Asymptotic theory of outlier detection algorithms for linear time series regression models

SØREN JOHANSEN
Department of Economics, University of Copenhagen and CREATES, University of Aarhus
BENT NIELSEN
Nuffield College & Department of Economics, University of Oxford
Running headline : Outlier detection for time series regression

June 27 2015

ABSTRACT: **Outlier detection algorithms are intimately connected with robust statistics that down-weight some observations to zero. We define a number of outlier detection algorithms related to the Huber-skip and the Least Trimmed Squares estimators, including the 1-step Huber skip estimator and the Forward Search. Next, we review a recently developed asymptotic theory of these. Finally, we analyze the gauge, the fraction of wrongly detected outliers, for a number of outlier detection algorithms and establish an asymptotic normal and a Poisson theory for the gauge.**
Keywords: Forward Search, gauge, Huber-skip, Impulse Indicator Saturation, iteration of 1-step estimators, iterated martingale inequality, gauge, 1-step Huber-skip, Robustified Least Squares, weighted and marked empirical processes

1 Introduction

We consider some outlier detection methods for linear regression models with regressors that are stationary or deterministically or stochastically trending. Outlier detection methods rely on cutoff values when classifying observations as outliers or not. We review some recent asymptotic results for such methods and apply the results to calibrate the cutoff values.

There is a close link between outlier detection methods and robust estimation methods that down-weight some of the observations to zero. Examples of such estimators are the Huber-skip by Huber (1964) and the Least Trimmed Squares by Rousseeuw (1984). Once the estimator has been calculated, the observations with weight zero are classified as outliers, and, conversely, if we start with an outlier detection method, then an estimator based on the remaining 'good' observations will be robust. When building a statistical model, the user can apply outlier detection methods in combination with considerations about the substantive context to decide which observations are 'good' and how to treat the 'outliers'.

In the regression model

$$y_i = \beta' x_i + \varepsilon_i, \tag{1}$$

where ε_i/σ are independent with common reference density f, outliers are pairs of observations (y_i, x_i) that do not conform with the model. In other words, a pair of observations (y_i, x_i) is an outlier if the scaled residual $r_i = (y_i - \beta' x_i)/\sigma$ does not conform with the reference density f. This has slightly different consequences for cross-sectional data and for

time series data. For cross-sectional data, the pairs of observations $(y_1, x_1), \ldots, (y_n, x_n)$ are unrelated. Thus, if the residual r_i is classified as an outlier, then the pair (y_i, x_i) is dropped. We can interpret this as a residual not conforming with the model, or that y_i or x_i or both are not correct. This is different for time-series data, where the regressors include lagged dependent variables. Consider for instance a first order autoregression where $x_i = y_{i-1}$. We then distinguish between innovative and additive outliers. Classifying the residual as an innovative outlier has the consequence that we discard the evaluation of the dynamics from y_{i-1} to y_i but not the observations y_{i-1} and y_i as such. Indeed, y_{i-1} appears as the dependent variable at time $i-1$ and the y_i as the regressor at time $i+1$, respectively. Thus, finding a single outlier in a time series context, implies that the observations are considered correct, but possibly not generated by the model. An additive outlier arises if an observation y_i is wrongly measured. For a first order autoregression, this is captured by two consecutive residuals r_i and r_{i+1}. Discarding these, the observation y_i will not appear.

It is open to discussion which outlier detection method to use, see for instance Section 1.4 in Hampel *et al.* (1986). A simple outlier detection method consists of testing if y_i has the mean given by the model. This applies a preliminary estimator $(\tilde{\beta}, \tilde{\sigma}^2)$ and residuals $\tilde{r}_i = (y_i - \tilde{\beta}' x_i)/\tilde{\sigma}$. An observation is classified as outlier if $|\tilde{r}_i| \geq c$, where c is a suitable cutoff value and we re-estimate the parameter β by regression based on the remaining observations. The new estimator is a *1-step Huber-skip estimator* or a reweighted least squares estimator with binary weights, see Welsh & Ronchetti (2002). This can of course be iterated to give, for instance, the *Forward Search* suggested by Hadi & Simonoff (1993), see also Atkinson & Riani (2000). These are analyzed in detail in this paper.

In order to use the algorithms with confidence, we need to understand their properties when all observations are 'good'. When classifying the observations we denote by $v_i = 1$ the observations classified as 'good' and $v_i = 0$ for the outliers. We define *the (empirical) gauge* as the fraction of outliers found

$$\hat{\gamma} = n^{-1} \sum_{i=1}^{n} (1 - v_i), \tag{2}$$

and the population gauge, γ, is the limit of its expected value $\mathsf{E}\hat{\gamma}$ when there are no outliers. This is similar to the size of a test, yet a slightly different attempt to control errors of the first kind. Similarly we need the asymptotic variance of the estimator based on the good observations, where the efficiency loss is the price paid for using a robust estimator.

The origins of the notion of gauge are as follows. Hoover & Perez (1999) studied the properties of a general-to-specific algorithm for variable selection through a simulation study. They considered various measures for the performance of the algorithm, that are related to what is now called the gauge. One of these, they referred to as the size, and this was the number of falsely significant variables divided by the difference between the total number of variables and the number of variables with non-zero coefficients. The Hoover-Perez idea for regressor selection was the basis of the *PcGets* and *Autometrics* algorithms, see for instance Hendry & Krolzig (2005), Doornik (2009) and Hendry & Doornik (2014). Through extensive simulation studies, the critical values of these algorithms have been calibrated in terms of the false detection rates for irrelevant regressors or irrelevant outliers. The term gauge was introduced in Hendry & Santos (2010) and Castle *et al.* (2011).

The paper has two parts. The first part starts with a motivating empirical example, where least squares is applied to find outliers. Next, we give an overview of recent asymptotic

results for outlier detection methods including 1-step Huber-skip estimators and iterations thereof, Impulse Indicator Saturation as well as the Forward Search. This builds in part on some of our own papers, Johansen and Nielsen (2009, 2010, 2013, 2015a). The results for the estimators are given as stochastic expansions. For instance, the simple 1-step Huber-skip estimators satisfy

$$N^{-1}(\hat{\beta}-\beta)=\frac{1}{\psi}\Sigma_n^{-1}N'\sum_{i=1}^{n}x_i\varepsilon_i 1_{(|\varepsilon_i|\leq\sigma c)}+\frac{2c\mathsf{f}(c)}{\psi}N^{-1}(\tilde{\beta}-\beta)+\mathrm{o_P}(1), \tag{3}$$

where $\psi=P(|\varepsilon_1|\leq\sigma c)$ and N is a normalization, such that $\Sigma_n=N'\sum_{i=1}^{n}x_ix_i'N=\mathrm{O_P}(1)$. This shows how the 1-step Huber-skip estimator, $\hat{\beta}$, depends on the initial estimator $\tilde{\beta}$. The advantage of this formulation in terms of an expansion, is that it unifies the theory for cases with stationary and non-stationary regressors. Limit distributions of $\hat{\beta}$ can be derived from this expansion for particular choices of the regressors. The expansion (3) also forms the basis for the analysis of iterated estimators.

In the second part we provide an asymptotic theory for setting the cutoff value c for the gauge. As a simple example, consider the 1-step Huber-skip estimator where the initial estimators $\tilde{\beta},\tilde{\sigma}$ are the least squares estimators. Then the empirical gauge, $\hat{\gamma}=n^{-1}\sum_{i=1}^{n}1_{(|y_i-\tilde{\beta}'x_i|\geq\tilde{\sigma}c)}$, converges in probability to $\mathsf{P}(|\varepsilon_1|\geq\sigma c)=1-\psi$, the size of the underlying test. Moreover it has the stochastic expansion

$$n^{1/2}\{\hat{\gamma}-(1-\psi)\}=n^{-1/2}\sum_{i=1}^{n}\{1_{(|\varepsilon_i|>\sigma c)}-(1-\psi)\}-\frac{c\mathsf{f}(c)}{\sigma^2}n^{1/2}(\tilde{\sigma}^2-\sigma^2)+\mathrm{o_P}(1), \tag{4}$$

where f is the density of the innovation ε_i/σ. Thus, for 1-step Huber-skip estimators, the asymptotic population gauge, γ, is the size of the underlying test, but the asymptotic distribution of the empirical gauge depends on the initial estimator. For the Forward Search the results for the gauge are completely different, see Theorem 9. The paper ends with a conclusion, and the main proofs are left for Appendix.

Part I
Review of recent asymptotic results

2 A motivating example

What is an outlier? How do we detect them? How should we deal with them? There is no simple, universally valid answer to these questions – it all depends on the context. We will therefore motivate our analysis with an example from time series econometrics.

Demand and supply is key to discussing markets in economics. To study this Graddy (1995, 2006) collected data on prices and quantities from the Fulton fish market in New York. For our purpose the following will suffice. The data consists of daily quantities of whiting sold by one wholesaler over the period 2 December 1991 to 8 May 1992. Figure 1(a) shows the daily aggregated quantity Q_t measured in pounds. The logarithm of the quantity, $q_t=\log Q_t$ is shown in panel (b). The supply of fish depends on the weather at sea, where

the fish is caught. Panel (c) shows a binary variable S_t taking value 1 if the weather is stormy. The present analysis is taken from Section 13.5 of Hendry & Nielsen (2007).

< Fig1 here>

A simple autoregressive model for log quantities q_t gives

$$\begin{array}{cccccc} \hat{q}_t & = & \underset{(0.8)}{7.0} & + \underset{(0.09)}{0.19} q_{t-1} & - \underset{(0.15)}{0.36}\, S_t, \\ \text{(standard error)} & & & & & \\ \text{[t-statistic]} & & [8.8] & [2.03] & [-2.39] & \end{array} \tag{5}$$

$$\begin{aligned} \hat{\sigma} = 0.72, \quad \hat{\ell} = -117.82, \quad R^2 = 0.090, \quad t = 2, \ldots, 111, \\ \chi^2_{norm}[2] = 6.9\ [p{=}\ 0.03], \quad \chi^2_{skew}[1] = 6.8\ [p{=}\ 0.01], \quad \chi^2_{kurt}[1] = 0.04\ [p{=}\ 0.84], \\ \mathsf{F}_{ar(1-2)}[2, 106] = 0.9\ [p{=}\ 0.40], \quad \mathsf{F}_{arch(1)}[1, 106] = 1.4\ [p{=}\ 0.24], \\ \mathsf{F}_{het}[3, 103] = 2.0\ [p{=}\ 0.12], \quad \mathsf{F}_{reset}[1, 106] = 1.8\ [p{=}\ 0.18]. \end{aligned}$$

Here $\hat{\sigma}^2$ is the residual variance, $\hat{\ell}$ is the log likelihood, T is the sample size. The residual specification tests include cumulant based tests for skewness, χ^2_{skew}, kurtosis, $\chi^2_{\mathsf{kurtosis}}$ and both, $\chi^2_{\mathsf{norm}} = \chi^2_{\mathsf{skew}} + \chi^2_{\mathsf{kurtosis}}$, a test F_{ar} for autoregressive temporal dependence, see Godfrey (1978), a test $\mathsf{F}_{\mathsf{arch}}$ for autoregressive conditional heteroscedasticity, see Engle (1982), a test $\mathsf{F}_{\mathsf{het}}$ for autoregressive conditional heteroscedasticity, see White (1980), and a test $\mathsf{F}_{\mathsf{reset}}$ for functional form, see Ramsey (1969). We note that the above references only consider stationary processes, but the specification tests also apply for non-stationary autoregressions, see Kilian & Demiroglu (2000) and Engler & Nielsen (2009) for $\chi^2_{\mathsf{skew}}, \chi^2_{\mathsf{kurtosis}}$ and Nielsen (2006) for F_{ar}. The computations were done using OxMetrics, see Doornik & Hendry (2013). Figure $1(b, d)$ shows the fitted values and the standardized residuals.

