Nuffield College
UNIVERSITY OF OXFORD

ECONOMICS DISCUSSION PAPERS

Paper No: 2023-W01

Least Trimmed Squares Asymptotics: Regression with leverage

By Vanessa Berenguer-Rico and Bent Nielsen

May 2023

Least Trimmed Squares Asymptotics: Regression with leverage

Vanessa Berenguer-Rico[*] & Bent Nielsen[†]

17 May 2023

Abstract

Least Trimmed Squares (LTS) regression is known to be robust to 'outliers' and in particular to bad leverage points. However, the current asymptotic theory for LTS is of limited use as its assumptions rule out leverage and the asymptotic distribution depends on the unknown contamination. We use a new model, where 'outlier' errors are extreme relative to the 'good' errors and where leverage effects are possible. We show that in this model the LTS estimator has an asymptotic distribution that is free of nuisance parameters. Thus, with the new model standard inference procedures apply while allowing a broad range of contamination.

1 Introduction

Least squares procedures are known to be highly sensitive to atypical observations. It is particularly so in the presence of leverage points, which attract the regression line towards them. In fact, leverage points can also affect other popular estimation procedures such as quantile regression (He et al., 1990). The influence of the leverage points can be bounded by using robust regression estimators such as the least trimmed of squares (LTS) estimator (Rousseeuw, 1984). However, a discussion of the asymptotic properties of such robust estimators in the presence of leverage points is largely absent in the literature. We derive asymptotic inference for LTS that is free of nuisance parameters and allows leverage, heteroscedasticity and temporal dependence.

Figure 1 illustrates the leverage effect for the ordinary least squares (OLS) and least absolute deviations (LAD) estimators. The data, from Rousseeuw and Leroy (1987), records log light intensity against log temperatures for $n = 47$ stars in a Hertzsprung-Russell diagram. The stars marked with bullet points are known as the main sequence while the four stars in the top right corner are red giants. The OLS and LAD regression lines are attracted by the giants. In contrast, the LTS line goes through the main sequence. The largest OLS or LAD residuals are associated with observations in the

[*] Mansfield College & Department of Economics, University of Oxford. Address for correspondence: Mansfield College, Oxford OX1 3TF, UK.E-mail:vanessa.berenguer-rico@economics.ox.ac.uk.

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

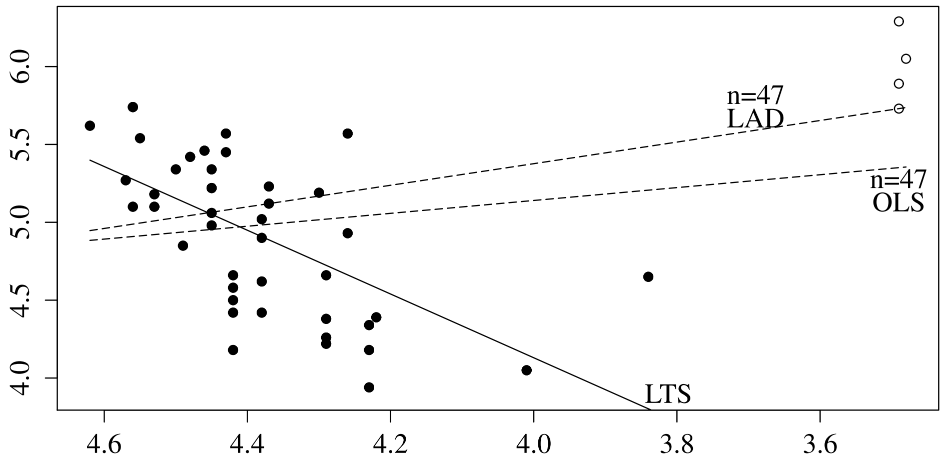

Figure 1: Leverage points. OLS, LAD, and LTS

main sequence rather than the giants. Thus, the data-analytic strategy of first estimating by OLS, removing observations with large residuals, and then re-estimating by OLS may not reveal that the giants are unusual (Welsh and Ronchetti, 2002).

While there are a number of approaches in the literature to deal with the presence of ‘outliers’, we will focus on the LTS estimator, which is commonly used. For instance, the MM estimator (Yohai, 1987) and the Forward Search algorithm (Atkinson et al., 2010) are often initialized by LTS. Variants include sparse LTS regression (Alfons et al., 2013) and a fraud detection algorithm (Rousseeuw et al., 2019). The LTS estimator is known to be very robust to leverage points with a high breakdown point (Rousseeuw and Leroy, 1987, Section 3.4). That is, the estimator remains bounded if, for a given sample, we distort nearly half of the observations in an arbitrary way.

The LTS estimator is computed as follows. The user specifies that a sample with n observations has h ‘good’ observations and $n-h$ ‘outliers’. The LTS estimator is the least squares estimator for the h sub-sample with the smallest residual sum of squares. In a location-scale model this search is of linear order, while in regression it is of binomial order, hence, making analysis harder in the regression context both from computational and theoretical viewpoints.

Figure 1 includes the LTS regression line for $h = 43$. LTS finds, precisely, the four giants as ‘outliers’. In contrast to OLS and LAD, the LTS line passes through the main sequence. In this example, with only two variables and four clear leverage points, the leverage points can be detected, or at least suspected, by graphical methods. However, in higher dimensions or with less evident atypical observations, LTS can be of great use in identifying ‘outliers’ and robustly estimating the regression line.

Our concern is how to conduct inference. The traditional statistical model in robust statistics is that of ϵ-contamination. In a regression context, the regression errors are assumed independent of the regressors and to be independent draws from a common ϵ-contaminated distribution, which mixes a normal distribution with a contamination distribution as popularized by Huber (1964). Such a model, with its independence of regression errors and regressors, cannot generate bad leverage points. Rather, bad leverage arises when ‘outliers’ are concentrated in a single point just as in Figure 1 where the giants essentially have a common regressor and large deviation from the LTS line.

The LTS estimator has been analyzed under ϵ-contamination assumptions by Butler (1982) for the location-scale case and by Rousseeuw (1985), Croux and Rousseeuw (1992), Čížek (2005), Víšek (2006), Johansen and Nielsen (2016) and Zuo (2022) for the regression case. Although normal, the asymptotic distribution depends on the contamination. This leaves the user with a serious nuisance parameter problem. Interestingly, the OLS procedure delivers consistent, efficient and nuisance parameter free inference for the slope estimators in the ϵ-contamination model. Finally, as a technical point, the current asymptotic theory for LTS assumes a compact parameter space, which goes against the tenet in robust statistics of seeking protection against arbitrary influence from ‘outliers’. For later reference, we term this approach as standard LTS, or in short SLTS.

Our analysis departs from this traditional approach and takes its starting point in the LTS model of Berenguer-Rico et al. (2023). The LTS model can generate a wide range of ‘outliers’ and in particular bad leverage points as those in Figure 1. The figure depicts data falling in two groups. The data closest to the LTS line appear to have a normal variation around that line. In contrast, the giant stars in the top right not only have a common regressor value, but their distance to the LTS line is also rather extreme relative to the more normal looking data closest to the LTS line. Thus, we need a model where the ‘outlier’ errors have regressor dependence and an extreme distributional behaviour. This cannot be captured by ϵ-contamination.

The LTS model has h ‘good’ regression errors which are normal and independent of the regressors, while the $n - h$ ‘outlier’ errors have support outside the range of the realized ‘good’ errors, but are otherwise unrestricted. The ‘outliers’ are therefore driven by the tail behaviour of the ‘good’ errors. Bad leverage is now possible. Berenguer-Rico et al. (2023) show that the LTS estimator maximizes the semi-parametric ϵ-likelihood of the LTS model in the sense of Scholz (1980). Moreover, they provide an asymptotic analysis of the LTS estimator in the location-scale case. Their proof relies on the fact that the LTS estimator is found by a linear search in the location-scale case and that the leverage feature is absent. Although informative, the location-scale model is of limited applicability. An asymptotic theory for the regression case that allows for leverage is central to the objective of robust statistics and highly relevant for practitioners but entirely missing in the literature. We present such a theory.

Specifically, we start by showing that the LTS estimator is bounded in probability. To the best of our knowledge, this result is new in the literature. Boundedness is derived under mild assumptions to the ‘good’ errors and the regressors. The proof adapts a recent argument for M-estimators with non-convex criterion functions (Johansen and Nielsen, 2019). The boundedness result resonates with the high breakdown point property of the LTS estimator and avoids a compact parameter space assumption.

Next, we show that the proportion of ‘good’ observations is consistently selected and derive the rate at which this consistent selection occurs. In doing so, we require mild tail conditions on the ‘good’ errors and the regressors. This allows normal and t errors, many forms of leverage, heteroscedasticity and non-stationary temporal dependence of the regressors.

The final result is an asymptotic expansion of the LTS estimator in terms of the infeasible OLS estimator for the 'good' observations. This oracle result is shown for both the regression parameters and the scale estimators. In contrast to the traditional approach to LTS, no nuisance correction and consistency factors are required, under the present assumptions. The usual asymptotic distribution theory for OLS estimators then applies under different assumptions to the 'good' observations such as i.i.d. or heteroscedastic structures and stationary or non-stationary time series.

In simulations, we consider OLS, LTS and SLTS inference for various contamined samples. The simulations confirm the asymptotic theory and the fact that the underlying model is of primary importance when conducting inferences with the LTS estimator. As an illustration, we revisit the stars data.

In practice, the user has to choose the number h of 'good' observations. Some estimators are implemented in software, but, again, asymptotic theory is largely absent in the literature. We hope to return to this in future research.

The asymptotic techniques presented here could be used for other robust regression estimators. For instance, the Least Trimmed sum of Absolute deviations (LTA) estimator is a robust version of the LAD estimator (Hössjer, 1994) that would be maximum likelihood in an LTS-type model with Laplace errors.

The paper is organized as follows. Section 2 describes the LTS estimator and the LTS model. Section 3 contains the asymptotic results: boundedness, consistent selection and asymptotic expansion. Section 4 gives examples of the wide range of regressors allowed by the theory. Section 5 illustrates the theory via simulations. Section 6 has an empirical illustration. Section 7 concludes. Proofs and technical derivations can be found in Appendices.

2 The LTS estimator and the LTS model

We consider the linear regression for a scalar y_i and a vector x_{in} of regressors given by

$$y_i = x_{in}'\beta + \sigma\varepsilon_i \qquad \text{for } i = 1, \ldots, n, \tag{2.1}$$

where x_{in} would usually include an intercept, but it does not have to. With this formulation of the model equation (2.1), all normalizations are built into the regressors x_{in} so that estimators for β will be $n^{1/2}$ consistent. For example, x_{in} could be an i.i.d. regressor, a level shift after a fraction of the sample $0 < \tau < 1$ so that $x_{in} = 1_{(i \leq \tau n)}$, or a normalized random walk, $x_{in} = n^{-1/2}\sum_{\ell=1}^{i}\psi_\ell$ with i.i.d. increments ψ_ℓ. In the notation for y_i, we suppress the dependence on n noting that in the asymptotic analysis y_i is always replaced by the right hand side of (2.1).

The LTS estimator can be defined as follows (Rousseeuw and van Driessen, 2000). Let ζ denote an h-subset of $(1, \ldots, n)$ with associated least squares estimators

$$\hat{\beta}_\zeta = (\sum_{i\in\zeta} x_{in}x_{in}')^{-1}\sum_{i\in\zeta} x_{in}y_i \qquad \text{and} \qquad \hat{\sigma}_\zeta^2 = h^{-1}\sum_{i\in\zeta}(y_i - x_{in}'\hat{\beta}_\zeta)^2, \tag{2.2}$$

where $\sum_{i\in\zeta} x_{in}x_{in}'$ is assumed invertible for any ζ. Then, the LTS estimator and the associated scale estimator are given by

$$\hat{\beta} = \hat{\beta}_{\hat{\zeta}} \qquad \text{and} \qquad \hat{\sigma}^2 = \hat{\sigma}_{\hat{\zeta}}^2 \qquad \text{where} \qquad \hat{\zeta} = \arg\min_{\zeta} \hat{\sigma}_\zeta^2. \tag{2.3}$$

That is, for a given number of 'good' observations h, the LTS estimator finds the h-subsample with the smallest residual sum of squares. The LTS estimator is maximum likelihood in the following model (Berenguer-Rico et al., 2023).

Model 1. *(The LTS model). Let $h \leq n$ be given. We condition on the random regressors $x_{1n}, \ldots, x_{nn}$. 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 independent with distribution functions $\mathsf{G}_j(z)$ for $z \in \mathbb{R}$, where G_j is continuous at 0, but may depend on x_{jn}. The 'outlier' errors are defined, for $j \notin \zeta$, by

$$\varepsilon_j = (\max_{i\in\zeta} \varepsilon_i + \xi_j)1_{(\xi_j>0)} + (\min_{i\in\zeta} \varepsilon_i + \xi_j)1_{(\xi_j<0)}. \tag{2.4}$$

The parameters are $\beta \in \mathbb{R}^{\dim x}$, $\sigma > 0$, ζ which is any h-subset of $i = 1, \ldots, n$ and G_j which are any $n - h$ arbitrary conditional distributions on $\mathbb{R}$, that are continuous at 0.

The LTS Model allows for 'outliers' in both the error term and the regressors. In particular, the 'outlier' errors are outside of the realized range of the 'good' errors and are characterized by an un-specified distribution $\mathsf{G}_j(z)$. We note that leverage can arise in the model. The model allows different distributions for the regressors for the good observations, x_{in} for $i \in \zeta$, and for the 'outlier' observations, x_{jn} for $j \notin \zeta$. Leverage can therefore arise in this model when the 'outlier' regressors are more concentrated than the 'good' regressors and the 'outlier' errors ε_j for $j \notin \zeta$ all have the same sign. Gallegos and Ritter (2009) present a related model, albeit without asymptotic analysis.

The asymptotic results in this paper use a series of assumptions that relax the structure of the LTS Model. Neither normality, i.i.d.ness nor full separation between 'outliers' and 'good' observations is needed in the asymptotic theory.

3 LTS Asymptotics

We present an asymptotic theory of the LTS estimator with h 'good' observations in the regression model $y_i = x_{in}'\beta + \sigma\varepsilon_i$ for $i = 1, \ldots, n$ and increasing values of h, n. The unknown parameters of the data generating process are denoted β, σ and ζ_n. Assumptions are given for the marginal distributions of the errors ε_i and the regressors x_{in} as we progress. Examples of permitted regressors are discussed in Section 4. We let $\#\zeta$ denote the count of elements in the set ζ.

3.1 Boundedness

A boundedness result is presented for the LTS estimator under assumptions to the second sample moment for the 'good' errors and to the frequency of small regressors.

