{\sigma}^2_{\delta_n+s}/\sigma^2 = \frac{1}{h_n}\sum_{i=s+1}^{h_n}\{\varepsilon_{(\delta_n+h_n)} - \varepsilon_{(\delta_n+i)}\}^2 + \frac{1}{h_n}\sum_{j=1}^{s}\overline{\varepsilon}^2_{(j)}$$
$$- [-\frac{1}{h_n}\sum_{i=s+1}^{h_n}\{\varepsilon_{(\delta_n+h_n)} - \varepsilon_{(\delta_n+i)}\} + \frac{1}{h_n}\sum_{j=1}^{s}\overline{\varepsilon}_{(j)}]^2.$$

Rearrange as

$$\begin{aligned}\hat{\sigma}^2_{\delta_n+s}/\sigma^2 = &\frac{1}{h_n}\sum_{i=s+1}^{h_n}\{\varepsilon_{(\delta_n+h_n)}-\varepsilon_{(\delta_n+i)}\}^2-[\frac{1}{h_n}\sum_{i=s+1}^{h_n}\{\varepsilon_{(\delta_n+h_n)}-\varepsilon_{(\delta_n+i)}\}]^2\\ &+\frac{1}{h_n}\sum_{j=1}^{s}\overline{\varepsilon}^2_{(j)}-\{\frac{1}{h_n}\sum_{j=1}^{s}\overline{\varepsilon}_{(j)}\}^2\\ &+2[\frac{1}{h_n}\sum_{i=s+1}^{h_n}\{\varepsilon_{(\delta_n+h_n)}-\varepsilon_{(\delta_n+i)}\}]\{\frac{1}{h_n}\sum_{j=1}^{s}\overline{\varepsilon}_{(j)}\}.\end{aligned}$$

The term in the second line satisfies

$$\frac{1}{h_n}\sum_{j=1}^{s}\overline{\varepsilon}^2_{(j)}-\{\frac{1}{h_n}\sum_{j=1}^{s}\overline{\varepsilon}_{(j)}\}^2=\frac{s}{h_n}(1-\frac{s}{h_n})\frac{1}{s}\sum_{j=1}^{s}\overline{\varepsilon}^2_{(j)}+(\frac{s}{h_n})^2[\frac{1}{s}\sum_{j=1}^{s}\overline{\varepsilon}^2_{(j)}-\{\frac{1}{s_n}\sum_{j=1}^{s}\overline{\varepsilon}_{(j)}\}^2].$$

Thus, we get

$$\hat{\sigma}^2_{\delta_n+s}/\sigma^2 = A_n = A_{n1}+A_{n2}+A_{n3}+2A_{n4},$$

which is (A.11), where

$$\begin{aligned}A_{n1} &= h_n^{-1}\sum_{i=s+1}^{h_n}\{\varepsilon_{(\delta_n+i)}-\varepsilon_{(\delta_n+h_n)}\}^2-[h_n^{-1}\sum_{i=s+1}^{h_n}\{\varepsilon_{(\delta_n+i)}-\varepsilon_{(\delta_n+h_n)}\}]^2,\\ A_{n2} &= (\frac{s}{h_n})^2[\frac{1}{s}\sum_{j=1}^{s}\overline{\varepsilon}^2_{(j)}-\{\frac{1}{s}\sum_{j=1}^{s}\overline{\varepsilon}_{(j)}\}^2],\\ A_{n3} &= \frac{s}{h_n}(1-\frac{s}{h_n})\frac{1}{s}\sum_{j=1}^{s}\overline{\varepsilon}^2_{(j)}\\ A_{n4} &= [h_n^{-1}\sum_{i=s+1}^{h_n}\{\varepsilon_{(\delta_n+h_n)}-\varepsilon_{(\delta_n+i)}\}]\{h_n^{-1}\sum_{j=1}^{s}\overline{\varepsilon}_{(j)}\}.\end{aligned}$$

This completes the proof of (A.11). □