The specification tests indicate that the residuals are skew. Indeed the time series plot of the residuals in Figure $1(d)$ shows a number of large negative residuals. The three largest residuals have an interesting institutional interpretation. The observations 18 and 34 are Boxing Day and Martin Luther King Day, which are public holidays, while observation 95 is Wednesday before Easter. Thus, from a substantive viewpoint it seems preferable to include dummy variables for each of these days, which gives

$$\begin{array}{cccccccc} \hat{q}_t = & \underset{(0.7)}{7.9} & + \underset{(0.08)}{0.09} q_{t-1} & - \underset{(0.14)}{0.36}\, S_t & - \underset{(0.66)}{1.94}\, D_t^{18} & - \underset{(0.66)}{1.82}\, D_t^{34} & - \underset{(0.66)}{2.38}\, D_t^{95}, \\ & [10.8] & [1.04] & [-2.68] & [-3.00] & [-2.75] & [-3.64] & \end{array} \tag{6}$$

$$\hat{\sigma} = 0.64, \quad \hat{\ell} = -104.42, \quad R^2 = 0.287, \quad t = 2, \ldots, 111.$$

Specification tests, which are not reported, indicate a marked improvement in the specification. Comparing the regressions (5) and (6) it is seen that the lagged quantities were marginally significant in the first, misspecified regression, but not significant in the second, better specified, regression. It is of course no surprise that outliers matter for statistical inference - and that institutions matter for markets.

The above modelling strategy blends usage of specification tests, graphical tools and substantive arguments. It points at robustifying a regression by removing outliers and then refitting the regression. We note that outliers are defined as observations that do not conform with the statistical model. In the following we consider some outlier detection algorithms that are inspired by this example. The algorithms are solely based on statistical information and we discuss their mathematical properties. In practice, outcomes should of course be assessed within the substantive context. We return to this example in Section 10.

3 The linear time series regression model

Throughout, we consider data (y_i, x_i), $i = 1, \ldots, n$, where y_i is univariate and x_i has dimension $\dim x$. The regressors are possibly trending in a deterministic or stochastic fashion. We assume that (y_i, x_i) satisfy the linear multiple regression equation

$$y_i = \beta' x_i + \varepsilon_i, \qquad i = 1, \ldots, n.$$

The innovations, ε_i, are independent of the filtration $\mathcal{F}_{i-1}$, which is the sigma-field generated by $x_1, \ldots, x_i$ and $\varepsilon_1, \ldots, \varepsilon_{i-1}$. We analyze properties of outlier detection methods when the innovations have no outliers. The methods we consider, combine least squares and classification using absolute residuals. Hence we assume that ε_i/σ have a known, symmetric density f and distribution function $\mathsf{F}(c) = \mathsf{P}(\varepsilon_i \leq \sigma c)$, see Assumption 1 for details. The results are formulated so that they generalize to other symmetric densities including t-distributions with sufficiently many moments. Non-symmetric densities will be relevant in future work discussing the situation where outliers are present.

We consider algorithms using absolute residuals, which implicitly assume a symmetric density of the non-outlying innovations. In case of symmetry, the absolute value errors $|\varepsilon_i|/\sigma$ have density $\mathsf{g}(c) = 2\mathsf{f}(c)$ and distribution function $\mathsf{G}(c) = \mathsf{P}(|\varepsilon_1| \leq \sigma c) = 1 - 2\mathsf{F}(c)$. We define $\psi = \mathsf{G}(c)$ so that $c = c_\psi$ is the ψ quantile of G

$$c = \mathsf{G}^{-1}(\psi) = \mathsf{F}^{-1}\{(1+\psi)/2\}, \qquad \psi \in [0, 1[. \tag{7}$$

In general we define the truncated moments

$$\tau = \int_{-c}^{c} u^2 \mathsf{f}(u) du, \qquad \varkappa = \int_{-c}^{c} u^4 \mathsf{f}(u) du, \tag{8}$$

and the conditional variance of ε_1/σ given $|\varepsilon_1| \leq \sigma c$ is

$$\varsigma^2 = \tau/\psi, \tag{9}$$

which will serve as a consistency correction for the variance estimators based on the truncated sample. Note that the parameters $\tau, \varkappa, \varsigma$ depend on c or ψ. We usually leave out this dependence for readability. For the normal density we find

$$\tau = \psi - 2c\mathsf{f}(c), \qquad \varkappa = 3\psi - 2c(c^2+3)\mathsf{f}(c). \tag{10}$$

4 The outlier detection algorithms

Outlier detection is closely linked to robust estimation. First, we discuss the Huber-skip and the Least Trimmed Squares (LTS) and show how they give rise to 1-step estimators and outlier detection algorithms. We then define two iterated versions: the m-step Huber-skip, and the Forward Search. We consider as outcome of an outlier detection algorithm, not only the set of outliers, but also the estimator based on the good observations.

4.1 Outlier detection based on two robust estimators

The Huber-skip is the minimizer of the objective function

$$R_n(\beta) = n^{-1}\sum_{i=1}^{n}(y_i - x_i'\beta)^2 1_{(|y_i - x_i'\beta| \leq \sigma c)} + c^2 1_{(|y_i - x_i'\beta| > \sigma c)}, \tag{11}$$

introduced by Huber (1964). The resulting estimator can be described as least squares among the observations with estimated residuals bounded by σc for known σ. If $\hat{\beta}$ denotes the minimizer, one can define outliers by $v_i = 1_{(|y_i - x_i'\hat{\beta}| \leq \sigma c)} = 0$.

The calculation of the Huber-skip is quite complicated. Figure 2 illustrates the non-convex objective function R_n for the fish data, using R 3.1.1, see R Development Core Team (2014). The specification is as in equation (5). All parameters apart from that on q_{t-1} are held fixed at the values in (5). Panel (a) shows that for a large cutoff c, R_n is quadratic in the central part. Panel (b) shows that for a smaller cutoff c, R_n is non-differentiable in a finite number of points.

< Fig2 here>

The asymptotic theory of the Huber-skip (and other M-estimators) has been studied in some detail for the situation without outliers. Huber (1964) gave a theory for M-estimation of location for convex objective functions. Chen & Wu (1988) showed strong consistency of M-estimators for general criterion functions and i.i.d. or deterministic regressors, while Johansen and Nielsen (2015b) analyze time series regression. See also pages 197 and 215 in Jurečková & Sen (1996) for alternative proofs for the location case.

The Least Trimmed Squares (LTS) estimator was introduced by Rousseeuw (1984). For a given β, we order the absolute residuals $\xi_i = |y_i - x_i'\beta|$, and let the k'th largest be $\xi_{(k)} = |y - x'\beta|_{(k)}$. The LTS is defined as the minimizer of

$$R_n(\beta) = \sum_{i=1}^{k}|y - x'\beta|_{(i)}^2 = \sum_{i=1}^{n}(y_i - x_i'\beta)^2 1_{(|y_i - x_i'\beta| \leq \xi_{(k)})}. \tag{12}$$

Compared to the Huber skip, the main difference is that weights are now based on order statistics, which are scale equivariant. If $\hat{\beta}$ denotes the estimator, we define the set of outliers as those observations for which $\hat{\xi}_i = |y_i - x_i'\hat{\beta}| > \hat{\xi}_{(k)}$, or $v_i = 1_{(|y_i - x_i'\hat{\beta}| \leq \hat{\xi}_{(k)})} = 0$. Thus, again a robust estimator gives rise to an outlier detection algorithm.

The Least Trimmed Squares is known to have breakdown point of $1 - \psi = 1 - k/n$ for $\psi < 1/2$, see Section 3.4 of Rousseeuw & Leroy (1987). An asymptotic theory is provided by Víšek (2006a,b,c). The estimator is computed through a binomial search algorithm which is infeasible in most practical situations, see Section 5.7 of Maronna, *et al.* (2006) for a discussion. A number of iterative approximations have been suggested such as the Fast LTS algorithm by Rousseeuw & van Driessen (1998). This leaves additional questions with respect to the properties of the approximating algorithms.

4.2 Outlier detection based on 1-step estimators

The main ingredient in our analysis is the notion of a 1-step estimator. When observations have already been classified as outliers or not, we can estimate the model using least squares on the non-outlying observations, that is calculate the 1-step estimator, and then reevaluate the outlier classification. As long as we have a good starting point, this addresses the computational difficulties in the Huber-skip and the Least Trimmed Squares. In the next section we study various algorithms based on 1-step estimators.

Consider an objective function with binary stochastic weights v_i for each observation, such that the 'good' observations satisfy $v_i = 1$, while the set of outliers is

$$\hat{\mathcal{O}} = (i : v_i = 0). \tag{13}$$

These weights define the method, and examples are given below. We then apply the least squares method on the set of 'good' observations to get

$$\hat{\beta} = (\sum_{i=1}^{n} v_i x_i x_i')^{-1} (\sum_{i=1}^{n} v_i x_i y_i), \tag{14}$$

as well as the scale estimator

$$\hat{\sigma}^2 = \varsigma^{-2} (\sum_{i=1}^{n} v_i)^{-1} \{\sum_{i=1}^{n} v_i (y_i - x_i'\hat{\beta})^2\}, \tag{15}$$

where $\varsigma^2 = \tau/\psi$ is the consistency correction factor defined in (9). We now use this setup to define two 1-step estimators: the 1-step Huber-skip and the 1-step LTS.

Definition 1 *Let $\tilde{\beta}, \tilde{\sigma}^2$ denote initial estimators. Then the 1-step Huber-skip estimators $\hat{\beta}, \hat{\sigma}^2$ are given by (14), (15) with a cutoff value c and weights*

$$v_i = 1_{(|y_i - x_i'\tilde{\beta}| \leq \tilde{\sigma} c)}. \tag{16}$$

Definition 2 *Let $\tilde{\beta}$ denote an initial estimator, while $\tilde{\xi}_{(k)}$ is the k-th smallest order statistic of the absolute residuals $\tilde{\xi}_i = |y_i - x_i'\tilde{\beta}|$. Then the 1-step LTS estimators $\hat{\beta}, \hat{\sigma}^2$ are given by (14), (15) with a cutoff $k \leq n$ and weights*

$$v_i = 1_{(|y_i - x_i'\tilde{\beta}| \leq \tilde{\xi}_{(k)})}. \tag{17}$$

The weights satisfy $\sum_{i=1}^{n} v_i = k$.