Assumption 3.1. *Suppose*
(i) ***Frequency of 'good' observations****: $h/n \to \lambda$ where $\lambda > 1/2$;*
(ii) ***'Good' errors****: $h^{-1}\sum_{i\in\zeta_n} \varepsilon_i^2 = \mathrm{O}_\mathsf{P}(1)$.*
(iii) ***Frequency of small regressors****: Define*

$$F_{nh}(a) = \max_{\zeta:\#\zeta=h} \sup_{\delta:|\delta|=1} h^{-1} \sum_{i\in\zeta} 1_{(|x_{in}'\delta|\leq a)}. \tag{3.1}$$

Let ξ satisfy $0 < \xi < 2 - \lambda^{-1}$ and suppose

$$\lim_{(a,n)\to(0,\infty)} \mathsf{P}\{F_{nh}(a) > \xi\} = 0, \tag{3.2}$$

that is $\forall \epsilon > 0$, $\exists (a_0, n_0) > 0$: $\forall a \leq a_0, n \geq n_0$ then $\mathsf{P}\{F_{nh}(a) > \xi\} < \epsilon$.

The first result bounds the difference of the LTS estimator and the infeasible OLS estimator $\hat{\beta}_{\zeta_n}$ on the unknown set of 'good' observations ζ_n. The asymptotic theory of the OLS estimator $\hat{\beta}_{\zeta_n}$ is of course widely studied. We note that the LTS estimator may not be unique, so we establish a uniform bound over the sets $\mathcal{M}_n$ of minimizers ζ of $\hat{\sigma}^2_\zeta$.

Theorem 3.1. *Suppose Assumption 3.1. Let $\mathcal{M}_n$ denote the set of minimizers ζ of $\hat{\sigma}^2_\zeta$. Then, $\max_{\zeta\in\mathcal{M}_n} |\hat{\beta}_\zeta - \hat{\beta}_{\zeta_n}| = \mathrm{O}_{\mathsf{P}}(1)$.*

The boundedness result in Theorem 3.1, its condition for the frequency of small regressors and its proof are inspired by the analysis of M-estimators in Johansen and Nielsen (2019). Two major differences are that, here, the criterion function has an explicit structure and we are searching for a set of 'good' observations of a known size h. This results in weaker assumptions to the errors and in a measure of the frequency of small observations $F_{nh}(a)$ that reflects the search for the 'good' observations. When $h = n$, then $F_{nn}(a)$ is the quantity $F_n(a)$ in Johansen and Nielsen (2019) and it is related to quantities used for M- and S-estimators (Chen and Wu, 1988; Davies, 1990).

The boundedness of the LTS estimator derived in Theorem 3.1, holds under very mild assumptions on the good errors and the frequency of small regressors. We note that no other structure is imposed. In particular, the defining feature of the LTS Model 1 of placing 'outlier' errors outside the range of 'good' errors is not needed.

Remark 3.1. *Assumption 3.1(ii) implies that $\hat{\sigma}$ is bounded. Indeed, since $\hat{\sigma}$ is a minimizer then $\hat{\sigma}^2 \leq \hat{\sigma}^2_{\zeta_n}$ where $\hat{\sigma}^2_{\zeta_n}/\sigma^2 \leq h^{-1}\sum_{i\in\zeta_n} \varepsilon_i^2$ by the model equation.*

Remark 3.2. *Assumption 3.1(iii) covers a wide range of regressors including leverage points – examples are given in Section 4. It implies that $\hat{\Sigma}_\zeta = h^{-1}\sum_{i\in\zeta} x_{in}x'_{in}$ is positive definite in large samples for all ζ as required in (2.2).*

3.2 Consistent selection of 'good' observations

We start by showing that the proportion of wrongly classified observations vanishes. The convergence rate is improved subsequently. We note that $\#(\zeta \cap \zeta_n)$ is the number of 'good' observations in ζ. The numbers of wrongly classified 'good' observations and wrongly classified 'outliers' satisfy $\#(\zeta^c \cap \zeta_n) = \#(\zeta \cap \zeta_n^c)$, since $h = \#(\zeta^c \cap \zeta_n) + \#(\zeta \cap \zeta_n)$ and $h = \#(\zeta \cap \zeta_n^c) + \#(\zeta \cap \zeta_n)$. The proportion of wrong classifications is then $\#(\zeta \cap \zeta_n^c)/h$. Let $\|m\|$ denote the spectral norm of a matrix m.

Assumption 3.2. *Let $m_n^2 = \min\{(\min_{i\in\zeta_n} \varepsilon_i)^2, (\max_{i\in\zeta_n} \varepsilon_i)^2\}$. Suppose*
(i) ***'Outlier' errors****: $\min_{j\notin\zeta_n} \varepsilon_j^2 \geq m_n^2\{1 + \mathrm{o}_{\mathsf{P}}(1)\}$.*
(ii) ***Regressors****: $\|\sum_{i=1}^n x_{in}x'_{in}\| = \mathrm{O}_{\mathsf{P}}(n)$.*
(iii) ***Infeasible OLS estimator****: $\hat{\beta}_{\zeta_n} = \mathrm{O}_{\mathsf{P}}(1)$.*

Theorem 3.2. *Suppose Assumptions 3.1, 3.2. Let $\mathcal{M}_n$ denote the set of minimizers ζ of $\hat{\sigma}^2_\zeta$. Then,* $\max_{\zeta\in\mathcal{M}_n} \#(\zeta \cap \zeta_n^c)/h = \mathrm{O}_{\mathsf{P}}(1/m_n^2)$.

In Assumption 3.2, part (i) is a relaxed version of the defining feature of the LTS model that 'outlier' errors are more extreme than 'good' errors. Specifically, it does not impose complete separation but allows for some overlap between 'good' and 'outlier' errors. The key feature in this assumptions is that the 'outlier' errors are driven by the extreme 'good' errors. Part (ii) is a mild assumption to the regressors but excludes diverging 'outliers' among the regressors. This is consistent with the recommendation of Rousseeuw (1994) to start an LTS analysis by detecting 'outliers' among the regressors. Part (ii) allows standard regressors, see Section 4.

Theorem 3.2 provides a consistency result whenever m_n^2 diverges, that is, whenever the 'good' errors have unbounded support. In that case, the Theorem shows that the proportion of wrong classifications vanishes in that $\#(\zeta \cap \zeta_n^c)/h = \mathrm{O}_{\mathsf{P}}(1/m_n^2) = \mathrm{o}_{\mathsf{P}}(1)$. Since $\#(\zeta \cap \zeta_n) + \#(\zeta \cap \zeta_n^c) = h$, we also get that the proportion of correctly classified 'good' observations goes to unity, that is $\#(\zeta \cap \zeta_n)/h = 1 + \mathrm{o}_{\mathsf{P}}(1)$.

3.3 Improving the rate of consistency

Theorem 3.2 gave conditions under which $\#(\zeta \cap \zeta_n^c)/h = \mathrm{O}_{\mathsf{P}}(1/m_n^2)$, which typically has a slow rate. Here, we improve the consistency rate. This requires assumptions to the intermediate extreme values of the 'good' errors and regressors.

Let $\lfloor . \rfloor$ denote the floor function. We define the square root of a symmetric, positive definite matrix m as follows. Decompose $m = vwv'$, where $v' = v^{-1}$ and w is diagonal and define $m^p = vw^pv'$ for any $p \in \mathbb{R}$.

Assumption 3.3. *Let* $m_n^2 = \min\{(\min_{i\in\zeta_n} \varepsilon_i)^2, (\max_{i\in\zeta_n} \varepsilon_i)^2\}$. *Suppose*
(i) ***'Good' errors****: ε_i for $i \in \zeta_n$ satisfy*
- (a) $1/m_n^2 = \mathrm{o}_{\mathsf{P}}(1)$;
- (b) $\max_{i\in\zeta_n} \varepsilon_i^2/m_n^2 = \mathrm{O}_{\mathsf{P}}(1)$;
- (c) *Let ε_i^2 for $i \in \zeta_n$ have order statistics $\psi_1 \leq \cdots \leq \psi_h$. Then the intermediate extreme values satify $\forall 0 < \rho < 1$, $\exists C_\rho < 1$: $\psi_{h-\lfloor h^\rho \rfloor}/m_n^2 \leq C_\rho + \mathrm{o}_{\mathsf{P}}(1)$;*
- (d) *Extremes are, at most, of polynomial order: $m_n^2 = \mathrm{o}_{\mathsf{P}}(n^\eta)$ for some $0 < \eta < 1/2$.*

(ii) ***Regressors****: Let $x'_{jn}(\sum_{i\in\zeta_n} x_{in}x'_{in})^{-1}x_{jn}$ for $j = 1, \ldots, n$ have order statistics $\phi_1 \leq \cdots \leq \phi_n$ satisfying*
- (a) *$\forall \delta > 0$, $\exists 0 < r < 1 - \eta$ so that $\phi_{n-\lfloor n^r \rfloor}/\phi_n \leq \delta + \mathrm{o}_{\mathsf{P}}(1)$;*
- (b) $\phi_n = \mathrm{O}_{\mathsf{P}}(m_n^2/h)$.

(iii) ***Infeasible OLS estimator****:* $(\sum_{i\in\zeta_n} x_{in}x'_{in})^{1/2}(\hat{\beta}_{\zeta_n} - \beta) = \mathrm{O}_{\mathsf{P}}(1)$.

Theorem 3.3. *Suppose Assumptions 3.1, 3.2, 3.3. Let $\mathcal{M}_n$ denote the set of minimizers ζ of $\hat{\sigma}^2_\zeta$. Then, for all $0 < \theta < 1$, it holds* $\max_{\zeta\in\mathcal{M}_n} \#(\zeta \cap \zeta_n^c)/h = \mathrm{O}_{\mathsf{P}}(h^{\theta-1})$.

Assumption 3.3(i) concerns the tail behaviour of the 'good' errors. Extreme tails can be assessed using the multiplicative strong law of large numbers (Galambos, 1978, Theorem 4.4.4). Intermediate tails can be assessed by modifying Chibisov (1964, Lemma 1). Details are given in Appendix B.

Example 3.1. *Assumption 3.3(i) holds in the following cases, see Appendix B for details.*
(i) Normal distribution with $m_n^2/2\log h \to 1$ *a.s. and* $C_\rho = 1-\rho$*;*
(ii) Laplace distribution with $m_n/\log h \to 1$ *a.s. and* $C_\rho = (1-\rho)^2$*;*
(iii) Double geometric distribution with $m_n/\log h \to 1$ *a.s.,* $C_\rho = (1-\rho)^2$*;*
(iv) t_d *distribution with* $d > \eta^{-1} > 2$ *degrees of freedom. In this case,* $m_n/h^{1/d}$ *converges in distribution and any choice of* C_ρ *function suffices.*

Assumption 3.3(*ii*) restricts the regressors' tails. Even so, it allows a variety of regressors and leverage effects. Examples follow in Section 4.

3.4 Main result

Next, we show that the asymptotic distribution of the normalized LTS estimator coincides with that of the normalized infeasible OLS estimator on the 'good' observations.

Theorem 3.4. *Suppose Assumptions 3.1, 3.2, 3.3. Let* $\mathcal{M}_n$ *denote the set of minimizers* ζ *of* $\hat\sigma^2_\zeta$*. Then*
(a) $\max_{\zeta\in\mathcal{M}_n} h^{1/2}|\hat\sigma^2_\zeta - \hat\sigma^2_{\zeta_n}| = \mathrm{o}_{\mathsf{P}}(1)$.
(b) $\max_{\zeta\in\mathcal{M}_n} |(\sum_{i\in\zeta} x_{in}x_{in}')^{1/2}(\hat\beta_\zeta - \beta) - (\sum_{i\in\zeta_n} x_{in}x_{in}')^{1/2}(\hat\beta_{\zeta_n} - \beta)| = \mathrm{o}_{\mathsf{P}}(1)$.

Theorem 3.4 generalizes the asymptotic theory for the location-scale case (Berenguer-Rico et al., 2023). The present assumptions are slightly different and allow t_d distributions with their polynomial tails.

The asymptotic distribution for the LTS estimator can be derived from standard OLS results applied to the infeasible OLS estimator on the 'good' observations. Under standard OLS assumptions to the 'good' observations, so that for $i \in \zeta_n$, suppose (x_{in}', ε_i) are i.i.d. with finite fourth moments while $\mathsf{E}(\varepsilon_i|x_{in}) = 0$ and $\mathsf{E}(\varepsilon_i^2|x_{in}) = \sigma^2$, then we get that $h^{-1}\sum_{i\in\zeta_n} x_{in}x_{in}' \to \Sigma_x$ in probability and

$$\hat\sigma \overset{\mathsf{P}}{\to} \sigma \quad \text{and} \quad \Big(\sum_{i\in\hat\zeta} x_{in}x_{in}'\Big)^{1/2}(\hat\beta - \beta)/\hat\sigma \overset{\mathsf{D}}{\to} \mathsf{N}(0, I_{\dim x}). \tag{3.3}$$

The 'good' errors can be heteroscedastic as long as Assumptions 3.1, 3.2, 3.3 are satisfied. For instance, suppose that $y_i = \alpha + \beta x_i + \sigma\varepsilon_i$ with $\varepsilon_i|x_i \sim \mathsf{N}(0, x_i^\omega)$ and $\omega > 2$ for $i \in \zeta_n$. Suppose, $x_i^{-\omega}$ is i.i.d. gamma with shape and inverse scale of $p/2$. Then, for $i \in \zeta_n$, $\varepsilon_i \sim$ *i.i.d.* t_p. If $p > 4$, then Assumptions 3.1, 3.2, 3.3 are satisfied, see Appendix C for details. Theorem 3.4 says that in this case the LTS estimator has the same asymptotic distribution as the OLS estimator on the 'good' observations. Since these present heteroscedasticity, valid inference requires Eicker-Huber-White standard errors for LTS in this case. In turn, these will be asymptotically equivalent to corrected standard errors for the infeasible OLS estimator.

4 Examples of regressors

We illustrate the regressor conditions.

4.1 Assumption to small regressors: General remarks

Recall the frequency of small regressors $F_{nh}(a)$ in (3.1). Assumption 3.1(iii) implies that $\hat{\Sigma}_\zeta = h^{-1}\sum_{i\in\zeta} x_{in}x_{in}'$ is positive definite in large samples for all ζ (Johansen and Nielsen, 2019). Indeed, for all $\delta \neq 0$,

$$\delta'\hat{\Sigma}_\zeta\delta \geq \min_\zeta h^{-1}\sum_{i\in\zeta}\delta' x_{in}x_{in}'\delta 1_{(|x_{in}'\delta|>a)} \geq a^2 \min_\zeta h^{-1}\sum_{i\in\zeta} 1_{(|x_{in}'\delta|>a)}.$$

Since $h^{-1}\sum_{i\in\zeta} 1_{(|x_{in}'\delta|>a)} = 1 - h^{-1}\sum_{i\in\zeta} 1_{(|x_{in}'\delta|\leq a)} \geq 1 - F_{nh}(a)$, we get

$$\delta'\hat{\Sigma}_\zeta\delta \geq a^2\{1 - F_{nh}(a)\} \geq a^2\{1-(\xi+\epsilon)\} > 0,$$

with large probability for large n and for $\epsilon < 1-\xi$ and some $a > 0$.