The cutoff values in the two definitions can be linked through $k/n = \psi = \mathsf{G}^{-1}(c)$. Thus, the methods can be calibrated through either of ψ, c, k.

The 1-step estimators relate to the 1-step M-estimators analyzed by Bickel (1975), although he was primarily concerned with more smooth weights v_i than those considered here. He also suggested iteration, but gave no results. The present 1-step estimators were considered by Ruppert & Carroll (1980) and Welsh & Ronchetti (2002). They are also reweighted least squares estimators with binary weights. He & Portnoy (1992) gave a theory for reweighted least squares estimators with smooth weights.

4.3 Some iterative outlier detection algorithms

We discuss some outlier detection algorithms defined by iteration of 1-step estimators.

The m-step Huber-skip algorithm is defined as follows.

Algorithm 1 *m-step Huber-skip.*
Choose initial estimators $\tilde{\beta} = \hat{\beta}^{(0)}, \tilde{\sigma} = \hat{\sigma}^{(0)}$, and a cutoff $c > 0$. Let $k = 0$. Apply the 1-step Huber-skip with initial estimators $\tilde{\beta} = \hat{\beta}^{(k)}$, $\tilde{\sigma} = \hat{\sigma}^{(k)}$ to get outliers $\hat{\mathcal{O}}^{(k+1)}$ and estimators $\hat{\beta}^{(k+1)}, \hat{\sigma}^{(k+1)}$ from (13), (14) and (15). If $k < m$ repeat with $k = k + 1$.

m-step Robustified least squares is a special case of the m-step Huber-skip, where the initial estimators are chosen as the full sample least squares estimators. The 1-step robustified least squares was applied in the analysis of the Fulton fish market data in Section 2. This approach is very common. This is fragile when there is a single high leverage outlier or when there are more than a few outliers, see Welsh & Ronchetti (2002) for a discussion.

Impulse indicator saturation improves the robustified least squares in a simple way. This is based on a suggestion by Hendry (1999), see also Chapter 15 of Hendry & Doornik (2014). The idea is to split the sample in two halves. Run regression on the first half and use this to find outliers in the second half. Then run regression on the second half and use this to find outliers in the first half. Then remove the two sets of outliers and run regression on the remaining observations. Now, suppose there is a leverage point in the second sample half. Then the first half estimator is consistent and will detect the leverage point. With several leverage points it can be necessary to split the sample in different ways. The Autometrics algorithm does this, see Doornik (2009).

Infinite iteration of 1-step estimators. Instead of iterating 1-step estimators a fixed number of times, we could iterate until we achieve a fixed point. An asymptotic theory is given in Johansen and Nielsen (2013). The Forward Search is an example of such iteration.

The Forward Search algorithm was suggested for the multivariate location model by Hadi (1992), for multiple regression by Hadi & Simonoff (1993) and developed further by Atkinson & Riani (2000), see also Atkinson, *et al.* (2010) and Johansen and Nielsen (2010). The algorithm starts with a robust estimator of the regression parameters. This is used to construct the set of observations with the smallest m_0 absolute residuals. We then estimate β, σ based on those m_0 observations and compute absolute residuals of all n observations. The observations with the $m_0 + 1$ smallest residuals are selected, and new estimates are found from these $m_0 + 1$ observations. This 1-step LTS estimation step is then iterated, so that the number of selected observations is gradually increased.

Algorithm 2 ***Forward Search.***
Choose an integer $m_0 < n$ and an initial estimator $\tilde{\beta} = \hat{\beta}^{(m_0)}$. Let $k = m_0$. Apply the 1-step LTS with initial estimators $\tilde{\beta} = \hat{\beta}^{(k)}$ and cutoff $k + 1$ to get the order statistic $\hat{z}^{(k)} = \xi^{(k)}_{(k+1)}$ for the $(k + 1)$ smallest absolute residual, outliers $\hat{\mathcal{O}}^{(k+1)}$ and estimators $\hat{\beta}^{(k+1)}, \hat{\sigma}^{(k+1)}$ using (13), (14) and (15). If $k < n$ repeat with $k = k + 1$. If the Forward Search is stopped at a stopping time $\hat{m}$ so $m_0 \leq \hat{m} \leq n$ we get outliers $\hat{\mathcal{O}}^{(\hat{m})}$ and estimators $\hat{\beta}^{(\hat{m})}, \hat{\sigma}^{(\hat{m})}$.

Applying the algorithm for $k = m_0, \ldots, n-1$ results in sequences of order statistics $\hat{z}^{(k)}$, least squares estimators $\hat{\beta}^{(k)}, \hat{\sigma}^{(k)}$, and sets of outliers $\hat{\mathcal{O}}^{(k)}$ with $n-k$ elements. We note that $\hat{\beta}^{(n)}, \hat{\sigma}^{(n)}$ are the full sample estimators, while $\hat{\mathcal{O}}^{(n)}$ is empty.

The idea of the Forward Search is to monitor the plot of scaled forward residuals $\hat{z}^{(k)}/\hat{\sigma}^{(k)}$ and stop when this is large. To do this we find the asymptotic distribution of $\hat{z}^{(k)}/\hat{\sigma}^{(k)}$ and add a curve of pointwise q-quantiles, $c_q(k)$, for $\hat{z}^{(k)}/\hat{\sigma}^{(k)}$ as a function of k for some q. We choose the stopping time $\hat{m}$ as the first exceedance time

$$\hat{m} = \min\{k : \hat{z}^{(k)}/\hat{\sigma}^{(k)} > c_q(k)\}, \tag{18}$$

with the convention that $\hat{m} = n$ if there is no exceedance. Asymptotic theory for the forward residuals, $\hat{z}^{(k)}/\hat{\sigma}^{(k)}$, is reviewed in Section 7.2. A theory for the estimator $\hat{m}$ and hence guidance for choosing q is given in Section 9.

A variant of the Forward Search advocated by Atkinson & Riani (2000) is to use the minimum deletion residuals

$$\hat{d}^{(k)} = \min_{i \in \hat{\mathcal{O}}^{(k-1)}} \hat{\xi}_i^{(k)} \tag{19}$$

instead of the forward residuals $\hat{z}^{(k)}$, where $\hat{\mathcal{O}}^{(k-1)}$ is based on the estimators $\hat{\beta}^{(k-1)}, \hat{\sigma}^{(k-1)}$.

5 An empirical process theory

The 1-step estimators are least squares estimators for observations that are selected by a previous estimator. We can analyze these using empirical process techniques. The central argument in the asymptotic analysis is to linearize the estimators with respect to the previous estimator. In the following, we describe the relevant empirical processes, outline the intuition behind their analysis, including a new iterated martingale inequality, and finish by stating the assumptions needed throughout the remainder of the paper. We refer to Johansen and Nielsen (2015a) for a detailed exposition.

5.1 Weighted and marked empirical processes

Definition. The 1-step estimators for β and σ^2, see (14) and (15) have estimation errors that can be expressed in terms of product moments of the form

$$\sum_{i=1}^{n} v_i, \qquad \sum_{i=1}^{n} v_i x_i \varepsilon_i, \qquad \sum_{i=1}^{n} v_i x_i x_i', \qquad \sum_{i=1}^{n} v_i \varepsilon_i^2, \tag{20}$$

where v_i are indicator functions for small residuals, like (16) and (17). Such sums of indicator functions are weighted and marked empirical processes. The $\mathcal{F}_{i-1}$-predictable factors x_i and $x_i x_i'$ are called weights in line with Koul (2002). The unbounded, $\mathcal{F}_i$-adapted factors ε_i and ε_i^2 are the marks.

Expansions of the product moments (20) can be found in Johansen and Nielsen (2015a) and form the basis for the results reviewed in Sections 6, 7. For the new developments in Part II we only need to consider a special case which we review below. The Assumptions needed throughout are listed in Section 5.2.

Expansion. We discuss the simplest case with weights and marks of unity for the 1-step Huber-skip, which is the one we need in the gauge considerations in Part II. For the moment we assume stationary regressors, which implies $N = n^{-1/2}$. When $n^{1/2}(\tilde{\beta} - \beta)$ and $n^{1/2}(\tilde{\sigma}^2 - \sigma^2)$ are both $\mathrm{O}_{\mathsf{P}}(1)$, and c is fixed, we get the expansion

$$\sum_{i=1}^{n} v_i = \sum_{i=1}^{n} 1_{(|y_i - x_i'\tilde{\beta}| \leq \tilde{\sigma} c)} = \sum_{i=1}^{n} 1_{(|\varepsilon_i| \leq \sigma c)} + \frac{c\mathsf{f}(c)}{\sigma^2} n^{1/2}(\tilde{\sigma}^2 - \sigma^2) + \mathrm{o}_{\mathsf{P}}(n^{1/2}). \tag{21}$$

We note two features. First, the expansion does not depend on the estimation error for the regression coefficient β, due to the absolute cutoff of residuals and the symmetric density. Second, under the Assumption 1 below, the expansion is valid for a wide range of regressors.

This particular expansion of the empirical distribution function for residuals is well-known; see Johansen and Nielsen (2009) with one-sided versions in Koul & Ossiander (1994), Koul (2002), see also Engler & Nielsen (2009) for autoregressive models. Theorems 4.1-4.4, Lemma D.5 of Johansen and Nielsen (2015a) give similar results for the product moments (20) with marks and weights, both for 1-step Huber-skip weights and LTS weights, for slowly converging initial estimators, and for a 1 or 2-sided cutoff c that may vary with n.

Sketch of the proof. First, the errors $n^{1/2}(\tilde{\sigma}^2 - \sigma^2)$ and $n^{1/2}(\tilde{\beta} - \beta)$ are bounded in probability, such that $n^{1/2}(\tilde{\sigma}^2 - \sigma^2) = 2\sigma n^{1/2}(\tilde{\sigma} - \sigma) + \mathrm{o}_{\mathsf{P}}(1)$. Thus, we replace $n^{1/2}(\tilde{\sigma} - \sigma)$, $n^{1/2}(\tilde{\beta} - \beta)$ and v_i with deterministic terms a, b, and $v_i^{abc} = 1_{\{|\varepsilon_i - n^{-1/2} x_i' b| \leq \sigma(1 + n^{-1/2} a)c\}}$, such that $v_i^{00c} = 1_{\{|\varepsilon_i| \leq \sigma c\}}$. The desired result follows if we show that

$$n^{-1/2} \sum_{i=1}^{n} v_i^{abc} - n^{-1/2} \sum_{i=1}^{n} v_i^{00c} = 2c\mathsf{f}(c)a + \mathrm{o}_{\mathsf{P}}(1), \tag{22}$$

uniformly for $|a|, |b| < B$ for some large B.