The Assumption to $F_{nh}(a)$ involves a supremum over all h-subsamples. We present two bounds for $F_{nh}(a)$ that avoid the supremum over sub-sets. We illustrate their use in Section 4.2 below. The first bound to $F_{nh}(a)$ involves the regressors for all observations:

$$F_{nh}(a) \leq (n/h)F_{nn}(a), \tag{4.1}$$

noting that $\sum_{i\in\zeta} 1_{(\cdot)} \leq \sum_{i=1}^n 1_{(\cdot)}$. In particular, Assumption 3.1(iii) holds whenever $F_{nn}(a) = \mathsf{o_P}(1)$. See Examples 4.1-4.4 below.

The second bound to $F_{nh}(a)$ only involves the regressors of the 'good' observations:

$$F_{nh}(a) \leq F_{hh}(a) + (n-h)/h, \tag{4.2}$$

with the convention $F_{hh}(a) = \sup_{\delta:|\delta|=1} h^{-1}\sum_{i\in\zeta_n} 1_{(|x_{in}'\delta|\leq a)}$. This bound follows through $\sum_{i\in\zeta} 1_{(\cdot)} = \sum_{i\in\zeta\cap\zeta_n} 1_{(\cdot)} + \sum_{i\in\zeta\cap\zeta_n^c} 1_{(\cdot)} \leq \sum_{i\in\zeta_n} 1_{(\cdot)} + \sum_{i\in\zeta_n^c} 1$. In particular, if $F_{hh}(a) = \mathsf{o_P}(1)$ then the right hand side of (4.2) has limit $\lambda^{-1} - 1$, which is strictly smaller than $2-\lambda^{-1}$ whenever $\lambda > 2/3$. This leaves space for choosing a ξ so that Assumption 3.1(iii) is satisfied. Thus, the LTS estimator will be bounded under a wide range of 'good' regressors, see Examples 4.1-4.4 below, while the 'outlier' regressors are arbitrary.

4.2 Examples

We analyze regressors with respect to the boundedness Assumption 3.1(iii) and the tail behavior condition in Assumption 3.3(ii). First, we consider five examples of regressors without 'outliers'. For Assumption 3.1(iii), it then suffices to analyze F_{nn} and then apply the inequality (4.1).

Example 4.1. ***Polynomial regressors***. *Let $x_{in}' = \{1, (i/n)^q\}$ for $q > 0$ or $0 > q > -1/2$. Then $F_{nn}(a) = \mathsf{o_P}(1)$ (Johansen and Nielsen, 2019, Example 3.2, 3.3) and Assumption 3.1(iii) follows. Since x_{in} is bounded, Assumption 3.3(ii) holds.*

Example 4.2. ***i.i.d. regressors***. *Let $x_{in}' = (1, z_{in})$ where z_{in} is i.i.d. with bounded, continuous density. Then $F_{nn}(a) = \mathsf{o_P}(1)$ (Johansen and Nielsen, 2019, Theorem 3.3). Assumption 3.3(ii) follows if z_{in} has thinner tails than or the same tails as the 'good' errors. For instance, z_{in} for $1 \leq i \leq n$ and ε_i for $i \in \zeta_n$ could be normal.*

Example 4.3. ***Stationary regressors***. *Let $x'_{in} = (1, z_{in})$ where z_{in} is a stationary, normal autoregression. Then $F_{nn}(a) = \mathrm{o_P}(1)$ (Johansen and Nielsen, 2019, Example 3.7). Assumption 3.3(ii) follows if the 'good' errors are also normal, since the distribution of the intermediate extreme values for stationary, normal autoregressions is same as for i.i.d. normal variables (Watts et al., 1982, Theorem 3.3).*

Example 4.4. ***Random walk***. *Let $x'_{in} = (1, z_{in})$ so $z_{in} = n^{-1/2}\sum_{\ell=1}^{i}\psi_i$ where ψ_i is i.i.d. multivariate, zero mean normal. Then $F_{nn}(a) = \mathrm{o_P}(1)$ (Johansen and Nielsen, 2019, Theorem 3.4) and Assumption 3.1(iii) follows. The maximum of a normalized random walk converges in distribution so that $\phi_n = \mathrm{O_P}(1/h)$ and Assumption 3.3(ii) follows. The normalized estimator $h^{1/2}(\hat{\beta}-\beta)$ will have an asymptotic Dickey-Fuller type distribution. The exact expression depends on how 'good' and 'outlier' errors alternate (Johansen and Nielsen, 2009).*

Example 4.5. ***Binary regressors***. *Let $x'_{in} = \{1, 1_{(1\leq\tau n)}\}$. Here, $F_{nn}(a) = \max(\tau, 1-\tau)$ for small $a > 0$ (Johansen and Nielsen, 2019, Example 3.1). Suppose $\max(\tau, 1-\tau) < 2\lambda - 1$, which is satified for instance when $\tau = 1/2$ and $\lambda > 3/4$. The inequality (4.1) then shows that $F_{nh}(a) \leq (n/h)F_{nn}(a) < 2 - 1/\lambda - \epsilon + \mathrm{o}(1)$ for small $\epsilon > 0$. Thus, an $\xi < 2 - 1/\lambda$ can be found so that $F_{nh}(a) \leq \xi$ with large probability. Assumption 3.1(iii) follows. The regressor is bounded and Assumption 3.3(ii) follows.*

If the 'good' regressors are regular they can be combined with 'outlier' regressors without much structure. In particular, if $F_{hh}(a) \to 0$ as $(a, n) \to (0, \infty)$ then Assumption 3.1(*iii*) is satisfied through the bound (4.2) and it suffices to check that the 'outlier' regressors do not drift too fast to satisfy Assumption 3.3(*iii*). We give a specific example, which we will consider in the simulation study in Section 5.

Example 4.6. ***Leverage***. *Let the 'good' regressors x_i be i.i.d. $\mathsf{U}[-10, 10]$, that is uniform on $[-10, 10]$, while the 'outlier' regressors satisfy $x_j = 10 + e_j + d$, where e_j are i.i.d.$\mathsf{U}[0, 1]$ while d is to be chosen. Here the 'outlier' regressors are spread out a bit to facilitate estimation by fast LTS (Rousseeuw and van Driessen, 2000).*

Assumption 3.1(iii) is satisfied through (4.2) whenever $\lambda > 2/3$. To see this, let $x_{in} = x_i$. The 'good' regressors satisfy $F_{hh}(a) \to 0$ as $(a, n) \to (0, \infty)$ by Example 4.2.

We now turn to Assumption 3.3(ii). First, suppose d is bounded. Apply the Law of Large Numbers separately to 'good' and 'outlier' regressors to see that $n^{-1}\sum_{i=1}^{n} x_{in}x'_{in}$ converges in probability and Assumption 3.2(ii) follows.

Second, let $d = \sqrt{n}$. Assumption 3.2(ii) fails, as $n^{-1}\sum_{i\in\zeta_n^c} x_{in}^2$ diverges at rate $n-h$, when choosing $x_{in} = x_i$. Similarly, Assumption 3.3(ii) fails since $H_n = hx_{jn}^2/\sum_{i\in\zeta_n} x_{in}^2$ diverges at order n for $j \in \zeta_n^c$, since $x_{jn}^2 = n\{1 + \mathrm{o_P}(1)\}$ and $h^{-1}\sum_{i\in\zeta_n} x_{in}^2 = 100/3 + \mathrm{o_P}(1)$. Thus, H_n exceeds m_n^2 for any permitted distribution for the 'good' errors.

We could force Assumption 3.2(ii) to hold by defining $x_{in} = x_i/n^{1/2}$. In that case, Assumption 3.1(iii) would fail in light of Example 4.5 with $\tau = \lambda$ as we would need $\max(\lambda, 1-\lambda) < 2\lambda - 1$ which is not possible for any $0 \leq \lambda \leq 1$.

5 Simulations

We study the finite sample properties of t-tests for $\beta_0 = \beta_1 = 0$ in the linear model

$$y_i = \beta_0 + \beta_1 z_i + \sigma \varepsilon_i. \tag{5.1}$$

We analyze three statistics and six data generating processes. We consider sample sizes $n = 25, 100, 400, 1600$ with $h/n = \lambda = 0.8$ and use 10^4 repetitions. The code was written in Matlab with LTS estimation done using the `mlts.m` code by Agullo et al. (2008).

Tests. We consider t-statistics $t_{k,s} = (\hat{\beta}_{k,s} - \mu)/\mathsf{se}_{k,s}$, where k and s denote parameter and estimation method, respectively. The t-tests rejects for $|t_{k,s}| > q$, where q is the normal 97.5% quantile giving a target level of 5%. We study three estimators so that $s \in \{OLS, LTS, SLTS\}$.

The OLS estimator is $\hat{\beta}_{OLS} = (\sum_{i=1}^n x_i x_i')^{-1}(\sum_{i=1}^n x_i y_i)$ with $x_i = (1, z_i)'$ and se^2_{OLS} is the product of $\hat{\sigma}^2_{OLS} = (n-2)^{-1}\sum_{i=1}^n (y_i - x_i'\hat{\beta}_{OLS})^2$ and the relevant diagonal element of $(\sum_{i=1}^n x_i x_i')^{-1}$.

The LTS estimators $\hat{\beta}_{LTS}$ and $\hat{\sigma}_{LTS}$ are given in (2.3). Further, se^2_{LTS} is the product of $\hat{\sigma}^2_{LTS}$ and the relevant diagonal element of $(\sum_{i \in \hat{\zeta}_{LTS}} x_i x_i')^{-1}$.

The SLTS test uses the standard LTS correction factor following common practice (Croux and Rousseeuw, 1992). Thus, $\hat{\beta}_{SLTS} = \hat{\beta}_{LTS}$. The correction factor assumes the 'good' observations are truncated normal so that $\varsigma^2_{h/n} = \int_{-c}^{c} x^2 \varphi(x)dx / \int_{-c}^{c} \varphi(x)dx$ with c chosen so that $\int_{-c}^{c} \varphi(x)dx = h/n$. In particular, $\varsigma^2_{0.8} = 0.438$. Then, $\hat{\sigma}^2_{SLTS} = \hat{\sigma}^2_{LTS}/\varsigma^2_{h/n}$ and $\mathsf{se}^2_{SLTS} = \mathsf{se}^2_{LTS}/(\varsigma^4_{h/n})$.

Data Generating Processes (DGPs). All DGPs are of the form (5.1), where $\beta_0 = \beta_1 = 0$ and $\sigma = 1$. In all cases, the 'good' errors are i.i.d. $\mathsf{N}(0,1)$ and the 'good' regressors are i.i.d. uniform on $[-10, 10]$, denoted $\mathsf{U}[-10, 10]$.

DGP 1 has no contamination, hence, all errors are i.i.d $\mathsf{N}(0,1)$ and all regressors are i.i.d $\mathsf{U}[-10, 10]$.

DGPs 2–6 have LTS-type contamination in the errors of the form (2.4), where $\varepsilon_j = (\max_{i \in \zeta} \varepsilon_i + \xi_j)1_{(\xi_j > 0)} + (\min_{i \in \zeta} \varepsilon_i + \xi_j)1_{(\xi_j < 0)}$.

DGPs 2–3 have no contamination in the regressors, which are all i.i.d $\mathsf{U}[-10, 10]$. The 'outlier' errors are defined so that $\xi_j - \nu^+ 1_{(\xi_j > 0)} + \nu^- 1_{(\xi_j < 0)}$ is i.i.d. normal $\mathsf{N}(0,1)$ and the constants ν^+ and ν^- separate 'good' and 'outlying' errors. DGP 2 has $\nu^+ = \nu^- = 0$ and DGP 3 has $\nu^+ = 3$, $\nu^- = 1$.

DGPs 4–6 have contamination in both errors and regressors following Example 4.6. 'Outlier' errors and regressors are positive and given by $\xi_j = u_j + c$ and $x_j = 10 + e_i + d$, where u_j and e_j are independent and i.i.d.$\mathsf{U}[0, 1]$, while DGP 4 has $c = d = 0$, DGP 5 has $c = 10$, $d = 0$, and DGP 6 has $c = 10$, $d = n^{0.5}$.

Tables 1,2 report simulated rejection frequencies for nominal 5% tests on the intercept and the slope, respectively. Results are based on 10^4 repetitions. The Monte Carlo standard error is 0.2% for correctly sized tests.

DGP 1 has no contamination. Both the OLS and the SLTS statistics perform well in small samples. The LTS statistic is not using the correct standard error and it is oversized for all samples sizes. These results are seen both for intercept and slope.

DGPs 2–3 have contamination in the errors, but not in the regressors. The LTS test has empirical size approaching 5% as the sample size increases for both intercept

Table 1: Simulated rejection frequencies for nominal 5% tests on intercepts.

method	n	DGP1	DGP2	DGP3	DGP4	DGP5	DGP6
OLS	25	0.060	0.083	0.074	0.218	0.378	0.262
	100	0.053	0.080	0.128	0.732	0.983	0.628
	400	0.050	0.104	0.323	1.000	1.000	0.888
	1600	0.055	0.162	0.741	1.000	1.000	0.948
	6400	0.049	0.293	0.979	1.000	1.000	0.900
LTS	25	0.377	0.268	0.084	0.581	0.063	0.064
	100	0.389	0.177	0.058	0.820	0.053	0.053
	400	0.392	0.113	0.052	0.674	0.050	0.050
	1600	0.400	0.069	0.048	0.263	0.049	0.067
	6400	0.389	0.053	0.050	0.052	0.047	0.918
SLTS	25	0.039	0.017	0.002	0.169	0.000	0.001
	100	0.042	0.003	0.000	0.608	0.000	0.000
	400	0.051	0.000	0.000	0.654	0.000	0.000
	1600	0.050	0.000	0.000	0.220	0.000	0.017
	6400	0.048	0.000	0.000	0.002	0.000	0.724

Table 2: Simulated rejection frequencies for nominal 5% tests on slopes.

method	n	DGP1	DGP2	DGP3	DGP4	DGP5	DGP6
OLS	25	0.064	0.057	0.059	0.754	0.999	1.000
	100	0.051	0.052	0.051	1.000	1.000	1.000
	400	0.053	0.049	0.048	1.000	1.000	1.000
	1600	0.047	0.049	0.049	1.000	1.000	1.000
	6400	0.050	0.051	0.050	1.000	1.000	1.000
LTS	25	0.366	0.290	0.092	0.905	0.065	0.066
	100	0.374	0.191	0.060	0.877	0.050	0.050
	400	0.386	0.135	0.053	0.683	0.052	0.052
	1600	0.390	0.098	0.051	0.279	0.054	0.069
	6400	0.398	0.084	0.053	0.065	0.049	1.000
SLTS	25	0.035	0.023	0.003	0.628	0.000	0.001
	100	0.039	0.003	0.000	0.859	0.000	0.000
	400	0.046	0.000	0.000	0.655	0.000	0.000
	1600	0.046	0.000	0.000	0.220	0.000	0.018
	6400	0.047	0.000	0.000	0.002	0.000	1.000

and slope. The LTS procedure works better in finite samples under DGP 3 than DGP 2, since DGP 3 has more separation of 'good' and 'outlier' errors. The OLS procedure performs differently for intercept and slope. Specifically, the empirical size for the intercept increases with sample size; whereas the empirical size for the slope statistic is approximately 5% for all sample sizes considered. The SLTS tests have empirical size close to zero for almost all sample sizes considered for both intercept and slope.

DGPs 4–5 have leverage points with positive contamination in the errors and contaminated regressors located around the largest ‘good’ regressors. The LTS test has empirical size approaching 5% as the sample size increases, for both intercept and slope. We note that LTS works better in finite samples under DGP 5 than DGP 4, since DGP 5 has more separation of ‘good’ and ‘outlier’ errors. The OLS test has empirical size approaching one for both intercept and slope. The SLTS test has a more complicated behaviour. For DGP 4, the size first increases for both intercept and slope and then decreases to near zero for large samples. For DGP 5, the size is near zero in all cases.

DGP 6 has positive contamination in the errors and positively contaminated regressors which are growing at the order of $n^{1/2}$ and larger than the largest ‘good’ regressors. The leverage points therefore become relatively closer to the regression line as n grows and are designed to violate Assumption 3.2(ii) in the LTS asymptotics. The LTS test has empirical size around 5% for most sample sizes, but the size jumps to near unity for the largest sample size. This supports the idea of looking for ‘outliers’ in the regressors before using the LTS estimator (Rousseeuw, 1994). The OLS test has an empirical size that is steadily growing with the sample size for the intercept and constantly at unity for the slope. The SLTS test has empirical size close to 0% for most sample sizes, but the size jumps to near unity for the largest sample size.

6 Empirical illustration

We consider the stars data of Rousseeuw and Leroy (1987) as shown in Figure 1. The data consists of a sample of $n = 47$ stars on the CYG OB1 cluster, for which the log light intensity and log temperature is recorded for each star. A detailed description can be found in Berenguer-Rico et al. (2023). In short, from the right, the first four stars are red supergiants of M-type, the fifth star is of F-type, the next 31 stars (1 doublet) are of B-type, and the final 11 stars (1 doublet) to the left are of O-type. The figure shows OLS and LAD fits, along with the LTS line for $h = 43$. Following Berenguer-Rico et al. (2023), we choose $h = 42$ pointing at $n - h = 5$ ‘outliers’. This choice is based on an estimator for λ developed in the above mentioned paper, where it was also also found that the ‘good’ errors may be approximately normal. With $h = 42$, the LTS estimator finds as outliers the four red giant stars and the F-type star that is unique in this sample and distinct from the B and O type stars in the main sequence.

Once we have estimated the $h = 42$ ‘good’ observations, we conduct inference according to the above theory, hence, computing standard least squares estimators and t-test statistics on those observations. This gives

$$\underset{\substack{(\mathsf{se}_{LTS})\\ [t-stat_{LTS}]}}{log.light} = \underset{\substack{(2.09)\\ [-3.54]}}{-\,7.40} + \underset{\substack{(0.48)\\ [5.09]}}{2.80}\, log.Te, \qquad \hat{\sigma}_{LTS} = 0.3761. \tag{6.1}$$

The t-statistics are asymptotically normal under the above assumptions. These assumptions are plausible as the supergiants seem to be of the leverage type of Example 4.6.

Alternatively, we may apply SLTS inference. This assumes an i.i.d. model for all n observations with errors following an ϵ-contaminated normal distribution. The proportion of ‘good’ observations is $42/47 = 0.89$. Thus, if there was no contamination, which seems implausible, then the nuisance scaling factor would be $\varsigma^2_{0.89} = 0.608$. Correcting $\hat{\sigma}^2_{SLTS} = \hat{\sigma}^2_{LTS}/\varsigma^2_{0.89}$ and $\mathsf{se}^2_{SLTS} = \mathsf{se}^2_{LTS}/\varsigma^4_{0.89}$ gives

$$\begin{array}{cccc} log.light & = -\,7.40 & +\,2.80\,log.Te, & \qquad \hat{\sigma}_{SLTS} = 0.4821. \\ {\scriptstyle (\mathsf{se}_{SLTS})} & {\scriptstyle (3.43)} & {\scriptstyle (0.78)} & \\ {\scriptstyle [t\text{-}stat_{SLTS}]} & {\scriptstyle [-2.16]} & {\scriptstyle [3.59]} & \end{array} \tag{6.2}$$

We see that the estimated uncertainty is considerably larger with the SLTS inference than when using the LTS model as in (6.1).

7 Discussion

We have provided an asymptotic theory of LTS regression estimation allowing for leverage points. The theory shows that the LTS estimator is bounded and consistent with the same asymptotic expansion as the infeasible least squares estimator on the 'good' observations. Thus, the asymptotic distribution does not depend on the 'outliers'.

The LTS model will be appropriate for some data sets. For other data sets, ϵ-contamination – the SLTS model – could be attractive despite the nuisance parameters in the inference. The simulations indicate that SLTS inference is not appropriate under the LTS model and vice versa. Mis-specification and model selection tools are therefore needed. We suspect that many traditional methods can be applied directly to the estimated sets of 'good' observations, but their properties need to be investigated.

A Proofs

A.1 Boundedness

Proof of Theorem 3.1. We adapt the proof of Johansen and Nielsen (2019).

(*a*) *Overview.* We want to prove that $\max_{\zeta\in\mathcal{M}_n} |\hat{\beta}_\zeta - \hat{\beta}_{\zeta_n}| = \mathsf{O}_\mathsf{P}(1)$, where $\mathcal{M}_n$ is the set of minimizers. That is, $\forall\epsilon > 0$, $\exists B_0, n_0 > 0$, $\forall n > n_0$ and with $\mathcal{A}_n = (\max_{\zeta\in\mathcal{M}_n} |\hat{\beta}_\zeta - \hat{\beta}_{\zeta_n}| > B_0)$, then $\mathsf{P}(\mathcal{A}_n) < \epsilon$.

Defining the set $\mathcal{A}_{n\zeta} = (|\hat{\beta}_\zeta - \hat{\beta}_{\zeta_n}| > B_0)$, we can write $\mathcal{A}_n = \cup_{\zeta\in\mathcal{M}_n}\mathcal{A}_{n\zeta}$.

Any minimizer $\zeta \in \mathcal{M}_n$ has a residual variance satisfying $\hat{\sigma}^2_\zeta \leq \hat{\sigma}^2_{\zeta_n}$. Let $\mathcal{Z}_n$ be the set of all possible ζ and define the set $\mathcal{B}_{n\zeta} = (\hat{\sigma}^2_\zeta \leq \hat{\sigma}^2_{\zeta_n})$. Since $\mathcal{B}_{n\zeta}$ contains all minimizers, $\zeta \in \mathcal{M}_n$ and some non-minimizers, we get $\mathcal{A}_n \subset \cup_{\zeta\in\mathcal{Z}_n}(\mathcal{A}_{n\zeta} \cap \mathcal{B}_{n\zeta})$.

Given an $\epsilon > 0$, we will find a $B_0 > 0$ and sets $\mathbb{C}_n$ with probability $\mathsf{P}(\mathbb{C}_n) \geq 1 - \epsilon$. On $\mathbb{C}_n$, we will argue deterministically that if $|\hat{\beta}_\zeta - \hat{\beta}_{\zeta_n}| > B_0$ for some ζ then $\hat{\sigma}^2_\zeta \geq (1+\epsilon)\hat{\sigma}^2_{\zeta_n} > \hat{\sigma}^2_{\zeta_n}$. Thus, such a ζ cannot be a minimizer. Hence, on $\mathbb{C}_n$, the intersection, $\mathcal{A}_{n\zeta} \cap \mathcal{B}_{n\zeta}$ is empty. We get

$$\mathcal{A}_n \subset \cup_{\zeta\in\mathcal{Z}_n}(\mathcal{A}_{n\zeta} \cap \mathcal{B}_{n\zeta}) \subset \cup_{\zeta\in\mathcal{Z}_n}\{(\mathcal{A}_{n\zeta} \cap \mathcal{B}_{n\zeta} \cap \mathbb{C}_n) \cup (\mathcal{A}_{n\zeta} \cap \mathcal{B}_{n\zeta} \cap \mathbb{C}^c_n)\} \subset \mathbb{C}^c_n.$$

We can then bound $\mathcal{A}_n \subset \mathbb{C}^c_n$, so that $\mathsf{P}(\mathcal{A}_n) \leq \mathsf{P}(\mathbb{C}^c_n) < \epsilon$.

(*b*) *Criterion function.* Given a set ζ we find the least squares estimator

$$\hat{\beta}_\zeta = (\sum_{i\in\zeta} x_{in}x'_{in})^{-1}\sum_{i\in\zeta} x_{in}y_i = \beta + (\sum_{i\in\zeta} x_{in}x'_{in})^{-1}\sum_{i\in\zeta} x_{in}\varepsilon_i\sigma$$

using the model equation (2.1). The scaled residuals are $\tilde{\varepsilon}_{\zeta i} = (y_i - x'_{in}\hat{\beta}_\zeta)/\sigma$. For $\zeta = \zeta_n$ write $\tilde{\varepsilon}_i$ for $\tilde{\varepsilon}_{\zeta_n i}$. For general ζ write $\tilde{\varepsilon}_{\zeta i} = \varepsilon_i - x'_{in}(\hat{\beta}_\zeta - \beta)/\sigma$. Add and subtract $x'_{in}\hat{\beta}_{\zeta_n}/\sigma$ to get $\tilde{\varepsilon}_{\zeta i} = \varepsilon_i - x'_{in}(\hat{\beta}_{\zeta_n} - \beta)/\sigma - x'_{in}(\hat{\beta}_\zeta - \hat{\beta}_{\zeta_n})/\sigma$ and in turn $\tilde{\varepsilon}_{\zeta i} = \tilde{\varepsilon}_i - x'_{in}(\hat{\beta}_\zeta - \hat{\beta}_{\zeta_n})/\sigma$.

Introduce polar coordinates with length $\hat{\ell}_\zeta = |\hat{\beta}_\zeta - \hat{\beta}_{\zeta_n}|/\sigma$ and direction $\hat{\delta}_\zeta = (\hat{\beta}_\zeta - \hat{\beta}_{\zeta_n})/|\hat{\beta}_\zeta - \hat{\beta}_{\zeta_n}|$ when $\hat{\ell}_\zeta > 0$. When $\hat{\ell}_\zeta = 0$ the direction $\hat{\delta}_\zeta$ can be chosen as an arbitrary vector of unit length. Thus, $\hat{\beta}_\zeta - \hat{\beta}_{\zeta_n} = \hat{\ell}_\zeta \hat{\delta}_\zeta \sigma$ and $|\hat{\delta}_\zeta| = 1$. The residuals satisfy $\tilde{\varepsilon}_{\zeta i} = \tilde{\varepsilon}_i - \hat{\ell}_\zeta x'_{in}\hat{\delta}_\zeta$, so that

$$h\hat{\sigma}^2_\zeta = \sigma^2 \sum_{i\in\zeta} (\tilde{\varepsilon}_i - \hat{\ell}_\zeta x'_{in}\hat{\delta}_\zeta)^2. \tag{A.1}$$

(c) *Bounding residuals under constraints to ε_i, $x'_{in}\hat{\delta}_\zeta$ and $\hat{\ell}_\zeta$.* We will later choose $A_0, a_0, C_0 > 0$. Let $B_{n0} = (A_0 + C_0\hat{\sigma}_{\zeta_n}/\sigma)/a_0$. Consider $|\tilde{\varepsilon}_i| \le A_0$ and $|x'_{in}\hat{\delta}_\zeta| > a_0$ and $\hat{\ell}_\zeta > B_{n0}$. Then, by the reverse triangle inequality, $|x - y| \ge |(|x| - |y|)| \ge |y| - |x|$,

$$|\tilde{\varepsilon}_i - \hat{\ell}_\zeta x'_{in}\hat{\delta}_\zeta| \ge \hat{\ell}_\zeta |x'_{in}\hat{\delta}_\zeta| - |\tilde{\varepsilon}_i| > B_{n0}a_0 - A_0 \ge C_0\hat{\sigma}_{\zeta_n}/\sigma. \tag{A.2}$$

(d) *Bounding residual variance for large $\hat{\ell}_\zeta$.* Apply the expression for $\hat{\sigma}^2_\zeta$ in (A.1). Delete summands of $\hat{\sigma}^2_\zeta$ for which $|\tilde{\varepsilon}_i| > A_0$ or $|x'_{in}\hat{\delta}_\zeta| \le a_0$ and consider only values of ζ with large $\hat{\ell}_\zeta > B_{n0}$ to get the lower bound

$$h\hat{\sigma}^2_\zeta \ge 1_{(\hat{\ell}_\zeta > B_{n0})}\sigma^2 \sum_{i\in\zeta} (\tilde{\varepsilon}_i - \hat{\ell}_\zeta x'_{in}\hat{\delta}_\zeta)^2 1_{(|\tilde{\varepsilon}_i| \le A_0)} 1_{(|x'_{in}\hat{\delta}_\zeta| > a_0)}.$$

Now, for $\hat{\ell}_\zeta > B_{n0}$ we can apply (A.2) to get the further bound

$$h\hat{\sigma}^2_\zeta \ge 1_{(\hat{\ell}_\zeta > B_{n0})} C_0^2 \hat{\sigma}^2_{\zeta_n} \sum_{i\in\zeta} 1_{(|\tilde{\varepsilon}_i| \le A_0)} 1_{(|x'_{in}\hat{\delta}_\zeta| > a_0)}.$$

Use that for sets $\mathbb{A}$ and $\mathbb{B}$ then $1_{\mathbb{A}\cap\mathbb{B}} = 1_{\mathbb{A}} - 1_{\mathbb{A}\cap\mathbb{B}^c} \ge 1_{\mathbb{A}} - 1_{\mathbb{B}^c}$ so that

$$h\hat{\sigma}^2_\zeta \ge 1_{(\hat{\ell}_\zeta > B_{n0})} C_0^2 \hat{\sigma}^2_{\zeta_n} \Big\{ \sum_{i\in\zeta} 1_{(|\tilde{\varepsilon}_i| \le A_0)} - \sum_{i\in\zeta} 1_{(|x'_{in}\hat{\delta}_\zeta| \le a_0)} \Big\}. \tag{A.3}$$

For each sum in (A.3), we find bounds not depending on ζ. The first sum satisfies, noting that $1_{(|\tilde{\varepsilon}_i| \le A_0)} = 1 - 1_{(|\tilde{\varepsilon}_i| > A_0)}$,

$$\sum_{i\in\zeta} 1_{(|\tilde{\varepsilon}_i| \le A_0)} \ge \sum_{i\in\zeta\cap\zeta_n} 1_{(|\tilde{\varepsilon}_i| \le A_0)} = \#(\zeta \cap \zeta_n) - \sum_{i\in\zeta\cap\zeta_n} 1_{(|\tilde{\varepsilon}_i| > A_0)}.$$

Since ζ has at most $n - h$ indices in ζ_n^c, then $\#(\zeta \cap \zeta_n) \ge h - (n - h) = 2h - n$. Further, by summing over additional non-negative elements, we have that $\sum_{i\in\zeta\cap\zeta_n} 1_{(|\tilde{\varepsilon}_i| > A_0)} \le \sum_{i\in\zeta_n} 1_{(|\tilde{\varepsilon}_i| > A_0)}$. The inequality $1_{(|\tilde{\varepsilon}_i| > A_0)} \le \tilde{\varepsilon}_i^2/A_0^2$ gives the further bound $\sum_{i\in\zeta_n} \tilde{\varepsilon}_i^2/A_0^2$. Since $\tilde{\varepsilon}_i$ are the residuals from OLS regression on ζ_n, we get $\sum_{i\in\zeta_n} \tilde{\varepsilon}_i^2 \le \sum_{i\in\zeta_n} \varepsilon_i^2$. Thus, the first sum in (A.3) satisfies $\sum_{i\in\zeta} 1_{(|\tilde{\varepsilon}_i| \le A_0)} \ge 2h - n - \sum_{i\in\zeta_n} \varepsilon_i^2/A_0^2$.