Second, we introduce the empirical process

$$\mathbb{G}_n(a, b, c) = n^{-1/2} \sum_{i=1}^{n} (v_i^{abc} - \mathsf{E}_{i-1} v_i^{abc}) \tag{23}$$

and decompose the left hand side of (22) as follows

$$n^{-1/2} \sum_{i=1}^{n} (v_i^{abc} - v_i^{00c}) = \mathbb{G}_n(a, b, c) - \mathbb{G}_n(0, 0, c) + n^{-1/2} \sum_{i=1}^{n} (\mathsf{E}_{i-1} v_i^{abc} - \mathsf{E}_{i-1} v_i^{00c}).$$

Third, we prove (22), and in turn (21), by showing that, uniformly in $|a|, |b| < B$,

$$n^{-1/2} \sum_{i=1}^{n} \{\mathsf{E}_{i-1} v_i(a, b, c) - \mathsf{E}_{i-1} v_i(0, 0, c)\} = 2c\mathsf{f}(c)a + \mathrm{o}_{\mathsf{P}}(1), \tag{24}$$

$$\mathbb{G}_n(a, b, c) - \mathbb{G}_n(0, 0, c) = \mathrm{o}_{\mathsf{P}}(1). \tag{25}$$

The expansion (24) is an application of the mean value theorem with an additional argument about uniformity in a, b. For this we exploit that, in contrast to the indicators, the expectation is a smooth function of a, b.

Finally, the expansion (25) is a statement of stochastic equicontinuity. We apply a chaining argument using the iterated martingale inequality given below. The idea is to cover

the compact interval $|a|, |b| < B$ with grid points. We can then study the variation across grid points and the variation within the small rectangles defined by the (a, b)-grid point using the iterated exponential martingale inequality outlined below.

In order to apply (21) to derive limit results, we have to prove that the process $\mathbb{G}_n(0, 0, c)$ is tight and converges to a Gaussian process.

An iterated martingale inequality. We are interested in the tail behaviour of the martingale $\mathbb{G}_n(a, b, c)$ in (23). This is an unbounded martingale. In the case discussed above we need to analyze the tail behaviour of the maximal value of $\mathbb{G}_n(a, b, c)$ over the grid points in a, b. This is done using the following iterated exponential martingale inequality.

Theorem 1 *(Johansen and Nielsen, 2015a, Theorem 5.2.) For $\ell = 1, \ldots, L$ let $z_{\ell,i}$ be $\mathcal{F}_i$-adapted so $\mathsf{E} z_{\ell,i}^{2^{\overline{r}}} < \infty$ for some $\overline{r} \in \mathbb{N}$. Let $D_r = \max_{1 \le \ell \le L} \sum_{i=1}^{n} \mathsf{E}(z_{\ell,i}^{2^r} | \mathcal{F}_{i-1})$ for $1 \le r \le \overline{r}$. Then, for all $\kappa_0, \kappa_1, \ldots, \kappa_{\overline{r}} > 0$,*

$$\mathsf{P}[\max_{1 \le \ell \le L} |\sum_{i=1}^{n} \{z_{\ell,i} - \mathsf{E}(z_{\ell,i} | \mathcal{F}_{i-1})\}| > \kappa_0] \le L \frac{\mathsf{E} D_{\overline{r}}}{\kappa_{\overline{r}}} + \sum_{r=1}^{\overline{r}} \frac{\mathsf{E} D_r}{\kappa_r} + 2L \sum_{r=0}^{\overline{r}-1} \exp(-\frac{\kappa_r^2}{14 \kappa_{r+1}}).$$

The bound in Theorem 1 involves the expectation of a variable D_r, which is the maximum of the quadratic variations. It can be used with advantage when D_r has a simple bound. The inequality is proved by iterating the exponential martingale inequality in Theorem 2.1 of Bercu & Touati (2008). It does not require the martingale difference sequences to be bounded, and so it can be used for analyzing the unbounded product moments (20). We have some flexibility in choosing the parameters $\kappa_0, \kappa_1, \ldots, \kappa_{\overline{r}}$. This is exploiting in different ways when showing that $\mathbb{G}_n(a, b, c) - \mathbb{G}_n(0, 0, c)$ vanishes and that $\mathbb{G}_n(0, 0, c)$ is tight.

5.2 Assumptions on density and regressors

For this presentation we assume a normal reference distribution, which, of course, is most used in practice, but formulate the results for more general symmetric densities. With normality we avoid a somewhat tedious discussion of existence of moments. The regressors can be temporally dependent and possibly deterministically or stochastically trending.

The minimal density assumption for the results presented is a symmetric density with derivative satisfying boundedness and tail conditions. This includes t-distributions, see Johansen and Nielsen (2015a) for a general discussion, and Johansen and Nielsen (2009, 2013, 2015a,b) for 1-step Huber-skip, the iterated 1-step Huber-skip, and for the Forward Search and for general M-estimators, respectively. Symmetry is natural when concerned with data generating processes without outliers. Non-symmetric densities are also discussed in Johansen and Nielsen (2015a,b). These arise when discussing data generating processes with outliers, but we leave this for future work.

Assumption 1 *Let $\mathcal{F}_i$ be the filtration generated by $x_1, \ldots, x_{i+1}$ and $\varepsilon_1, \ldots, \varepsilon_i$. Assume*
(i) innovations ε_i/σ are independent of $\mathcal{F}_{i-1}$ and standard normal;
(ii) regressors x_i satisfy, for some non-stochastic normalisation matrix $N \to 0$ and random matrices V, Σ, μ, the following joint convergence results
(a) $V_n = N' \sum_{i=1}^{n} x_i \varepsilon_i \xrightarrow{\mathsf{D}} V$;

(*b*) $\Sigma_n = N'\sum_{i=1}^n x_i x_i' N \xrightarrow{\mathsf{D}} \Sigma \overset{a.s.}{>} 0$;
(*c*) $n^{-1/2}N'\sum_{i=1}^n x_i \xrightarrow{\mathsf{D}} \mu$;
(*d*) $\max_{i\le n} |n^{1/2}N'x_i| = \mathrm{o}_{\mathsf{P}}(n^{\phi})$ *for all* $\phi > 0$;
(*e*) $n^{-1}\mathsf{E}\sum_{i=1}^n |n^{1/2}N'x_i|^q = \mathrm{O}(1)$ *for some* $q > 9$.

Assumption 1(*ii*) for the regressors is satisfied in a range of situations, see Johansen and Nielsen (2009). For instance, x_i could be vector autoregressive with stationary roots or roots at one. It could also include deterministically trending regressors. The normalisation is $N = n^{-1/2}I_{\dim x}$ for stationary regressors and $N = n^{-1}I_{\dim x}$ for random walk regressors.

6 Asymptotic results for Huber-skip estimators

This section contains the asymptotic properties of the Huber-skip, and the 1-step and m-step Huber-skip defined in Algorithm 1, with applications to the robustified least squares. The result is formulated as a stochastic expansion of the updated estimation error in terms of a kernel and the original estimation error, and the proof is as outlined in Section 5.1.

The Huber-skip estimator is the minimizer of the criterion (11), where the scale σ is known. Since this minimization problem is non-convex, we need an additional assumption that bounds the frequency of small regressors.

Assumption 2 *Define* $z_{ni} = n^{1/2}N'x_i$ *and*

$$F_n(a) = \sup_{|\delta|=1} F_{n\delta}(a) = \sup_{|\delta|=1} n^{-1}\sum_{i=1}^n 1_{(|z_{ni}'\delta|\le a)}. \tag{26}$$

(*a*) *Assume* $n^{-1}\mathsf{E}\sum_{i=1}^n |n^{1/2}N'x_i|^{2^{r+1}} = \mathrm{O}(1)$ *where* $2^{r+1} > 2(\dim x + 2)$.
(*b*) *Assume, for* $(a, n) \to (0, \infty)$, *that*

$$\sup_{|\delta|=1}\{F_{n\delta}(a) - F_{n\delta}(0)\} \xrightarrow{\mathsf{P}} 0, \tag{27}$$

and there exists $0 \le \xi < 1$ *and* $n_0 > 0$, *such that for all* $\epsilon > 0$ *and all* $n \ge n_0$

$$\mathsf{P}\{F_n(0) \le \xi\} \ge 1 - \epsilon. \tag{28}$$

Theorem 2 *(Johansen and Nielsen, 2015b, Theorems 1,2,3) Consider the Huber-skip defined as the minimizer of (11). Suppose Assumptions 1, 2 are satisfied and that* f *is bounded and* $\dot{\mathsf{f}}$ *exists. Then any minimizer of the objective function (11) has a measurable version and satisfies*

$$N^{-1}(\hat\beta - \beta) = \frac{1}{\psi - 2c\mathsf{f}(c)}\Sigma_n^{-1}N'\sum_{i=1}^n x_i\varepsilon_i 1_{(|\varepsilon_i|\le\sigma c)} + \mathrm{o}_{\mathsf{P}}(1). \tag{29}$$

A condition similar to Assumption 2 was introduced by Davies (1990) in the analysis of S-estimators with deterministic regressors. He defined

$$\lambda_n(\xi) = \min_{|S|=\text{int}(n\xi)} \min_{|\delta|=1} \max_{i\in S} |z'_{ni}\delta|, \tag{30}$$

where S are subsets of the indices $i = 1, \ldots, n$. The function $\lambda_n(\xi)$ is related to $F_n(a)$ by

$$\{F_n(a) > \text{int}(n\xi)/n\} \subset \{\lambda_n(\xi) \leq a\} \subset \{F_n(a) \geq \text{int}(n\xi)/n\},$$

and is thus an approximative inverse of $F_n(a)$. Chen and Wu (1988) show, when $\mu_\rho = 0$, that for deterministic regressors $\hat{\beta} \xrightarrow{a.s.} \beta_0$ if $F_n(a) \to 0$ as $(a, n) \to (0, \infty)$.

The conditions (27) and (28) are satisfied for some stationary and non-stationary regressors. The condition is used to prove that the objective function is uniformly bounded below for large values of the parameter, a property that implies existence and tightness of the estimator. For full descriptions of the bound to the regressors and extensions to a wider class of M-estimators, see Johansen and Nielsen (2015b).

Theorem 2 was conjectured by Huber (1964) for pure location problems. The regularity conditions on the regressors are much weaker than those normally considered in for instance Chen & Wu (1988), Liese & Vajda (1994), Maronna *et al.* (2006), Huber & Ronchetti (2009), and Jurečková & Sen (1996). Theorem 2 extends to non-normal, but symmetric densities and even to non-symmetric densities and objective functions, by introducing a bias correction for the constant term in the regression.

<Fig3 here>

From Theorem 2 we can derive the asymptotic distribution of the estimator for specific types of regressors. For stationary regressors we can apply the Central Limit Theorem to the expansion (29), see Lemma 2, to get

$$n^{1/2}(\hat{\beta} - \beta) \xrightarrow{\mathsf{D}} \mathsf{N}[0, \Sigma^{-1}\frac{\tau\sigma^2}{\{\psi - 2c\mathsf{f}(c)\}^2}]. \tag{31}$$

With a normal reference distribution, $\tau = \psi - 2c\mathsf{f}(c)$ by (8), so that

$$n^{1/2}(\hat{\beta} - \beta) \xrightarrow{\mathsf{D}} \mathsf{N}\{0, \Sigma^{-1}\frac{\sigma^2}{\psi - 2c\mathsf{f}(c)}\}. \tag{32}$$

The efficiency $1/\eta_\beta = \psi - 2c\mathsf{f}(c)$ relative to the least squares estimator is shown in Figure 3. In the situation with deterministically trending regressors, for instance $x_i = i$, the normalisation N does not reduce to $n^{1/2}$, but the same limiting distribution applies. If, instead, the regressor x_i is a random walk the limiting distribution is non-normal. Suppose x_i converges jointly with the partial sum of the truncated innovation to a Brownian motion

$$n^{-1/2}\begin{Bmatrix} x_{[nu]} \\ \sum_{i=1}^{[nu]} \varepsilon_i 1_{(|\varepsilon_i| \leq \sigma c)} \end{Bmatrix} \xrightarrow{\mathsf{D}} \begin{Bmatrix} W_x(u) \\ W_\varepsilon(u) \end{Bmatrix}, \quad \mathsf{Var}\begin{Bmatrix} W_x \\ W_\varepsilon \end{Bmatrix} = \begin{pmatrix} \Sigma_x & \delta \\ \delta' & \sigma^2\tau \end{pmatrix},$$

on the space $D[0,1]^{1+\dim x}$ of right continuous functions with limits from the left. Here $\delta = \tau \mathsf{Cov}(\varepsilon_i, x_i - x_{i-1})$ when the innovations are normal. Following Johansen and Nielsen (2009) we then get

$$n(\hat{\beta} - \beta) \xrightarrow{\mathsf{D}} \frac{1}{\psi - 2c\mathsf{f}(c)} (\int_0^1 W_x W_x' du)^{-1} \int_0^1 W_x (dW_\varepsilon). \tag{33}$$

The 1-step Huber-skip estimator. The estimation error for this estimator can be expanded in terms of the estimation error of the original estimator. This was done in Welsh & Ronchetti (2002), without proof, and in Johansen and Nielsen (2009).

Theorem 3 *Consider the 1-step Huber-skip, $\hat{\beta}, \hat{\sigma}$, see Definition 1. Suppose Assumption 1 holds and that the initial estimators $N^{-1}(\tilde{\beta} - \beta)$ and $n^{1/2}(\tilde{\sigma} - \sigma)$ are $\mathrm{O}_\mathsf{P}(1)$. Recall the coefficients ψ, τ from (7),(8) and define*

$$\rho_\beta = 2c\mathsf{f}(c)/\psi, \qquad \rho_\sigma = (c^2 - \tau/\psi)c\mathsf{f}(c)/\tau. \tag{34}$$

Then the $\hat{\beta}, \hat{\sigma}$ satisfy

$$N^{-1}(\hat{\beta} - \beta) = \rho_\beta N^{-1}(\tilde{\beta} - \beta) + \frac{1}{\psi} \Sigma_n^{-1} N' \sum_{i=1}^n x_i \varepsilon_i 1_{(|\varepsilon_i| \le \sigma c)} + \mathrm{o}_\mathsf{P}(1), \tag{35}$$

$$n^{1/2}\{\hat{\sigma}^2 - \sigma^2\} = \rho_\sigma n^{1/2}(\tilde{\sigma}^2 - \sigma^2) + \frac{1}{\tau}\sigma^2 n^{-1/2} \sum_{i=1}^n (\frac{\varepsilon_i^2}{\sigma^2} - \frac{\tau}{\psi}) 1_{(|\varepsilon_i| \le \sigma c)} + \mathrm{o}_\mathsf{P}(1). \tag{36}$$

Theorem 3 shows that the updated regression estimator, $\hat{\beta}$, only depends on the initial regression estimator $\tilde{\beta}$ and not on the initial scale estimator $\tilde{\sigma}$. This is a consequence of the symmetry imposed on the problem. Johansen and Nielsen (2009) also analyze situations, where the reference distribution, f, is non-symmetric and the cutoff is made in a matching non-symmetric way. In that situation, both expansions involve the initial estimation uncertainty for $\tilde{\beta}$ and $\tilde{\sigma}^2$. The assumption that the normalised initial estimators are bounded in probability can be relaxed using the techniques of Johansen and Nielsen (2015a).

The m-step Huber-skip estimator is a finite iteration of the 1-step estimator. The expansions in Theorem 3 can be iterated a finite number of times without difficulty since the combination of a finite number of remainder terms of order $\mathrm{o}_\mathsf{P}(1)$ remains $\mathrm{o}_\mathsf{P}(1)$.

m-step robustified least squares. A situation of special interest is when the initial estimators are the full sample least squares estimators. These have the expansion

$$N^{-1}(\tilde{\beta} - \beta) = \Sigma_n^{-1} N' \sum_{i=1}^n x_i \varepsilon_i, \qquad n^{1/2}(\tilde{\sigma}^2 - \sigma^2) = n^{-1/2} \sum_{i=1}^n (\varepsilon_i^2 - \sigma^2) + \mathrm{o}_\mathsf{P}(1),$$

see also Johansen and Nielsen (2009, Theorems 1.3, 1.7). We then have the following result.

Corollary 1 *Consider m-step robustified least squares, see Algorithm 1, with full sample least squares as initial estimators. Introduce* $\eta_\beta^{(0)} = \eta_\sigma^{(0)} = 1$ *and, for* $m = 1, 2, \ldots,$

$$\eta_\beta^{(m)} = [\{\frac{1-\rho_\beta^m}{(1-\rho_\beta)\psi}\}^2 + 2\{\frac{1-\rho_\beta^m}{(1-\rho_\beta)\psi}\}\rho_\beta^m]\tau + \rho_\beta^{2m}, \tag{37}$$

$$\eta_\sigma^{(m)} = [\{\frac{1-\rho_\sigma^m}{(1-\rho_\sigma)\tau}\}^2 + 2\{\frac{1-\rho_\sigma^m}{(1-\rho_\sigma)\tau}\}\rho_\sigma^m]\frac{\varkappa - \tau^2/\psi}{\kappa - 1} + \rho_\sigma^{2m}, \tag{38}$$

using the coefficients $\tau, \varkappa$ *from (8),* ρ_β, ρ_σ *from (34), and* $\kappa = \mathsf{E}(\varepsilon_1/\sigma)^4$. *Then, under the assumption of Theorem 3, we find for stationary regressors*

$$n^{1/2}\begin{pmatrix} \hat\beta - \beta \\ \hat\sigma^2 - \sigma^2 \end{pmatrix} \xrightarrow{\mathsf{D}} \mathsf{N}\left\{0, \begin{pmatrix} \sigma^2\eta_\beta^{(m)}\Sigma^{-1} & 0 \\ 0 & \sigma^4\eta_\sigma^{(m)}(\kappa-1) \end{pmatrix}\right\}. \tag{39}$$

Note, however, that the asymptotic distribution of $n^{1/2}(\hat\sigma^2 - \sigma^2)$ *is valid for general regressors.*

The efficiencies $1/\eta_\beta^{(1)}, 1/\eta_\beta^{(2)}$ are plotted as the top curves in Figure 3. Johansen and Nielsen (2009, Figure 1.1) plot the efficiency for the variance, $1/\eta_\sigma^{(1)}$.

Impulse indicator saturation. This estimator has the same distribution as the 1-step robustified least squares estimator in the situation without outliers and stationary regressors, see Theorems 1.5, 1.7 in Johansen and Nielsen (2009).

Infinite iteration of 1-step Huber-skip estimators. A necessary condition for convergence of infinite iteration of mappings (35), (36) is that these are contractions. Johansen and Nielsen (2013) prove this for a range of unimodal distributions. Moreover, the combination of the infinitely many remainder terms of order $\mathsf{o_P}(1)$ remains $\mathsf{o_P}(1)$. The fixed point for the regression estimator has the same expansion as the Huber-skip estimator in Theorem 2.

7 Asymptotic results for LTS type estimators

The LTS-type estimators have a cutoff determined from the order statistics of the absolute residuals as opposed to the fixed cutoff for the Huber-skip estimators. The asymptotic results appear to be the same, but the argument to get there is a bit more convoluted because of the quantiles involved. We give an overview of the result for the 1-step LTS, and its consequences for LTS and for the Forward Search.

7.1 LTS type estimators

The LTS estimator has the same asymptotic expansion (29) as the Huber-skip.

Theorem 4 *(Víšek, 2006c, Theorem 1) Consider the LTS estimator* $\hat\beta_{LTS}$ *defined as minimizer of (12). Suppose Assumption 1 holds, that the regressors are fixed, and that their empirical distribution can be suitably approximated by a continuous distribution function, see Víšek (2006c) for details. Then*

$$N^{-1}(\hat\beta_{LTS} - \beta) = \frac{1}{\psi - 2c\mathsf{f}(c)}\Sigma_n^{-1}N'\sum_{i=1}^n x_i\varepsilon_i 1_{(|\varepsilon_i| \le \sigma c)} + \mathsf{o_P}(1). \tag{40}$$

The 1-step LTS estimator satisfies an expansion that is similar to the 1-step Huber-skip, albeit with a slight difference in the expansion for the variance estimator. The reason is that the LTS is based on quantiles rather than an initial scale estimator. The proof is given in the Appendix. We refer to Johansen and Nielsen (2015a) for minimal conditions and statement of uniformity in ψ. Ruppert & Carroll (1980) state a similar result for a related estimator, but omit the details of the proof.

Theorem 5 *Consider 1-step LTS with a cutoff $k = [\psi n]$, see Definition 2. Suppose Assumption 1 holds, and that the initial estimator $N^{-1}(\tilde{\beta} - \beta)$ is $\mathrm{O_P}(1)$. Recall the coefficients ψ, τ from (7),(8) and ρ_β, ρ_σ from (34). Then*

$$N^{-1}(\hat{\beta} - \beta) = \rho_\beta N^{-1}(\tilde{\beta} - \beta) + \frac{1}{\psi}\Sigma_n^{-1} N \sum_{i=1}^{n} x_i \varepsilon_i 1_{(|\varepsilon_i| \leq \sigma c)} + \mathrm{o_P}(1), \tag{41}$$

$$n^{1/2}(\hat{\sigma}^2 - \sigma^2) = \rho_\sigma \frac{2\sigma^2}{c} n^{1/2}(\frac{\hat{z}_\psi}{\sigma} - c) + \frac{1}{\tau}\sigma^2 n^{-1/2} \sum_{i=1}^{n} (\frac{\varepsilon_i^2}{\sigma^2} - \frac{\tau}{\psi}) 1_{(|\varepsilon_i| \leq \sigma c)} + \mathrm{o_P}(1), \tag{42}$$

where the quantile $\hat{z}_\psi$ satisfies

$$2\mathsf{f}(c) n^{1/2}(\frac{\hat{z}_\psi}{\sigma} - c) = -n^{-1/2} \sum_{i=1}^{n} \{1_{(|\varepsilon_i| \leq \sigma c)} - \psi\} + \mathrm{o_P}(1). \tag{43}$$

Asymptotic variances are reported in Lemma 2.