For the second sum in (A.3), replace $\hat{\delta}_\zeta$ by an arbitrary δ, take supremum over δ and take maximum over sets ζ of length h to get the bound

$$\sum_{i\in\zeta} 1_{(|x_{in}'\hat{\delta}_\zeta|\le a_0)} \le \max_{\zeta:\#\zeta=h} \sup_{|\delta|=1} \sum_{i\in\zeta} 1_{(|x_{in}'\delta|\le a_0)} = hF_{nh}(a_0).$$

Insert the bounds in (A.3) to get, uniformly in ζ satisfying $\hat{\ell}_\zeta > B_{n0}$, that

$$\hat{\sigma}_\zeta^2 \ge 1_{(\hat{\ell}_\zeta > B_{n0})} C_0^2 \hat{\sigma}_{\zeta_n}^2 \Big\{ \frac{2h-n}{h} - A_0^{-2} \frac{1}{h} \sum_{i\in\zeta_n} \varepsilon_i^2 - F_{nh}(a_0) \Big\}. \tag{A.4}$$

(*e*) *Probability argument.* We construct sets $\mathbb{C}_n$ with large probability.
Assumption 3.1(*i*) has $h/n \to \lambda > 1/2$. Thus, $(2h-n)/h \to 2-\lambda^{-1} > 0$.
Assumption 3.1(*ii*) states that $h^{-1}\sum_{i\in\zeta_n} \varepsilon_i^2 = \mathrm{O}_\mathsf{P}(1)$. This implies that $\hat{\sigma}_{\zeta_n}^2/\sigma^2 = \mathrm{O}_\mathsf{P}(1)$, see Remark 3.1.
Assumption 3.1(*iii*) states that $\lim_{(a,n)\to(0,\infty)} \mathsf{P}\{F_{nh}(a) \ge \xi\} = 0$ for some $\xi < 2-\lambda^{-1}$.

These assumptions show that for all $\epsilon > 0$ there exists $a_0, A_0, n_0 > 0$ and sets $\mathbb{C}_n$ with $\mathsf{P}(\mathbb{C}_n) \ge 1-\epsilon$ for all $n > n_0$ so that on $\mathbb{C}_n$ we have

$$\frac{1}{h}\sum_{i\in\zeta_n} \varepsilon_i^2 \le \epsilon A_0^2 \qquad \text{and} \qquad \hat{\sigma}_{\zeta_n}^2/\sigma^2 \le A_0 \qquad \text{and} \qquad F_{nh}(a_0) \le \xi.$$

Now, choose $C_0^2 = (1+\epsilon)/(2-\lambda^{-1}-2\epsilon-\xi)$, noting that $C_0^2 > 0$ for small ϵ since $\xi < 2-\lambda^{-1}$. Let $B_0 = (A_0 + C_0 A_0)/a_0$ so that $B_0 \ge B_{n0}$ on $\mathbb{C}_n$.

(*f*) *Bound residual variance on* $\mathbb{C}_n$. As argued in (*a*), consider any ζ with $\hat{\ell}_\zeta = |\hat{\beta}_\zeta - \hat{\beta}_{\zeta_n}| > B_0 \ge B_{n0}$. Apply the constraints defining $\mathbb{C}_n$ to the lower bound for $\hat{\sigma}_\zeta^2$ in (A.4) to get the bound

$$\hat{\sigma}_\zeta^2 \ge C_0^2 \hat{\sigma}_{\zeta_n}^2 \{(2-\lambda^{-1}-\epsilon) - \epsilon - \xi\} = (1+\epsilon)\hat{\sigma}_{\zeta_n}^2$$

on the set $\mathbb{C}_n$. Thus, this ζ cannot be a minimizer since minimizers satisfy $\hat{\sigma}_\zeta^2 \le \hat{\sigma}_{\zeta_n}^2$. This is what had to be proved as outlined in item (*a*). □

A.2 Consistent selection of 'good' observations

For a matrix m let $\|m\|$ be the spectral norm. Thus, $\|m\|^2 = \max \mathrm{eigen}(m'm)$. If the matrices m_1, m_2 are conformable then $\|m_1 m_2\| \le \|m_1\|\|m_2\|$.

Proof of Theorem 3.2. We note that for any minimizer, $\zeta \in \mathcal{M}_n$, then $\hat{\sigma}_\zeta^2 \le \hat{\sigma}_{\zeta_n}^2$.

We construct a high probability set $\mathbb{D}_n$, where we can deterministically bound certain statistics. Assumptions 3.1, 3.2(*iii*) along with Remark 3.1 and Theorem 3.1 show that $\max_{\zeta\in\mathcal{M}_n} |\hat{\beta}_\zeta|$ and $\hat{\sigma}_{\zeta_n}^2$ are $\mathrm{O}_\mathsf{P}(1)$. Assumption 3.2(*ii*) is that $\|\sum_{i=1}^n x_{in}x_{in}'\| = \mathrm{O}_\mathsf{P}(n) = \mathrm{O}_\mathsf{P}(h)$. Thus, for all $\epsilon > 0$ there exist $C, n_0 > 0$ and a sequence of sets $\mathbb{D}_n$ with $\mathsf{P}(\mathbb{D}_n) > 1-\epsilon$ for all $n > n_0$, so that on $\mathbb{D}_n$

$$\max_{\zeta\in\mathcal{M}_n} |\hat{\beta}_\zeta - \beta|/\sigma \le C, \quad \hat{\sigma}_{\zeta_n}^2/\sigma^2 \le C, \quad \|\sum_{i=1}^n x_{in}x_{in}'\| \le Ch. \tag{A.5}$$

For a minimizer ζ, we expand the least squares residual variance as

$$h\hat{\sigma}^2_\zeta = \sigma^2 \sum_{i\in\zeta} \varepsilon_i^2 - (\hat{\beta}_\zeta - \beta)'(\sum_{i\in\zeta} x_{in}x_{in}')(\hat{\beta}_\zeta - \beta). \tag{A.6}$$

The first term satisfies $\sum_{i\in\zeta} \varepsilon_i^2 \geq \sum_{i\in\zeta\cap\zeta_n^c} \varepsilon_i^2 \geq \#(\zeta \cap \zeta_n^c) m_n^2 \{1 + \mathrm{o_P}(1)\}$ by Assumption 3.2(i). For the second term, $\| \sum_{i\in\zeta} x_{in}x_{in}' \| \leq \| \sum_{i=1}^n x_{in}x_{in}' \| \leq Ch$ and $|\hat{\beta}_\zeta - \beta| \leq C\sigma$ while $C\sigma^2 \geq \hat{\sigma}^2_{\zeta_n} \geq \hat{\sigma}^2_\zeta$ on the set $\mathbb{D}_n$. Thus, we get

$$hC \geq h\hat{\sigma}^2_\zeta/\sigma^2 \geq \#(\zeta \cap \zeta_n^c) m_n^2 \{1 + \mathrm{o_P}(1)\} - C^3 h. \tag{A.7}$$

On $\mathbb{D}_n$, solve to get $\#(\zeta \cap \zeta_n^c) \leq (C + C^3)(h/m_n^2)\{1 + \mathrm{o_P}(1)\}$, uniformly in $\zeta \in \mathcal{M}_n$. □

A.3 Improving the rate of consistency

When improving the consistency rate we will bound terms like $\sum_{i\in\zeta\cap\zeta_n^c} x_{in}x_{in}'$. The bounds should be uniform in ζ where the number of misclassifications, $\#(\zeta \cap \zeta_n^c)$, is bounded by some sequence g_n. Thus, everywhere, $g_n > 0$ is a sequence in n not depending on ζ. When applying the bounds we will first choose $g_n = Ch/m_n^2$ and later $g_n = Ch^\theta$. The bounds will expressed in terms of

$$\mathcal{S}_{g_n} = \sum_{i>n-g_n}^n \phi_i,$$

where $\phi_1, \ldots, \phi_n$ are non-decreasing order statistics of $x_{in}'(\sum_{i\in\zeta_n} x_{in}x_{in}')^{-1}x_{in}$ as introduced in Assumption 3.3(ii).

Lemma A.1. *Suppose Assumption 3.3(id, ii). Then $\forall 0 < C < \infty$ and $g_n = Ch/m_n^2$ we get $S_{g_n} = \mathrm{o_P}(1)$. Also, $\forall 0 < C < \infty$, $\forall \theta > 0$ and $g_n = Ch^\theta$ we get $S_{g_n} = \mathrm{O_P}(n^{\eta+\theta-1})$.*

Remark A.1. *For Lemma A.1 and hence for Theorem 3.3 it suffices that $0 < \eta < 1$ in Assumption 3.3(id, ii).*

Proof of Lemma A.1. The case $g_n = Ch/m_n^2$. Given a $\delta > 0$ choose $0 < r < 1 - \eta$ so that, by Assumption 3.3(iia), $\phi_{n-\lfloor n^r \rfloor}/\phi_n \leq \delta + \mathrm{o_P}(1)$. If $\lfloor n^r \rfloor < g_n$, decompose

$$\mathcal{S}_{g_n} = \sum_{i>n-g_n}^n \phi_i = \sum_{i>n-g_n}^{n-\lfloor n^r \rfloor} \phi_i + \sum_{i=n-\lfloor n^r \rfloor+1}^n \phi_i.$$

The sums have summands bounded by $\phi_{n-\lfloor n^r \rfloor}$ and ϕ_n, while their number of summands are bounded by g_n and n^r, respectively. Thus, we can bound

$$\mathcal{S}_{g_n} \leq g_n \phi_{n-\lfloor n^r \rfloor} + n^r \phi_n. \tag{A.8}$$

The first term of (A.8) has $g_n = Ch/m_n^2$ by construction. Moreover, as noted above, $\phi_{n-\lfloor n^r \rfloor}/\phi_n \leq \delta + \mathrm{o_P}(1)$ where $\phi_n = \mathrm{O_P}(m_n^2/h)$ by Assumption 3.3(iib). Thus, the first term is $\mathrm{O_P}(\delta)$. Since $\delta > 0$ can be chosen arbitrarily small, it is actually $\mathrm{o_P}(1)$.

For the second term of (A.8) use again that $\phi_n = \mathrm{O}_{\mathsf{P}}(m_n^2/h)$ while $m_n^2 = \mathrm{o}_{\mathsf{P}}(n^\eta)$ for $0 < \eta < 1/2$ by Assumption 3.3(id). Since $h/n \to \lambda$, the second term is $\mathrm{O}_{\mathsf{P}}(h^{r+\eta-1}) = \mathrm{o}_{\mathsf{P}}(1)$ as $r < 1-\eta$. Then $\mathcal{S}_{g_n} = \mathrm{o}_{\mathsf{P}}(1)$.

If $\lfloor n^r \rfloor \geq g_n$, then $\mathcal{S}_{g_n} \leq \sum_{i>n-\lfloor n^r \rfloor}^n \phi_i \leq n^r \phi_n = \mathrm{O}_{\mathsf{P}}(h^{r+\eta-1}) = \mathrm{o}_{\mathsf{P}}(1)$.

The case $g_n = Ch^\theta$. Recall $\phi_n = \mathrm{O}_{\mathsf{P}}(m_n^2/h) = \mathrm{O}_{\mathsf{P}}(h^{\eta-1})$ by Assumption 3.3(iia). Thus, we bound $\mathcal{S}_{g_n} = \sum_{i>n-g_n}^n \phi_i \leq (g_n+1)\phi_n = \mathrm{O}_{\mathsf{P}}(h^\theta)\mathrm{O}_{\mathsf{P}}(h^{\eta-1}) = \mathrm{O}_{\mathsf{P}}(h^{\theta+\eta-1})$. □

The following notation is convenient. Let

$$\begin{aligned}
z_{jn} &= (\sum_{i\in\zeta_n} x_{in}x_{in}')^{-1/2}x_{jn},\\
A_\zeta &= (\sum_{i\in\zeta} x_{in}x_{in}')^{1/2}(\hat\beta_\zeta - \beta)/\sigma = (\sum_{i\in\zeta} x_{in}x_{in}')^{-1/2}\sum_{i\in\zeta} x_{in}\varepsilon_i,\\
B_\zeta &= \sum_{i\in\zeta} z_{in}\varepsilon_i - \sum_{i\in\zeta_n} z_{in}\varepsilon_i \qquad\qquad (A.9)\\
&= C_\zeta = (\sum_{i\in\zeta} x_{in}x_{in}')^{-1/2}(\sum_{i\in\zeta_n} x_{in}x_{in}')^{1/2} - I_{\dim x}.
\end{aligned}$$

so that $A_{\zeta_n} = \sum_{i\in\zeta_n} z_{in}\varepsilon_i$.

Lemma A.2. *The squared difference* $|A_\zeta - A_{\zeta_n}|^2$ *can be bounded as follows*

$$\frac{1}{3}|A_\zeta - A_{\zeta_n}|^2 \leq |B_\zeta|^2(1+\|C_\zeta\|^2) + \|C_\zeta\|^2|\sum_{i\in\zeta_n} z_{in}\varepsilon_i|^2. \qquad (A.10)$$

Proof of Lemma A.2. By definition

$$A_\zeta - A_{\zeta_n} = (\sum_{i\in\zeta} x_{in}x_{in}')^{-1/2}(\sum_{i\in\zeta_n} x_{in}x_{in}')^{1/2}(\sum_{i\in\zeta} z_{in}\varepsilon_i') - (\sum_{i\in\zeta_n} z_{in}\varepsilon_i').$$

Rewrite as $A_\zeta - A_{\zeta_n} = B_\zeta + C_\zeta B_\zeta + C_\zeta(\sum_{i\in\zeta_n} z_{in}\varepsilon_i')$. The triangle and Jensen's inequalities and the spectral norm sub-multiplicativity give the desired result. □

Lemma A.3. *Let* $\#(\zeta\cap\zeta_n^c) \leq g_n$. *Then* $\sum_{i\in\zeta\cap\zeta_n^c} z_{in}'z_{in}$, $\sum_{i\in\zeta^c\cap\zeta_n} z_{in}'z_{in}$ *are at most* $\mathcal{S}_{g_n}$.

Proof of Lemma A.3. By definition $z_{in}'z_{in} = x_{in}'(\sum_{i\in\zeta_n} x_{in}x_{in}')^{-1}x_{in}$. As remarked in Section 3.2, we have $\#(\zeta\cap\zeta_n^c) = \#(\zeta^c\cap\zeta_n)$. Since ϕ_i are the increasing order statistics of $z_{in}'z_{in}$ and $\#(\zeta\cap\zeta_n^c) \leq g_n$ both sums are bounded by $\mathcal{S}_{g_n}$. □

Lemma A.4. *Let* $\#(\zeta\cap\zeta_n^c) \leq g_n$. *The term* $|B_\zeta|^2$ *can be bounded by*

$$|B_\zeta|^2 \leq 2\Big(\sum_{i\in\zeta\cap\zeta_n^c}\varepsilon_i^2 + \sum_{i\in\zeta^c\cap\zeta_n}\varepsilon_i^2\Big)\mathcal{S}_{g_n}.$$

Proof of Lemma A.4. Decompose B_ζ along the lines of (A.11) to get

$$B_\zeta = \sum_{i\in\zeta} z_i\varepsilon_i - \sum_{i\in\zeta_n} z_i\varepsilon_i = \sum_{i\in\zeta\cap\zeta_n^c} z_{in}\varepsilon_i - \sum_{i\in\zeta^c\cap\zeta_n} z_{in}\varepsilon_i.$$

Apply the triangle, Jensen and Cauchy-Schwarz inequalities to get

$$\begin{aligned}|B_\zeta|^2 &\leq \Big(\sum_{i\in\zeta\cap\zeta_n^c} |z_{in}\varepsilon_i| + \sum_{i\in\zeta^c\cap\zeta_n} |z_{in}\varepsilon_i| \Big)^2 \\ &\leq 2\Big\{ \Big(\sum_{i\in\zeta\cap\zeta_n^c} |z_{in}\varepsilon_i| \Big)^2 + \Big(\sum_{i\in\zeta^c\cap\zeta_n} |z_{in}\varepsilon_i| \Big)^2 \Big\} \\ &\leq 2\Big(\sum_{i\in\zeta\cap\zeta_n^c} |z_{in}|^2 \sum_{i\in\zeta\cap\zeta_n^c} \varepsilon_i^2 + \sum_{i\in\zeta^c\cap\zeta_n} |z_{in}|^2 \sum_{i\in\zeta^c\cap\zeta_n} \varepsilon_i^2 \Big).\end{aligned}$$

Lemma A.3 bounds $\sum_{i\in\zeta\cap\zeta_n^c} |z_{in}|^2$ and $\sum_{i\in\zeta^c\cap\zeta_n} |z_{in}|^2$ by $\mathcal{S}_{g_n}$. □

Lemma A.5. *Let* $M_\zeta = (\sum_{i\in\zeta_n} x_{in}x_{in}')^{-1/2} \sum_{i\in\zeta} x_{in}x_{in}' (\sum_{i\in\zeta_n} x_{in}x_{in}')^{-1/2}$ *and* $\#(\zeta \cap \zeta_n^c) \leq g_n$. *Then,* $\|M_\zeta - I_{\dim x}\| \leq 2\mathcal{S}_{g_n}$.

Proof of Lemma A.5. Since $z_{jn} = (\sum_{i\in\zeta_n} x_{in}x_{in}')^{-1/2} x_{jn}$, we get $M_\zeta = \sum_{i\in\zeta} z_{in}z_{in}'$. Write

$$M_\zeta = \sum_{i\in\zeta_n} z_{in}z_{in}' + (\sum_{i\in\zeta} z_{in}z_{in}' - \sum_{i\in\zeta n} z_{in}z_{in}'),$$

and note that the first sum satisfies $\sum_{i\in\zeta_n} z_{in}z_{in}' = I_{\dim x}$ while, in the last two sums, we can cancel elements with index in $\zeta \cap \zeta_n$. Hence,

$$M_\zeta = I_{\dim x} + \sum_{i\in\zeta\cap\zeta_n^c} z_{in}z_{in}' - \sum_{i\in\zeta^c\cap\zeta_n} z_{in}z_{in}'. \tag{A.11}$$

Use the spectral norm and the triangle inquality to get that

$$\|M_\zeta - I_{\dim x}\| \leq \sum_{i\in\zeta\cap\zeta_n^c} \|z_{in}z_{in}'\| + \sum_{i\in\zeta^c\cap\zeta_n} \|z_{in}z_{in}'\| = \sum_{i\in\zeta\cap\zeta_n^c} z_{in}'z_{in} + \sum_{i\in\zeta^c\cap\zeta_n} z_{in}'z_{in}.$$

By Lemma A.3, each of the sums is bounded by $\mathcal{S}_{g_n}$. The desired bound follows. □

Lemma A.6. *If* $\mathcal{S}_{g_n} = \mathrm{o}_{\mathsf{P}}(1)$ *then* $\max_{\zeta:\#(\zeta\cap\zeta_n^c)\leq g_n} \|C_\zeta\| = \mathrm{O}_{\mathsf{P}}(\mathcal{S}_{g_n})$.

Proof of Lemma A.6. Let $M_\zeta = (\sum_{i\in\zeta_n} x_{in}x_{in}')^{-1/2} \sum_{i\in\zeta} x_{in}x_{in}' (\sum_{i\in\zeta_n} x_{in}x_{in}')^{-1/2}$ as before. We express C_ζ in terms of M_ζ as follows

$$C_\zeta = \Big\{ \Big(\sum_{i\in\zeta_n} x_{in}x_{in}' \Big)^{1/2} M_\zeta \Big(\sum_{i\in\zeta_n} x_{in}x_{in}' \Big)^{1/2} \Big\}^{-1/2} \Big(\sum_{i\in\zeta_n} x_{in}x_{in}' \Big)^{1/2} - I.$$

Expanding the curly bracket, we get

$$C_\zeta = \Big(\sum_{i\in\zeta_n} x_{in}x_{in}' \Big)^{-1/4} M_\zeta^{-1/2} \Big(\sum_{i\in\zeta_n} x_{in}x_{in}' \Big)^{1/4} - I. \tag{A.12}$$

By Lemma A.5, $\|M_\zeta - I_{\dim x}\| \leq 2\mathcal{S}_{g_n}$. Write $M_\zeta = I_{\dim x} + (M_\zeta - I_{\dim x})$ so

$$M_\zeta = I_{\dim x}\{1 + \mathrm{O}(\|M_\zeta - I_{\dim x}\|)\} = I_{\dim x}\{1 + \mathrm{O}(\mathcal{S}_{g_n})\}.$$

Since $\mathcal{S}_{g_n} = \mathrm{o}_{\mathsf{P}}(1)$, we get $M_\zeta^{-1/2} = I_{\dim x}\{1 + \mathrm{O}_{\mathsf{P}}(\mathcal{S}_{g_n})\}$. Inserting into (A.12) gives $C_\zeta = I_{\dim x}\mathrm{O}_{\mathsf{P}}(\mathcal{S}_{g_n})$, so that $\max_\zeta \|C_\zeta\| = \mathrm{O}_{\mathsf{P}}(\mathcal{S}_{g_n})$ as desired. □

Lemma A.7. *Suppose Assumption 3.3(iii). Let $\mathcal{S}_{g_n} = \mathrm{o}_{\mathsf{P}}(1)$. Then,*

$$|A_\zeta - A_{\zeta_n}|^2 \leq \Big(\sum_{i \in \zeta \cap \zeta_n^c} \varepsilon_i^2 + \sum_{i \in \zeta^c \cap \zeta_n} \varepsilon_i^2 \Big) \mathrm{O}_{\mathsf{P}}(\mathcal{S}_{g_n}) + \mathrm{o}_{\mathsf{P}}(1),$$

where the remainder terms are uniform in ζ so that $\#(\zeta \cap \zeta_n^c) \leq g_n$.

Proof of Lemma A.7. By Lemma A.2,

$$|A_\zeta - A_{\zeta_n}|^2 \leq 3|B_\zeta|^2\{1 + \|C_\zeta\|^2\} + 3\|C_\zeta\|^2 |\sum_{i \in \zeta_n} z_{in}\varepsilon_i|^2. \tag{A.13}$$

By Assumption 3.3(iii), $\sum_{i \in \zeta_n} z_{in}\varepsilon_i' = \mathrm{O}_{\mathsf{P}}(1)$. By Lemma A.6 using the Assumption that $\mathcal{S}_{g_n} = \mathrm{o}_{\mathsf{P}}(1)$, we get $\|C_\zeta\| = \mathrm{O}_{\mathsf{P}}(\mathcal{S}_{g_n}) = \mathrm{o}_{\mathsf{P}}(1)$ uniformly in ζ. By Lemma A.4, $|B_\zeta|^2 \leq 2(\sum_{i \in \zeta \cap \zeta_n^c} \varepsilon_i^2 + \sum_{i \in \zeta^c \cap \zeta_n} \varepsilon_i^2)\mathcal{S}_{g_n}$. Insert these results into (A.13). □

Lemma A.8. *Suppose Assumption 3.3(iii). Let $\mathcal{S}_{g_n} = \mathrm{o}_{\mathsf{P}}(1)$. Then,*

$$h(\hat{\sigma}_\zeta^2 - \hat{\sigma}_{\zeta_n}^2)/\sigma \geq \{1 + \mathrm{o}_{\mathsf{P}}(1)\} \sum_{i \in \zeta \cap \zeta_n^c} \varepsilon_i^2 - \{1 + \mathrm{o}_{\mathsf{P}}(1)\} \sum_{i \in \zeta^c \cap \zeta_n} \varepsilon_i^2 + \mathrm{O}_{\mathsf{P}}(1),$$

where all remainder terms are uniform in ζ so that $\#(\zeta \cap \zeta_n^c) \leq g_n$.

Proof of Lemma A.8. Write

$$Q_\zeta \ = h(\hat{\sigma}_\zeta^2 - \hat{\sigma}_{\zeta_n}^2)/\sigma = \sum_{i \in \zeta} \varepsilon_i^2 - A_\zeta' A_\zeta - \sum_{i \in \zeta_n} \varepsilon_i^2 + A_{\zeta_n}' A_{\zeta_n}. \tag{A.14}$$

Manipulate the sums as in (A.11) and note $A_{\zeta_n}' A_{\zeta_n} \geq 0$ to bound

$$Q_\zeta \geq \sum_{i \in \zeta \cap \zeta_n^c} \varepsilon_i^2 - \sum_{i \in \zeta^c \cap \zeta_n} \varepsilon_i^2 - A_\zeta' A_\zeta. \tag{A.15}$$

Write $A_\zeta = A_{\zeta_n} + (A_\zeta - A_{\zeta_n})$ to get

$$A_\zeta' A_\zeta \leq 2\{A_{\zeta_n}' A_{\zeta_n} + (A_\zeta - A_{\zeta_n})'(A_\zeta - A_{\zeta_n})\} = 2A_{\zeta_n}' A_{\zeta_n} + 2|A_\zeta - A_{\zeta_n}|^2. \tag{A.16}$$

Assumption 3.3(iii) has $A_{\zeta_n}' A_{\zeta_n} = \mathrm{O}_{\mathsf{P}}(1)$, while Lemma A.7 using the assumption $\mathcal{S}_{g_n} = \mathrm{o}_{\mathsf{P}}(1)$ uniformly in ζ bounds $|A_\zeta - A_{\zeta_n}|^2 \leq (\sum_{i \in \zeta \cap \zeta_n^c} \varepsilon_i^2 + \sum_{i \in \zeta^c \cap \zeta_n} \varepsilon_i^2)\mathrm{o}_{\mathsf{P}}(1) + \mathrm{o}_{\mathsf{P}}(1)$, where the remainders are uniform in ζ. Therefore, the bound (A.16) becomes

$$A_\zeta' A_\zeta \leq \mathrm{O}_{\mathsf{P}}(1) + \Big(\sum_{i \in \zeta \cap \zeta_n^c} \varepsilon_i^2 + \sum_{i \in \zeta^c \cap \zeta_n} \varepsilon_i^2 \Big) \mathrm{o}_{\mathsf{P}}(1) + \mathrm{o}_{\mathsf{P}}(1). \tag{A.17}$$

Insert (A.17) in (A.15) to get the desired result. □

Lemma A.9. *Suppose Assumptions 3.2(i), 3.3 with $0 < \eta < 1$. Then, $\forall C > 0$, $0 < \theta < 1 - \eta$, we have that $\min_{\zeta : h^\theta \leq \#(\zeta \cap \zeta_n^c) \leq hC/m_n^2} h^{1-\theta}(\hat{\sigma}_\zeta^2 - \hat{\sigma}_{\zeta_n}^2) \to \infty$ in probability.*

Proof of Lemma A.9. Let # be short hand for $\#(\zeta^c \cap \zeta_n) = \#(\zeta \cap \zeta_n^c)$.

We consider $h^\theta \leq \# \leq g_n$ where $g_n = hC/m_n^2$. We have that $\mathcal{S}_{g_n} = \mathrm{o}_{\mathsf{P}}(1)$, by Lemma A.1 using Assumption 3.3(id, ii). Thus, Lemma A.8 using Assumption 3.3(iii) shows

$$h(\hat{\sigma}_\zeta^2 - \hat{\sigma}_{\zeta_n}^2)/\sigma \geq \{1 + \mathrm{o}_{\mathsf{P}}(1)\} \sum_{i \in \zeta \cap \zeta_n^c} \varepsilon_i^2 - \{1 + \mathrm{o}_{\mathsf{P}}(1)\} \sum_{i \in \zeta^c \cap \zeta_n} \varepsilon_i^2 + \mathrm{O}_{\mathsf{P}}(1), \tag{A.18}$$

where all remainder terms are uniform in ζ. We show that the lower bound diverges.

The first sum in (A.18) relates to 'outliers', which satisfy $\varepsilon_j^2 \geq m_n^2\{1 + \mathrm{o}_{\mathsf{P}}(1)\}$ for $j \notin \zeta_n$ by Assumption 3.2(i). Thus, $\sum_{i \in \zeta \cap \zeta_n^c} \varepsilon_i^2 \geq m_n^2 \#\{1 + \mathrm{o}_{\mathsf{P}}(1)\}$.