The expansion (43) for the quantile is similar to the Bahadur (1966) representation, see also Section 6 of Csörgő (1983), which links the empirical distribution function with empirical quantiles. In other words, the quantile $\hat{\xi}_{(k)}$, that is computed in a complicated way, has the same asymptotic behaviour as the kth order statistic of the absolute errors $|\varepsilon_i|$.

The LTS scale estimator is the consistency corrected minimum of the LTS criterion function (12), see Croux & Rousseeuw (1992). This estimator is a 1-step LTS estimator with the LTS estimator as initial estimator. Theorem 5 has the following Corollary.

Corollary 2 *Let $\tilde{\beta}$ be the LTS estimator with a cutoff $k > 0$, corresponding to a breakdown point $1 - k/n$. Apply the 1-step LTS to get the LTS scale estimator $\hat{\sigma}$. Suppose the Assumptions of Theorems 4, 5 hold. Then $\hat{\sigma}^2$ has expansion (42).*

7.2 Forward Search

The Forward Search is an iterated 1-step LTS, where the cutoff changes slightly in each step. We highlight asymptotic expansions for the forward regression estimators $\hat{\beta}^{(m)}$ and for the scaled forward residuals $\hat{z}^{(m)}/\hat{\sigma}^{(m)}$. The results are formulated in terms of embeddings of the time series $\hat{\beta}^{(m)}, \hat{\sigma}^{(m)}, \hat{z}^{(m)}$ for $m = m_0 + 1, \ldots, n$ into the space $D[0,1]$ of right continuous functions with limits from the left. As an example consider $\hat{\beta}^{(m)}$:

$$\hat{\beta}_\psi = \begin{cases} \hat{\beta}^{(m)} & \text{for } m = \text{integer}(n\psi) \text{ and } \psi_0 = m_0/n \leq \psi \leq 1, \\ 0 & \text{otherwise.} \end{cases}$$

Theorem 6 *(Johansen and Nielsen, 2015a, Theorems 3.1, 3.2, 3.4, 3.5) Consider the Forward Search in Algorithm 2. Suppose Assumption 1 holds and that* $N^{-1}(\hat{\beta}^{(m_0)} - \beta)$ *is* $\mathrm{O_P}(1)$. *Let* $\psi_1 > \psi_0 > 0$. *Then, as processes in* $\psi = \mathsf{G}(c_\psi)$,
(i) $\hat{\beta}_\psi$ *has the same expansion as the LTS (40), uniformly in* $\psi_1 \leq \psi \leq 1$;
(ii) $\hat{\sigma}_\psi, \hat{z}_\psi$ *have the same expansions as the 1-step LTS estimators* $\hat{\sigma}, \hat{\xi}_k$ *in (42), (43), uniformly in* $\psi_0 \leq \psi \leq n/(n+1)$;
(iii) $\hat{d}_\psi$ *has the same expansion as* $\hat{\xi}_k$ *in (43), uniformly in* $\psi_1 \leq \psi \leq n/(n+1)$.

The idea of the Forward Search is to monitor the plot of the sequence of scaled forward residuals. The expansions for $\hat{\sigma}_\psi$ and $\hat{z}_\psi$ in Theorem 6 combine as follows.

Corollary 3 *(Johansen and Nielsen 2015a, Theorem 3.3). Consider the Forward Search-estimator in Algorithm 2. Suppose Assumption 1 holds and that* $N^{-1}(\hat{\beta}^{(m_0)} - \beta)$ *is* $\mathrm{O_P}(1)$. *Then*

$$\mathbb{Z}_n(\psi) = 2\mathsf{f}(c_\psi)n^{1/2}(\frac{\hat{\tilde{z}}_\psi}{\hat{\sigma}_\psi} - c_\psi) = 2\mathsf{f}(c_\psi)n^{1/2}(\frac{\hat{\tilde{z}}_\psi}{\sigma} - c_\psi) - \frac{c_\psi \mathsf{f}(c_\psi)}{\sigma^2} n^{1/2}(\hat{\sigma}^2_\psi - \sigma^2) + \mathrm{o_P}(1), \quad (44)$$

uniformly in $\psi_0 \leq \psi \leq n/(n+1)$. *Here,* $\mathbb{Z}_n(\psi)$ *converges on* $D[\psi_0, 1]$ *to a Gaussian process* $\mathbb{Z}(\psi)$ *with variance given in Lemma 2.*

Part II
Gauge as a measure of false detection

We now explore the gauge as a means of controlling the cutoff in outlier detection algorithms, when in fact there are no outliers. The gauge therefore controls errors of the first type. When there are undetected outliers, we get errors of the second kind. We leave this discussion of the influence of outliers to future work. Proofs follow in the appendix.

The empirical gauge was defined, see (2), as the fraction of detected outliers. We show that the population gauge for 1-step Huber-skip outlier detection relates in a simple way to the size of an underlying statistical test. In general, the population gauge will, however, be of a more complicated nature. An example is the Forward Search, see Section 9.

<Table 1 here>

The gauge concept is related to, but also distinct from the false discovery rate for multiple tests of Benjamini & Hochberg (1995). To illustrate this, Table 1 shows the number of errors when testing m hypotheses. The false discovery rate is concerned with those tests that reject the hypothesis. Suppose there are R of those, of which V are false rejections. Then the false discovery rate is defined as $\mathsf{E}(V/R)$. The gauge is concerned with those observations that are not outliers. In our setting there are no outliers so $m = m_0 = n$. Out of these, V observations are falsely declared outliers. Then the gauge is $\mathsf{E}(V/m_0)$.

8 The gauge of Huber-skip estimators

We derive an asymptotic theory for the gauge of outlier detection methods based on estimators of the Huber-skip type. We consider initial estimators $(\tilde{\beta}, \tilde{\sigma})$ so that observations are classified as outliers if the absolute residuals $|y_i - x_i'\tilde{\beta}|/\tilde{\sigma}$ are large. The empirical gauge is

$$\hat{\gamma} = \frac{1}{n}\sum_{i=1}^{n} 1_{(|y_i - x_i'\tilde{\beta}| > \tilde{\sigma}c)}. \tag{45}$$

We prove below that $\mathsf{E}\hat{\gamma} \to \mathsf{P}(|\varepsilon_1| > \sigma c) = 1 - \psi$. This probability equals $\mathsf{P}(|y_1 - x_1'\beta| > \sigma c)$, which is the size of a test for the hypothesis that the first observation is an outlier when the parameters β, σ are known. Further, we show that for a given ψ, the empirical gauge is asymptotically normal around $\gamma = 1 - \psi$. If, instead, we fix the expected number of incorrectly determined outliers, $n\gamma = n(1 - \psi) = \lambda$, then $n\hat{\gamma}$ is asymptotically Poisson.

8.1 Normal approximations to gauge

The first result is an asymptotic expansion of the sample gauge $\hat{\gamma}$, using (21).

Theorem 7 *Consider the sample gauge $\hat{\gamma}$ of the form (45). Suppose Assumption 1 is satisfied and that $N^{-1}(\tilde{\beta} - \beta)$, $n^{1/2}(\tilde{\sigma}^2 - \sigma^2)$ are $\mathrm{O}_{\mathsf{P}}(1)$. Then, for fixed $\psi = \mathsf{P}(|\varepsilon_1| \leq \sigma c)$,*

$$n^{1/2}\{\hat{\gamma} - (1 - \psi)\} = n^{-1/2}\sum_{i=1}^{n}\{1_{(|\varepsilon_i| > \sigma c)} - (1 - \psi)\} - \frac{c\mathsf{f}(c)}{\sigma^2} n^{1/2}(\tilde{\sigma}^2 - \sigma^2) + \mathrm{o}_{\mathsf{P}}(1). \tag{46}$$

It follows that $\hat{\gamma} \to 1 - \psi$ in mean, such that the population gauge is $\gamma = 1 - \psi$.

The result in Theorem 7 does not depend on the type of regressors. We can apply it to the range of Huber-skip estimators and derive an asymptotically normal distribution theory. The asymptotic variance is analyzed case by case since the expansion in Theorem 7 depends on the variance of the initial estimator $\tilde{\sigma}^2$.

The Huber-skip. Theorem 2 shows that $N^{-1}(\hat{\beta} - \beta)$ is tight. In this case the variance is assumed known, $\hat{\sigma}^2 = \sigma^2$ and therefore only the first binomial term in Theorem 7 matters.

Corollary 4 *Let $\tilde{\beta}$ be the Huber-skip estimator $\hat{\beta}$ while $\tilde{\sigma} = \sigma$ is known. Suppose Assumptions 1, 2 are satisfied and $\psi = \mathsf{P}(|\varepsilon_1| \leq \sigma c)$, then*

$$n^{1/2}\{\hat{\gamma} - (1 - \psi)\} \xrightarrow{\mathsf{D}} \mathsf{N}\{0, \psi(1 - \psi)\}.$$

m-step robustified least squares has a similar expression. But now the estimation of the variance contributes to the asymptotic distribution.

Corollary 5 *Let $\tilde{\beta} = \hat{\beta}^{(m)}, \tilde{\sigma} = \hat{\sigma}^{(m)}$ be the m-step robustified least squares estimators, see Algorithm 1, with full sample least squares as initial estimators. Suppose Assumption 1 is satisfied. Then, for fixed $\psi = \mathsf{P}(|\varepsilon_1| \leq \sigma c)$,*

$$n^{1/2}\{\hat{\gamma} - (1 - \psi)\} \xrightarrow{\mathsf{D}} \mathsf{N}\{0, \psi(1 - \psi) + \eta_\gamma^{(m)}\},$$

where, with τ, ρ_σ, $\eta_\sigma^{(m)}$ defined in (8), (34), (38), $\kappa = \mathsf{E}(\varepsilon_1/\sigma)^4$,

$$\eta_\gamma^{(m)} = \{c\mathsf{f}(c)\}^2 \eta_\sigma^{(m-1)}(\kappa - 1) + 2c\mathsf{f}(c)\rho_\sigma^{m-1}(\tau - \psi). \tag{47}$$

Impulse indicator saturation. The gauge for the outliers detected by the original split half least squares estimators is the same as for the robustified least squares, when the regressors are stationary. Details are given in Theorem 9.4 of Johansen and Nielsen (2014).

Infinite iteration of 1-step Huber-skip estimators. Theorem 9.5 of Johansen and Nielsen (2014) show that the gauge of the m-step estimator converges to γ uniformly in m. Their Theorem 9.6 provides a fixed point result for the gauge of the fully iterated estimator.