The second sum in (A.18) relates to 'good' errors. Let $\psi_1 \leq \cdots \leq \psi_h$ be the order statistics of ε_i^2 for $i \in \zeta_n$. Given $\theta > 0$ choose $0 < \rho < \theta$. Since $\lfloor h^\rho \rfloor < \#$, then

$$\sum_{i \in \zeta^c \cap \zeta_n} \varepsilon_i^2 \leq \sum_{i=h+1-\#}^{h} \psi_i = \sum_{i=h+1-\#}^{h-\lfloor h^\rho \rfloor} \psi_i + \sum_{i=h-\lfloor h^\rho \rfloor+1}^{h} \psi_i \leq \# \psi_{h-\lfloor h^\rho \rfloor} + h^\rho \psi_h.$$

For the first term, Assumption 3.3(ic) shows a $C_\rho < 1$ exists so that $\psi_{h-\lfloor h^\rho \rfloor}/m_n^2 \leq C_\rho + \mathrm{o}_{\mathsf{P}}(1)$. Thus, the first term is bounded by $m_n^2 \#\{C_\rho + \mathrm{o}_{\mathsf{P}}(1)\}$.

For the second term, we have $\rho < \theta$ so that $h^\rho = \mathrm{o}(h^\theta)$ while $h^\theta \leq \#$ by construction. Further, Assumption 3.3(ib) shows $\psi_h/m_n^2 = \mathrm{O}_{\mathsf{P}}(1)$. Thus, the second term is bounded by $m_n^2 \# \mathrm{o}_{\mathsf{P}}(1)$. Overall, we get $\sum_{i \in \zeta^c \cap \zeta_n} \varepsilon_i^2 \leq m_n^2 \#\{C_\rho + \mathrm{o}_{\mathsf{P}}(1)\}$.

Inserting the above bounds in (A.18), we find

$$h(\hat{\sigma}_\zeta^2 - \hat{\sigma}_{\zeta_n}^2)/\sigma \geq m_n^2 \#(1 - C_\rho)\{1 + \mathrm{o}_{\mathsf{P}}(1)\}.$$

Since m_n^2 diverges due to Assumption 3.3(ia), $\# \geq h^\theta$ and $C_\rho < 1$, then $h^{1-\theta}(\hat{\sigma}_\zeta^2 - \hat{\sigma}_{\zeta_n}^2) \to \infty$ in probability. □

Proof of Theorem 3.3. First, Theorem 3.2 using Assumptions 3.1 and 3.2, shows that $\max_{\zeta \in \mathcal{M}_n} \#(\zeta \cap \zeta_n^c) = \mathrm{O}_{\mathsf{P}}(h/m_n^2)$.

Second, Lemma A.9 using Assumptions 3.2(i), 3.3, considers estimators $\hat{\sigma}_\zeta^2$ for index sets ζ that contain a positive number of 'outliers', in the range, $h^\theta \leq \#(\zeta \cap \zeta_n^c) \leq Ch/m_n^2$ for any $C > 0$, $0 < \theta < 1 - \eta$. This set of ζ does not include the true set of 'good' observations, ζ_n. Lemma A.9 states that $h^{1-\theta}(\hat{\sigma}_\zeta^2 - \hat{\sigma}_{\zeta_n}^2)$ diverges to positive infinity uniformly in values of ζ in the set. Since the function $(\hat{\sigma}_\zeta^2 - \hat{\sigma}_{\zeta_n}^2)$ is zero at ζ_n, the considered set of ζ values cannot contain a minimizer in the limit.

In combination, all minimizers, $\zeta \in \mathcal{M}_n$, satisfy $\max_{\zeta \in \mathcal{M}_n} \#(\zeta \cap \zeta_n^c) = \mathrm{O}_{\mathsf{P}}(h^\theta)$. □

A.4 Main result

Proof of Theorem 3.4. (a) Theorem 3.3, using the Assumptions 3.1, 3.2, 3.3, gives that $\#(\zeta \cap \zeta_n^c) = \mathrm{O}_{\mathsf{P}}(h^\theta)$ for any $0 < \theta < 1$. Hence, consider minimizers ζ so that $\#(\zeta \cap \zeta_n^c) \leq g_n = Ch^\theta$. Then, by Lemma A.1 using Assumption 3.3(id, ii), we have that $\mathcal{S}_{g_n} = \mathrm{O}_{\mathsf{P}}(h^{\theta+\eta-1})$. Since $\eta < 1/2$ and $\theta > 0$ is arbitrary, we have $\mathcal{S}_{g_n} = \mathrm{o}_{\mathsf{P}}(1)$.

For any minimizer $\hat{\sigma}_\zeta^2 - \hat{\sigma}_{\zeta_n}^2 \leq 0$. Thus we need to show that $\hat{\sigma}_\zeta^2 - \hat{\sigma}_{\zeta_n}^2 \geq -\epsilon h^{-1/2}$ with large probability for any small $\epsilon > 0$. Lemma A.8, using Assumption 3.3(id, ii, iii)

and the fact that $\mathcal{S}_{g_n} = \mathsf{o}_\mathsf{P}(1)$, gives the lower bound

$$h(\hat{\sigma}_\zeta^2 - \hat{\sigma}_{\zeta_n}^2) \geq \{1 + \mathsf{o}_\mathsf{P}(1)\} \sum_{i \in \zeta \cap \zeta_n^c} \varepsilon_i^2 - \{1 + \mathsf{o}_\mathsf{P}(1)\} \sum_{i \in \zeta^c \cap \zeta_n} \varepsilon_i^2 + \mathrm{O}_\mathsf{P}(1), \tag{A.19}$$

where all remainder terms are uniform in ζ. First, we bound $\sum_{i \in \zeta \cap \zeta_n^c} \varepsilon_i^2 \geq 0$. Second,

$$\sum_{i \in \zeta^c \cap \zeta_n} \varepsilon_i^2 \leq (\max_{i \in \zeta_n} \varepsilon_i^2)\{\#(\zeta^c \cap \zeta_n)\} = \mathsf{o}_\mathsf{P}(h^{\theta+\eta}), \tag{A.20}$$

uniformly in ζ, as $\max_{i \in \zeta_n} \varepsilon_i^2 = \mathrm{O}_\mathsf{P}(m_n^2)$ and $m_n^2 = \mathsf{o}_\mathsf{P}(n^\eta)$ by Assumptions 3.3(ib, id), while $\#(\zeta^c \cap \zeta_n) \leq Ch^\theta$. Thus, for $\eta < 1/2$ and small $\theta > 0$, we get, with large probability that $0 \geq \hat{\sigma}_\zeta^2 - \hat{\sigma}_{\zeta_n}^2 \geq -\epsilon h^{\theta+\eta-1} > -\epsilon h^{-1/2}$.

(b) We show for all ζ so that $\#(\zeta \cap \zeta_n^c) \leq Ch^\theta$ that

$$\mathcal{D}_\zeta = |(\sum_{i \in \zeta} x_{in} x_{in}')^{1/2}(\hat{\beta}_\zeta - \beta) - (\sum_{i \in \zeta_n} x_{in} x_{in}')^{1/2}(\hat{\beta}_{\zeta_n} - \beta)| = \mathsf{o}_\mathsf{P}(1).$$

Lemma A.7 using Assumption 3.3(iii) gives that

$$\mathcal{D}_\zeta^2 = |A_\zeta - A_{\zeta_n}|^2 \leq \Big(\sum_{i \in \zeta \cap \zeta_n^c} \varepsilon_i^2 + \sum_{i \in \zeta^c \cap \zeta_n} \varepsilon_i^2\Big) \mathrm{O}_\mathsf{P}(\mathcal{S}_{g_n}) + \mathsf{o}_\mathsf{P}(1). \tag{A.21}$$

For the first sum, we use (A.19) in part (a) to bound

$$\{1 + \mathsf{o}_\mathsf{P}(1)\} \sum_{i \in \zeta \cap \zeta_n^c} \varepsilon_i^2 \leq h(\hat{\sigma}_\zeta^2 - \hat{\sigma}_{\zeta_n}^2) + \{1 + \mathsf{o}_\mathsf{P}(1)\} \sum_{i \in \zeta^c \cap \zeta_n} \varepsilon_i^2 + \mathrm{O}_\mathsf{P}(1), \tag{A.22}$$

where all reminder terms are uniform in ζ. Insert the bound $\hat{\sigma}_\zeta^2 - \hat{\sigma}_{\zeta_n}^2 \leq 0$ and use that $\{1+\mathsf{o}_\mathsf{P}(1)\}^{-1} = 1+\mathsf{o}_\mathsf{P}(1)$ and $\{1+\mathsf{o}_\mathsf{P}(1)\}\{1+\mathsf{o}_\mathsf{P}(1)\} = 1+\mathsf{o}_\mathsf{P}(1)$ while $\{1+\mathsf{o}_\mathsf{P}(1)\}\mathrm{O}_\mathsf{P}(1) = \mathrm{O}_\mathsf{P}(1)$ to get the further bound

$$\sum_{i \in \zeta \cap \zeta_n^c} \varepsilon_i^2 \leq \{1 + \mathsf{o}_\mathsf{P}(1)\} \sum_{i \in \zeta^c \cap \zeta_n} \varepsilon_i^2 + \mathrm{O}_\mathsf{P}(1). \tag{A.23}$$

Insert this bound into (A.21) and use $\mathcal{S}_{g_n} = \mathsf{o}_\mathsf{P}(1)$ to get

$$\mathcal{D}_\zeta^2 \leq \Big(\sum_{i \in \zeta^c \cap \zeta_n} \varepsilon_i^2\Big) \mathrm{O}_\mathsf{P}(\mathcal{S}_{g_n}) + \mathsf{o}_\mathsf{P}(1). \tag{A.24}$$

As noted above, $\mathcal{S}_{g_n} = \mathrm{O}_\mathsf{P}(n^{\theta+\eta-1})$ and $\sum_{i \in \zeta^c \cap \zeta_n} \varepsilon_i^2 = \mathsf{o}_\mathsf{P}(h^{\theta+\eta})$ in (A.20). Noting that $\eta < 1/2$ and that $\theta > 0$ can be chosen small, we get $\mathcal{D}_\zeta^2 = \mathsf{o}_\mathsf{P}(1)$. □

B On extreme and intermediate quantiles

We verify Assumption 3.3 (i) for some common distributions. We let $\varepsilon_1, \ldots, \varepsilon_n$ be i.i.d. from a symmetric, unbounded distribution F with extremes $\varepsilon_{(n)} = \max_{1 \leq i \leq n} \varepsilon_i$ and $\varepsilon_{(1)} = \min_{1 \leq i \leq n} \varepsilon_i$. We write $a_n \sim b_n$ if $a_n/b_n \to 1$ in probability.

B.1 Extreme quantiles

If F has exponential tails, we can establish Assumption 3.3 (ib, id) as follows. It suffices to show that $\varepsilon_{(n)}/a_n \to 1$ in probability for some increasing sequence a_n of logarithmic rate, so that $a_n = \mathrm{o}(n^\eta)$ for all $\eta > 0$. In that case, $\varepsilon_{(1)}/a_n \to -1$ by symmetry so that $\max\{\varepsilon^2_{(n)}, \varepsilon^2_{(1)}\}/\min\{\varepsilon^2_{(n)}, \varepsilon^2_{(1)}\} \to 1$ and Assumption 3.3 (ib, id) follows. We can check the sufficient condition using the following multiplicative strong law of large numbers.

Lemma B.1. *(Galambos, 1978, Theorem 4.4.4) Let $a_n = \inf\{y : \mathsf{F}(y) \geq 1 - 1/n\}$. Then, $\varepsilon_{(n)}/a_n \to 1$ a.s. if and only if, for any $k > 1$,*

$$\sum_{n=3}^{\infty}\{1 - \mathsf{F}(ka_n)\} < \infty. \tag{B.1}$$

Example B.1. *Let F be standard normal. Condition (B.1) is satisfied and $a_n \sim \sqrt{2\log(n)}$ (DasGupta, 2008, Example 8.13). Thus, $\varepsilon^2_{(n)} \sim 2\log(n)$ a.s.*

Example B.2. *Let F be standard Laplace. This symmetric distribution has $\mathsf{F}(x) = 1 - \exp(-x)/2$ for $x \geq 0$ so that $a_n = \mathsf{F}^{-1}(1 - 1/n) = -\log(2/n)$ for $n > 2$, Thus, $1 - \mathsf{F}(ka_n) = (2/n)^k/2$ for $n > 2$. Since $\sum_{n=3}^{\infty} n^{-k} < \infty$ for $k > 1$ then condition (B.1) is satisfied. We note that $a_n \sim \log n$, so that $\varepsilon_{(n)} \sim \log n$ a.s.*

Example B.3. *Let F be double geometric with $\mathsf{f}(x) = (1-p)^{|x|-1}p/2$ for $x \in \mathbb{Z}\backslash\{0\}$, so that $\mathsf{F}(x) = 1 - (1-p)^x/2$ for $x \in \mathbb{N}$ and $a_n = \lfloor \log(2/n)/\log(1-p)\rfloor$ for $n > 2$ where $\lfloor\cdot\rfloor$ is the floor. Note, $a_n \sim \log n$. We note that this distribution is not of an extremal type. To see this, modify Example 1.7.15 for the geometric distribution in Leadbetter et al. (1982). To apply Lemma B.1 note that $\lfloor x\rfloor > x - 1 > x - \log n$ for $n > 2$, so that $a_n > \log(2/n^2)/\log(1-p) = \tilde{a}_n$ for $n > 2$. Thus, $1 - \mathsf{F}(ka_n) \leq 1 - \mathsf{F}(k\tilde{a}_n) = (2/n^2)^k/2$ and the argument is completed as in Example B.2.*

If F has polynomial tail behaviour, we need a different argument.

Example B.4. *Let F be t_d with $d > 0$ degrees of freedom. The extremal quotient $\varepsilon_{(n)}/\varepsilon_{(1)}$ converges to a non-degenerate, positive distribution with median 1 (Gumbel and Keeney, 1950). Assumption 3.3 (ib) follows. Next, $1 - \mathsf{F}(x) \sim C_d x^{-d}$ for $x \to \infty$ and $\mathsf{F}^{-1}(1 - 1/n) \sim c_d n^{1/d}$ for $n \to \infty$ for some constants C_d, c_d depending on d (Soms, 1976). Thus, $\{1 - \mathsf{F}(tx)\}/\{1 - \mathsf{F}(t)\} \to x^{-d}$ for $t \to \infty$ so that $\varepsilon_{(n)} \sim n^{1/d}$ (Galambos, 1978, Theorem 2.1.1). Assumption 3.3 (id) follows for $\eta < 1/d$.*

B.2 Intermediate quantiles

We now consider the Assumption 3.3 (ic, iia) concerning intermediate quantiles.

Lemma B.2. *Let $\varepsilon_1, \ldots, \varepsilon_n$ be i.i.d. with distribution function F where $\inf\{x : \mathsf{F}(x) > 0\} = -\infty$. Let $n \to \infty$ and $0 < \rho < 1$. Define $C_n = \mathsf{F}^{-1}(n^{\rho-1}/\log n)/\mathsf{F}^{-1}(n^{-1}\log n)$. If $\limsup_{n\to\infty} C_n \leq C_\rho < 1$ then $\varepsilon_{(n^\rho)}/\varepsilon_{(1)} \leq C_\rho + \mathrm{o}_{\mathsf{P}}(1)$.*

Proof. Theorem 1.8.1 in Leadbetter et al. (1982) with $v_n = \mathsf{F}^{-1}(n^{-1}\log n)$ so that $n\mathsf{F}(v_n) = \log n \to \infty$ shows that $\mathsf{P}\{\varepsilon_{(1)} > v_n\} \to \exp(-\infty) = 0$.