<Table 2 here>

Numerical comparison of gauges Table 2 shows the asymptotic standard deviations of $\hat{\gamma}$ for the Huber-skip and the robustified least squares for $m = 1$ taken from Corollaries 4, 5, respectively. The Table also shows results for the fixed point of the fully iterated Huber-skip where the variance correction equals $\lim_{m\to\infty} \eta_\gamma^{(m)} = (\varkappa - \tau^2/\psi)\{c\mathsf{f}(c)\}^2/\{(1-\rho_\sigma)\tau\}^2$. For gauges of 1% or lower, the standard deviations are very similar. If the gauge is chosen as $\gamma = 0.05$ and $n = 100$, then the sample gauges $\hat{\gamma}$ will be asymptotically normal with mean $\gamma = 0.05$ and a standard deviation of about $0.2/n^{1/2} = 0.02$. This suggests that it is not unusual to find up to 8-9 outliers when in fact there are none. Lowering the gauge to $\gamma = 0.01$ or $\gamma = 0.0025$, the standard deviation is about $0.1/n^{1/2} = 0.01$ and $0.05/n^{1/2} = 0.005$, respectively, when $n = 100$. Thus, it is not unusual to find up to 2-3 and up to 1 outliers, respectively, when in fact there are none. This suggests that the gauge should be chosen rather small in line with the discussion in Section 7.6 of Hendry & Doornik (2014).

8.2 Poisson approximation to gauge

If we set the cutoff so as to accept the same fixed *number* of falsely discovered outliers regardless of the sample size, then a Poisson exceedance theory arises. The idea is to choose the cutoff c_n so that, for some $\lambda > 0$,

$$\mathsf{P}(|\varepsilon_i| > \sigma c_n) = \lambda/n. \tag{48}$$

The cutoff c_n appears both in the definition of the gauge and in the definition of the estimators, so some care is needed. We analyze again the m-step Huber-skip. Let $\tilde{\beta}_n$ and $\tilde{\sigma}_n$ be sequences of initial estimators that may depend on c_n, hence the subscript n in the notation for the estimators. We assume that the estimation errors $N^{-1}(\tilde{\beta}_n - \beta)$ and $n^{1/2}(\tilde{\sigma}_n - \sigma)$ are tight. Thus, the result immediately applies to robustified least squares, where the initial estimators $\tilde{\beta}_n$ and $\tilde{\sigma}_n$ are the full sample least squares estimators, which do not depend on the cutoff c_n. But, in general we need to check this tightness condition. We choose to prove the result assuming a more general density function and replace Assumption 1(i) by the following assumption, which is satisfied for the normal distribution, see Remark 2.

Assumption 1(i') *The innovations ε_i/σ are independent of $\mathcal{F}_{i-1}$, and the density f is symmetric with decreasing tails and support on R so that $c_n \to \infty$ and*

(a) $\mathsf{E}|\varepsilon_i|^r < \infty$ *for some* $r > 4$;
(b) $\mathsf{f}(c_n)/[c_n\{1 - \mathsf{F}(c_n)\}] = \mathrm{O}(1)$;
(c) $\mathsf{f}(c_n - n^{-1/4}A)/\mathsf{f}(c_n) = \mathrm{O}(1)$ *for all* $A > 0$;

Theorem 8 *Consider m-step Huber-skip, see Algorithm 1, where c_n is given by (48). If the initial estimation errors $N^{-1}(\tilde{\beta}-\beta), n^{1/2}(\tilde{\sigma}^2-\sigma^2)$ are $\mathrm{O_P}(1)$ and Assumption 1(i′, ii) is satisfied, then the m-step Huber-skip estimators, $\hat{\beta}_n^{(m)}, \hat{\sigma}_n^{(m)}$ say, have the same asymptotic properties as the full sample least squares estimators $\hat{\beta}_{LS}, \hat{\sigma}_{LS}^2$:*

$$N^{-1}(\hat{\beta}_n^{(m)}-\hat{\beta}_{LS}), n^{1/2}(\hat{\sigma}_n^{(m)}-\hat{\sigma}_{LS}) = \mathrm{o_P}(1), \tag{49}$$

and the sample gauge $\hat{\gamma}$ in (45) satisfies

$$n\hat{\gamma} \xrightarrow{\mathsf{D}} \mathsf{Poisson}(\lambda). \tag{50}$$

<Table 3 here>

Table 3 shows the Poisson approximation to the probability of finding at most x outliers for different values of λ. For small λ and n this approximation is possibly more accurate than the normal approximation, although that would have to be investigated in a detailed simulation study. The Poisson distribution is left skew so the probability of finding at most $x = \lambda$ outliers increases from 62% to 90% for λ decreasing from 5 to 0.1. In particular, for $\lambda = 1$, $n = 100$ so the cutoff is $c_n = 2.58$, the probability of finding at most one outlier is 74% and the probability of finding at most two outliers is 92%. In other words, the chance of finding 3 or more outliers is small when in fact there are none.

9 The gauge of the Forward Search

We now consider the gauge for the Forward Search. The forward plot consists of the scaled forward residuals $\hat{z}^{(m)}/\hat{\sigma}^{(m)}$ for $m = m_0, \ldots, n-1$. Along with this, we plot point-wise confidence bands derived from Theorem 3. The idea is to stop the algorithm once the scaled forward residuals exceed a suitable quantile. We choose the quantile from the gauge.

Consider a stopping time $\hat{m}$ based on this information, so that $n-\hat{m}$ is the number of the outliers. The sample gauge (2) then simplifies as

$$\hat{\gamma} = \frac{n-\hat{m}}{n} = \frac{1}{n}\sum_{m=m_0}^{n-1}(n-m)1_{(\hat{m}=m)} = \frac{1}{n}\sum_{m=m_0}^{n-1}1_{(\hat{m}\le j)}, \tag{51}$$

by substituting $n-m = \sum_{j=m}^{n-1} 1$ and changing summation order. If the stopping time is an exit time, then the event $(\hat{m} \le j)$ is true if $\hat{z}^{(m)}/\hat{\sigma}^{(m)}$ has exited at the latest by $m = j$.

An example of a stopping time is the following. Theorem 3 shows that

$$\mathbb{Z}_n(\psi) = 2\mathsf{f}(c_\psi)n^{1/2}(\hat{z}_\psi/\hat{\sigma}_\psi - c_\psi) \xrightarrow{\mathsf{D}} \mathbb{Z}(\psi) \qquad \text{on } D[\psi_0, 1]. \tag{52}$$

We now choose the stopping time as the first time greater than or equal to $m_1(\ge m_0)$, where $\hat{z}^{(m)}/\hat{\sigma}^{(m)}$ exceeds q times its pointwise asymptotic standard deviation, that is,

$$\hat{m} = \underset{m_1\le m<n}{\arg\min}[\mathbb{Z}_n(c_{m/n}) > q\mathsf{sdv}\{\mathbb{Z}_n(c_{m/n})\}]. \tag{53}$$

We insert this into expression into (51) noting that

$$(\hat{m} \leq j) = [\max_{m_1 \leq m \leq j} \frac{\mathbb{Z}_n(c_{m/n})}{\mathsf{sdv}\{\mathbb{Z}_n(c_{m/n})\}} > q].$$

The convergence (52) then lead to the following result; see Appendix for details.

Theorem 9 *Consider the Forward Search stopped at $\hat{m}$ in (53) for some $q \geq 0$ and $m_0 = \text{int}(\psi_0 n)$ and $m_1 = \text{int}(\psi_1 n)$ for $\psi_1 \geq \psi_0 > 0$. If Assumption 1 is satisfied then*

$$\mathsf{E}\hat{\gamma} = \mathsf{E}\frac{n-\hat{m}}{n} \to \gamma = \int_{\psi_1}^{1} \mathsf{P}[\sup_{\psi_1 \leq \psi \leq u} \frac{\mathbb{Z}(c_\psi)}{\mathsf{sdv}\{\mathbb{Z}(c_\psi)\}} > q]du.$$

If $\psi_1 > \psi_0$, the same limit holds for the Forward Search when replacing $\hat{z}^{(m)}$ by the deletion residual $\hat{d}^{(m)}$, see (19), in the definition of $\hat{m}$ in (53).

<Table 4 here>

The integral in Theorem 9 cannot be computed analytically in an obvious way. Instead we simulated it using Ox 7, see Doornik (2007). For a given n, we draw normal ε_i. From this, we compute the process $\mathbb{Z}_n$ and then the maximum of $\mathbb{Z}_n(c_{m/n})/\mathsf{sdv}\{\mathbb{Z}(c_{m/n})\}$ over $m_1 \leq m \leq j$ for all j so that $m_1 \leq j \leq n$. Repeating this n_{rep} times we can approximate the probability appearing as the integrand for given values of q and u . From this the integral γ is computed. This expresses γ as a function of q and ψ_1. Inverting this for fixed ψ_1 expresses q as a function of γ and ψ_1. Table 4 reports results for $n_{rep} = 10^5$ and $n = 1600$.

10 Application of the Forward Search to the fish data

We next apply the methods analyzed above to the fish data. We need to choose the initial estimator, the fractions ψ_0, ψ_1 and the gauge. As initial estimator we chose the fast LTS estimator with breakdown point ψ_0 by Rousseeuw & van Driessen (1998) as implemented in the `ltsReg` function of the R-package `robustbase`, see Rousseuw *et al.* (2013). There is no asymptotic analysis of this estimator. It is meant to be an approximation to the Least Trimmed Squares estimator, for which we have Theorem 4 based on Víček (2006c). That result requires fixed regressors. Nonetheless, we apply it to the fish data where the two regressors are the lagged dependent variable and the binary variable S_t which is an indicator for stormy weather. We choose $\psi_0 = \psi_1$ as either 0.95 or 0.8.

Figure 4 shows the forward plots of a renormalized version of the scaled forward residuals, $\hat{\xi}^{(m)}_{(m+1)}/\varsigma_{m/n}\hat{\sigma}^{(m+1)}$, where the scaling is chosen in line with Atkinson *et al.* (2010).

Panel (*a*) has $\psi_0 = \psi_1 = 0.95$. Choose the gauge as, for instance, $\gamma = 0.01$, in which case we need to consider the third exit band from the top. This is exceeded for $\hat{m} = 107$, pointing at $n - \hat{m} = 3$ outliers. These are the three holiday observations 18, 34, 95 discussed in Section 2. If the gauge is set to $\gamma = 0.001$ we find no outliers. If the gauge is set to $\gamma = 0.05$ we find $\hat{m} = 104$, pointing at $n - \hat{m} = 6$, which is 5% of the observations.

Panel (*b*) has $\psi_0 = \psi_1 = 0.80$. With a gauge of $\gamma = 0.01$, we find $\hat{m} = 96$, pointing at $n - \hat{m} = 14$ outliers. These include the three holiday observations along with 11 other

observations. This leaves some uncertainty about the best choice of the number of outliers. The present analysis is based on asymptotic analysis of the expected gauge. It does not consider sampling variation for the sample gauge and it could be distorted in finite samples.

<Fig4 here>

11 Conclusion and further work

We have presented the relation between some outlier detection algorithms and robust statistics, which discard some observations. We have exploited the Huber-skip and the LTS to construct 1-step versions of these as outlier detection algorithms, with robustified least squares and Impulse Indicator Saturation as special cases, and we have analyzed the Forward Search. We have given an overview of the asymptotic theory of these 1-step estimators and the stochastic expansion that allows to derive asymptotic distributions. The outlier detection algorithms are discussed in terms of their gauge, and we have shown that for the 1-step Huber-skip, the population gauge is the size of the underlying test, whereas for the Forward Search, a different relation was derived which can be used for calibrating the algorithm.