Lemma 1 in Chibisov (1964) with $a_n = \mathsf{F}^{-1}(n^{\rho-1}/\log n)$, $x = 1$, $b_n = 0$, $k_n = n^\rho$ and $u_n(x) = \{n\mathsf{F}(a_n x + b_n) - k_n\}/k_n^{1/2}$ shows that $\mathsf{P}\{\varepsilon_{(k_n)} \leq a_n x + b_n\} - \Phi\{u_n(x)\} \to 0$. In our case, $u_n(x) = n^{\rho/2}\{(\log n)^{-1} - 1\} \to -\infty$, so that $\Phi\{u_n(x)\} \to 0$ and $\mathsf{P}\{\varepsilon_{(k_n)} \leq a_n\} \to 0$.

Let $\epsilon > 0$ be given. Consider the set $A_n = \{\varepsilon_{(k_n)}/\varepsilon_{(1)} \leq C + \epsilon\}$. We must show that $\mathsf{P}(A_n) \to 1$. Rewrite $A_n = \{(\varepsilon_{(k_n)}/a_n) < (Cv_n/a_n)(\varepsilon_{(1)}/v_n)\}$. Let $B_n = \{\varepsilon_{(1)}/v_n \geq 1\}$ and $D_n = \{\varepsilon_{(k_n)}/a_n < 1\}$, so that $\mathsf{P}(B_n), \mathsf{P}(D_n) \to 1$, noting that v_n, a_n are negative. By assumption, $\limsup_{n\to\infty} a_n/v_n \leq C$. Thus, $\forall\epsilon > 0$ then $a_n/v_n \leq C + \epsilon$ for large n. Hence, $(C + \epsilon)v_n/a_n > 1$. Thus, A_n holds on $B_n \cap C_n$, so that $\mathsf{P}(A_n) \geq \mathsf{P}(B_n \cap C_n) \to 1$. □

Example B.5. *Let* F *be standard normal. By Mill's ratio,* $x\Phi(x) \sim -\varphi(x)$ *for* $x \to -\infty$, *so that* $\Phi^{-1}(s_n^{-1}) \sim -(2\log s_n)^{1/2} \to \infty$ *for* $s_n \to \infty$. *We find, for* $0 < \rho < 1$, *that* $C_n = \{\log(n^{\rho-1}/\log n)/\log(n^{-1}\log n)\}^{1/2} \sim (1-\rho)^{1/2} = C_\rho < 1$. *Assumption 3.3* (ic) *follows by Lemma B.2. Example B.1 shows that* $\varepsilon_{(n)}^2 \sim 2\log n = \mathrm{o}(n^\eta)$ *for any* $\eta > 0$. *Thus, for all* $\delta > 0$, *we can choose* η *so small that a* $\rho < 1 - \eta$ *exists so that* $\varepsilon_{(n^\rho)}^2/\varepsilon_{(1)}^2 \leq C_\rho + \mathrm{o}_{\mathsf{P}}(1) < \delta$ *and Assumption 3.3* (iia) *follows.*

Example B.6. *Let* F *be Laplace. Then* $\mathsf{F}(x) = \exp(x)/2$ *for* $x < 0$ *and* $\mathsf{F}^{-1}(\psi) = \log(2\psi)$ *for* $\psi < 1/2$ *Thus,* $C_n = \log(2n^{\rho-1}/\log n)/\log(2n^{-1}\log n)$ *so that* $C_n \sim 1 - \rho = C_\rho < 1$. *Assumption 3.3* (ic, iia) *follow by Lemma B.2.*

Example B.7. *Let* F *be double geometric. Then* $\mathsf{F}(x) = (1-p)^x/2$ *for* $x \in -\mathbb{N}$ *and* $\mathsf{F}^{-1}(\psi) \sim -\log(2\psi)/\log(1-p)$ *for* $\psi \to 0$. *Assumption 3.3* (ic, iia) *follow by Lemma B.2 since* $C_n \sim \{\log(2n^{\rho-1}/\log n)\}/\{\log(2n^{-1}\log n)\} \sim 1 - \rho = C_\rho < 1$.

Example B.8. *Let* F *be the* t_d *with* d *degrees of freedom, so that* $\mathsf{F}^{-1}(\psi) \sim -c_d\psi^{-1/d}$ *for* $\psi \to 0$ *and some constant* c_d *depending on* d *(Soms, 1976). Thus, for any* $0 < \rho < 1$, *we get* $C_n \sim \{(n^{\eta-1}/\log n)/(n^{-1}\log n)\}^{-1/d} = n^{-\eta/d}(\log n)^{2/d} \to 0$. *Thus, by Lemma B.2 we have that* $\varepsilon_{(n^\rho)}/\varepsilon_{(1)}$ *vanishes for any* ρ.

C Heteroscedastic example

Let $z = x^{-\omega}$ be gamma distributed with shape and inverse scale of $\nu = p/2$ and some $\omega > 2$. Let ε given x be $\mathsf{N}(0, 1/z)$. We will require that $p > 4$ so that x, ε have the fourth moments needed for heteroscedastic inference.

We show that ε is t_p distributed. Using a gamma integral, the density is found to be

$$\begin{aligned}
\mathsf{f}_\varepsilon(\varepsilon) &= \int_0^\infty \frac{1}{\sqrt{2\pi/z}}\exp(-z\varepsilon^2/2)\frac{\nu^\nu}{\Gamma(\nu)}z^{\nu-1}\exp(-\nu z)dz \\
&= \frac{\nu^\nu}{\Gamma(\nu)\sqrt{2\pi}}\int_0^\infty z^{\nu-1+1/2}\exp\{-z(\nu+\varepsilon^2/2)\}dz \qquad \text{(C.1)} \\
&= \Big\{\frac{\nu^\nu}{\Gamma(\nu)\sqrt{2\pi}}\Big\}\Big\{\frac{\Gamma(\nu+1/2)}{(\nu+\varepsilon^2/2)^{-\nu-1/2}}\Big\} = \frac{\Gamma\{(p+1)/2\}}{\Gamma(p/2)\sqrt{\pi p}}(1+\varepsilon^2/p)^{-(p+1)/2}.
\end{aligned}$$

We show that $x = z^{-1/\omega}$ has a bounded density so that Assumption 3.1(iii) is satisfied through Example 4.2. By the change-of-variable formula with mapping $z \mapsto z^{-1/\omega} = x$, inverse mapping $x \mapsto x^{-\omega}$ and Jacobean $\omega x^{-\omega-1}$ then x has density

$$\mathsf{f}_x(x) = \mathsf{f}_z(x^{-\omega})\omega x^{-\omega-1} = \frac{\omega\nu^\nu}{\Gamma(\nu)} x^{-\omega\nu-1}\exp(-\nu x^{-\omega}).$$

The density is positive and continuous for $x > 0$ with $\mathsf{f}(x) \to 0$ for $x \to 0$ since the exponential function dominates the power function. Thus, the density is bounded.

We show that $\mathsf{E}x^4 < \infty$ so that Assumption 3.2(ii) is satisfied by the Law of Large Numbers and x has the required moments. With $\nu = p/2 > 2$ and $\omega > 2$ we get

$$\begin{aligned}
(\mathsf{E}x^4)^{\omega/2} &\le \mathsf{E}x^{2\omega} = \mathsf{E}(1/z^2) \\
&= \frac{\nu^\nu}{\Gamma(\nu)}\int_0^\infty \frac{1}{z^2} z^{\nu-1}\exp(-\nu z)dz = \left\{\frac{\nu^\nu}{\Gamma(\nu)}\right\}\left\{\frac{\Gamma(\nu-2)}{\nu^{\nu-2}}\right\} \\
&= \frac{\nu^\nu}{\nu^{\nu-2}(\nu-1)(\nu-2)} = \frac{\nu^2}{(\nu-1)(\nu-2)} < \infty.
\end{aligned}$$

We study the tail behaviour of x as required in Assumption 3.3. It suffices to show that x has thinner tails than ε. Consider n i.i.d. repetitions of x, ε. Example B.4 implies that $\max_{1\le i\le n}\varepsilon_i^2 \sim n^{2/p}$ since ε_i is t_p. Thus, we show $\mathcal{P}_n = \mathsf{P}(\max_{1\le i\le n} x_i^2 \le n^{2/p}) \to 1$. Exploiting the i.i.d. structure, we get

$$\mathcal{P}_n = \mathsf{P}\cap_{1\le i\le n}(x_i^2 \le n^{2/p}) = \{\mathsf{P}(x_1^2 \le n^{2/p})\}^n = \exp\{n\log\mathsf{P}(x_1^2 \le n^{2/p})\}. \tag{C.2}$$

Exploiting that $z = x^{-\omega}$ where $y = \nu z$ is gamma with shape $\nu = p/2$ and scale 1 gives

$$\begin{aligned}
\mathsf{P}(x_1^2 \le n^{2/p}) &= \mathsf{P}(z \ge n^{-\omega/p}) = \mathsf{P}(y \ge \nu n^{-\omega/p}) \\
&= \frac{1}{\Gamma(\nu)}\int_{\nu n^{-\omega/p}}^\infty y^{\nu-1}\exp(-y)dy.
\end{aligned}$$

Expand the gamma integral (Gradshteyn and Ryzhik, 1965, 8.354.2) to get

$$\mathsf{P}(x_1^2 \le n^{-2/p}) = 1 - \frac{(\nu n^{-\omega/p})^\nu}{\nu\Gamma(\nu)} + \mathrm{o}\{(n^{-\omega/p})^\nu\} = 1 - \frac{\nu^{\nu-1}n^{-\omega/2}}{\Gamma(\nu)} + \mathrm{o}(n^{-\omega/2}).$$

Insert in (C.2) and expand the logarithm to see that $\mathcal{P}_n \to 1$ when $\omega > 2$.

References

Agullo, J., Croux, C., and Van Aelst, S. (2008). The multivariate least-trimmed squares estimator. *J. Multivariate Anal.*, 99:311–338.

Alfons, A., Croux, C., and Gelper, S. (2013). Sparse least trimmed squares regression for analyzing high-dimensional large data sets. *Ann. Appl. Stat.*, 7:226–248.

Atkinson, A. C., Riani, M., and Cerioli, A. (2010). The forward search: Theory and data analysis (with discussion). *J. Korean Statist. Soc.*, 39:117–163.

Berenguer-Rico, V., Johansen, S., and Nielsen, B. (2023). A model where the least trimmed squares estimator is maximum likelihood. *J. Roy. Statist. Soc. Ser. B.* To appear.

Butler, R. (1982). Nonparametric interval and point prediction using data trimmed by a Grubbs-type outlier rule. *Ann. Statist.*, 10:197–204.

Chen, X. R. and Wu, Y. H. (1988). Strong consistency of M-estimators in linear models. *J. Multivariate Anal.*, 27:116–130.

Chibisov, D. M. (1964). On limit distributions for order statistics. *Theory Probab. Appl.*, 9:142–147.

Čížek, P. (2005). Least trimmed squares in nonlinear regression under dependence. *J. Statist. Plann. Inference*, 136:3967–3988.

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

DasGupta, A. (2008). *Asymptotic Theory of Statistics and Probability.* Springer, New York.

Davies, L. (1990). The asymptotics of S-estimators in the linear regression model. *Ann. Statist.*, 18:1651–1675.

Galambos, J. (1978). *The Asymptotic Theory of Extreme Order Statistics.* John Wiley & Sons, New York.

Gallegos, M. T. and Ritter, G. (2009). Trimmed ML estimation of contaminated mixtures. *Sankhya A*, 71:164–220.

Gradshteyn, I. S. and Ryzhik, I. M. (1965). *Table of Integrals, Series and Products.* Academic Press, New York.

Gumbel, E. J. and Keeney, R. D. (1950). The extremal quotient. *Ann. Math. Statist.*, 21:523–538.

He, X., Jurečková, J., Koenker, R., and Portnoy, S. (1990). Tail behavior of regression estimators and their breakdown points. *Econometrica*, 58:1195–1214.

Hössjer, O. (1994). Rank-based estimates in the linear model with high breakdown point. *J. Amer. Statist. Assoc.*, 89:149–158.

Huber, P. J. (1964). Robust estimation of a location parameter. *Ann. Math. Statist.*, 35:73–101.

Johansen, S. and Nielsen, B. (2009). Saturation by indicators in regression models. In *The Methodology and Practice of Econometrics: Festschrift in Honour of David F. Hendry*, pages 1–36. Oxford University Press, Oxford.

Johansen, S. and Nielsen, B. (2016). Asymptotic theory of outlier detection algorithms for linear time series regression models (with discussion). *Scand. J. Stat.*, 43:321–81.

Johansen, S. and Nielsen, B. (2019). Boundedness of M-estimators for multiple linear regression in time series. *Econometric Theory*, 35:653–683.

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

Rousseeuw, P. (1985). Least median of squares regressions. In Grossmann, W., Pflug, G., Vincze, I., and Wertz, W., editors, *Mathematical Statistics and Applications*, pages 283–297. Reidel, Dordrecht.

Rousseeuw, P. J. (1984). Least median of squares regressions. *J. Amer. Statist. Assoc.*, 79:871–880.

Rousseeuw, P. J. (1994). Unconventional features of positive-breakdown estimators. *Statist. Probab. Lett.*, 19:417–431.

Rousseeuw, P. J. and Leroy, A. M. (1987). *Robust Regression and Outlier Detection.* John Wiley & Sons, Hoboken, NJ.

Rousseeuw, P. J., Perrotta, D., Riani, M., and Hubert, M. (2019). Robust monitoring of time series with application to fraud detection. *Econom. Stat.*, 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., and Schader, M., editors, *Data Analysis: Scientific Modeling and Practical Application*, pages 335–346. Springer Verlag.

Scholz, F. W. (1980). Towards a unified definition of maximum likelihood. *Canad. J. Statist.*, 8:193–203.

Soms, A. P. (1976). An asymptotic expansion for the tail area of the t-distribution. *J. Amer. Statist. Assoc.*, 71:728–730.

Víšek, J. A. (2006). The least trimmed squares; part III: Asymptotic normality. *Kybernetika*, 42:203–224.

Watts, V., Rootzén, H., and Leadbetter, M. R. (1982). On limiting distributions of intermediate order statistics from stationary sequences. *Ann. Probab.*, 10:653–662.

Welsh, A. H. and Ronchetti, E. (2002). A journey in single steps: robust one-step M-estimation in linear regression. *J. Statist. Plann. Inference*, 103:287–310.

Yohai, V. J. (1987). High breakdown-point and high efficiency robust estimates for regression. *Ann. Math. Statist.*, 15:642–656.

Zuo, Y. (2022). Asymptotics for the least trimmed squares estimator. Technical report, arXiv:2210.06460.