In future research we will look at situations, where there actually are outliers. Various configurations of outliers will be of interest: single outliers, clusters of outliers, level shifts, symmetric and non-symmetric outliers, or ϵ-contamination.

12 Acknowledgment

We would like to thank the organizers of the NordStat meeting in Turku, Finland, June 2014, for giving us the opportunity to present these lectures on outlier detection, and Elvezio Ronchetti, Christophe Croux, Jurgen Doornik and Silvelyn Zwanzig for comments to the manuscript. The first author is grateful to CREATES - Center for Research in Econometric Analysis of Time Series (DNRF78), funded by the Danish National Research Foundation.

References

Atkinson, A. C. & Riani, M. (2000). *Robust diagnostic regression analysis.* Springer, New York.

Atkinson, A. C., Riani, M. & Cerioli, A. (2010). The forward search: Theory and data analysis (with discussion). *J. Korean Stat. Soc.* **39**, 117–134.

Bahadur, R. R. (1966). A note on quantiles in large samples. *Ann. Math. Statist.* **37**, 577–580.

Benjamini, Y. & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc.* B **57**, 289–300.

Bercu, B. & Touati, A. (2008). Exponential inequalities for self-normalized martingales with applications. *Ann. Appl. Probab.* **18**, 1848–1869.

Bickel, P. J. (1975). One-step Huber estimates in the linear model. *J. Amer. Statist. Assoc.* **70**, 428–434.

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

Castle, J. L., Doornik, J. A. & Hendry, D. F. (2011). Evaluating automatic model selection. *J. Time Ser. Econom.* **3**, Issue 1, Article 8.

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

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

Csörgő, M. (1983). *Quantile Processes with Statistical Applications.* CBMS-NFS Regional Conference Series in Applied Mathematics **42**, Society for Industrial and Applied Mathematics.

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

Doornik, J. A. (2007). *Object-oriented matrix programming using Ox*, 3rd ed. Timberlake Consultants Press, London, and Oxford: www.doornik.com.

Doornik, J. A. (2009). Autometrics. In *The methodology and practice of econometrics: A festschrift in honour of David F. Hendry (*eds J. L. Castle, & N. Shephard), 88–121. Oxford University Press, Oxford.

Doornik, J. A. & Hendry, D. F. (2013). *Empirical econometric modelling - PcGive 14, volume 1.* Timberlake Consultants, London.

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica* **50**, 987–1108.

Engler, E. & Nielsen, B. (2009). The empirical process of autoregressive residuals. *Econom. J.* **12**, 367–381.

Godfrey, L. G. (1978). Testing against general autoregressive and moving average error models when the regressors include lagged dependent variables. *Econometrica* **46**, 1293–1302.

Graddy, K. (1995). Testing for imperfect competition at the Fulton fish market. *RAND J. Econom.* **26**, 75–92.

Graddy, K. (2006). The Fulton fish market. *J. Econom. Perspec.* **20**, 207–220.

Hadi, A. S. (1992). Identifying multiple outliers in multivariate data. *J. R. Stat. Soc. Ser. B* **54**, 761–771.

Hadi, A. S. & Simonoff, J. S. (1993). Procedures for the identification of multiple outliers in linear models *J. Amer. Statist. Assoc.* **88**, 1264-1272.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. & Stahel, W. A. (1986). *Robust statistics*: *The approach based on influence functions.* John Wiley & Sons, New York.

He, X. & Portnoy, S. (1992). Reweighted LS estimators converge at the same rate as the initial estimator. *Ann. Statist.* **20**, 2161–2167.

Hendry, D. F. & Doornik, J. A. (2014). *Empirical model discovery and theory evaluation.* MIT Press, Cambridge MA.

Hendry, D. F. & Krolzig, H. -M. (2005). The properties of automatic *GETS* modelling. *Econom. J.* **115**, C32–61.

Hendry, D. F. & Nielsen, B. (2007). *Econometric modelling.* Princeton University Press, Princeton NJ.

Hendry, D. F. & Santos, C. (2010). An automatic test of super exogeneity. In Bollerslev, T., Russell, J.R. & Watson, M.W. (eds.) *Volatility and time series econometrics: Essays in honor of Robert F. Engle*, pp. 164–193. Oxford University Press, Oxford.

Hoover, K. D. & Perez, S. J. (1999). Data mining reconsidered: encompassing and the general-to-specific approach to specification search (with discussion). *Econom. J.* **2**, 167–191.

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

Huber, P. J. & Ronchetti, E.M. (2009). *Robust statistics.* Wiley, New York.

Johansen, S. & Nielsen, B. (2009). An analysis of the indicator saturation estimator as a robust regression estimator. In *The methodology and practice of econometrics: A festschrift in honour of David F. Hendry* (eds J. L. Castle, & N. Shephard), 1-36. Oxford University Press, Oxford.

Johansen, S. & Nielsen, B. (2010). Discussion: The Forward Search: Theory and data analysis. *J. Korean Statist. Soc.* **39**, 137–145.

Johansen, S. & Nielsen, B. (2013). Asymptotic theory for iterated one-step Huber-skip estimators. *Econometrics* **1**, 53–70.

Johansen, S. & Nielsen, B. (2015a). Analysis of the Forward Search using some new results for martingales and empirical processes. *Bernoulli* DOI: 10.3150/14-BEJ689.

Johansen, S. & Nielsen, B. (2015b). Asymptotic theory of M-estimators in linear time series regression models. University of Copenhagen.

Jurečková, J. & Sen, P. K. (1996). *Robust statistical procedures: Asymptotics and Interrelations.* John Wiley & Sons, New York.

Kilian, L. & Demiroglu, U. (2000). Residual based tests for normality in autoregressions: asymptotic theory and simulations. *J. Bus. Econom. Statist.* **18**, 40–50.

Koul, H. L. (2002). *Weighted empirical processes in dynamic nonlinear models,* 2nd edn. Springer, New York.

Koul, H. L. & Ossiander, M. (1994). Weak convergence of randomly weighted dependent residual empiricals with applications to autoregression. *Ann. Statist.* **22**, 540–582.

Liese, F. & Vajda, I. (1994). Consistency of M-estimates in general regression models. *J. Multivariate Anal.* **50,** 93–114.

Maronna, R. A., Martin, R. D. & Yohai, V. J. (2006). *Robust statistics: theory and methods.* John Wiley & Sons, Chichester.

Nielsen, B. (2006). Order determination in general vector autoregressions. In *Time series and related topics: In memory of Ching-Zong Wei* (eds H. -C. Ho, C. -K. Ing & T. L. Lai), 93-112. IMS Lecture Notes and Monograph Series 52.

Nielsen, B. (2014). Outlier detection algorithms for least squares time series regression. Discussion paper, Nuffield College.

R Development Core Team (2014). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria.

Ramsey, J. B. (1969). Tests for specification errors in classical linear least squares regression analysis. *J. R. Stat. Soc. Ser. B* **31**, 350–371.

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

Rousseeuw, P. J. & van Driessen, K. (1998). A fast algorithm for the minimum covariance determinant estimator. *Technometrics* **41**, 212–223.

Rousseeuw, P. J. & Leroy, A. M. (1987). *Robust regression and outlier detection.* Wiley, New York.

Rousseeuw, P., Croux, C., Todorov, V., Ruckstuhl, A., Salibian-Barrera, M., Verbeke, T., Koller, M. & Maechler M. (2013). robustbase: Basic Robust Statistics. R package version 0.9-10. URL http://CRAN.R-project.org/package=robustbase

Ruppert, D. & Carroll, R. J. (1980). Trimmed least squares estimation in the linear model. *J. Amer. Statist. Assoc.* **75**, 828–838.

Sampford, M. R. (1953). Some inequalities on Mill's ratio and related functions. *Ann. Math. Statist.* **24**, 130–132.

Víšek, J. Á. (2006a). The least trimmed squares. Part I: Consistency. *Kybernetika* **42**, 1–36.

Víšek, J. Á. (2006b). The least trimmed squares. Part II: $\sqrt{n}$-consistency. *Kybernetika* **42**, 181–202.

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

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

White, H. (1980). A heteroskedastic-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica* **48**, 817–838.

Corresponding author:
Bent Nielsen, Nuffield College & Department of Economics, University of Oxford & Programme on Economic Modelling, INET, Oxford. Address for correspondence: Nuffield College, Oxford OX1 1NF, UK. E-mail: bent.nielsen@nuffield.ox.ac.uk.

A Proofs

We start by stating variances of various moments of the innovations.

Lemma 1 *Suppose ε_i has symmetric density f with finite fourth moment. Recall the notation $\psi, \tau, \varkappa$ defined in Section 3 and let $\kappa = \mathsf{E}(\varepsilon_1/\sigma)^4$. Then*

$$\mathsf{Var}\begin{pmatrix} \varepsilon_i \\ \varepsilon_i 1_{(|\varepsilon_i|\leq \sigma c_\psi)} \end{pmatrix} = \sigma^2 \begin{pmatrix} 1 & \tau \\ \tau & \tau \end{pmatrix},$$

$$\mathsf{Var}\begin{pmatrix} \varepsilon_i^2/\sigma^2 - 1 \\ 1_{(|\varepsilon_i|\leq \sigma c_\psi)} - \psi \\ (\varepsilon_i^2/\sigma^2 - \tau_\psi/\psi)1_{(|\varepsilon_i|\leq \sigma c_\psi)} \end{pmatrix} = \begin{pmatrix} \kappa - 1 & \tau - \psi & \varkappa - \tau^2/\psi \\ \tau - \psi & \psi(1-\psi) & 0 \\ \varkappa - \tau^2/\psi & 0 & \varkappa - \tau^2/\psi \end{pmatrix}.$$

The stochastic expansions are given in terms of two empirical processes. The next result shows asymptotic normality as processes on $D[0,1]$ and find their variances.

Lemma 2 *Suppose Assumption 1 holds and that $c_\psi = \mathsf{G}(\psi)$. Then the processes*

$$\mathbb{A}_n(\psi) = n^{-1/2}\sum_{i=1}^{n}\{1_{(|\varepsilon_i|\leq \sigma c_\psi)} - \psi\}, \qquad \mathbb{B}_n(\psi) = n^{-1/2}\sum_{i=1}^{n}(\frac{\varepsilon_i^2}{\sigma^2} - \frac{\tau_\psi}{\psi})1_{(|\varepsilon_i|\leq \sigma c_\psi)}. \tag{54}$$

converge to continuous Gaussian limits $\mathbb{A}, \mathbb{B}$ *on* $D[0,1]$ *endowed with the uniform metric. The processes have covariance matrix*

$$Var \left\{ \begin{array}{c} \mathbb{A}_n(\psi) \\ \mathbb{B}_n(\psi) \end{array} \right\} = \left\{ \begin{array}{cc} \psi(1-\psi) & 0 \\ 0 & \varkappa - \tau^2/\psi \end{array} \right\}.$$

The asymptotic variances in Theorem 5 are then

$$as\mathsf{Var} \left\{ \begin{array}{c} 2\mathsf{f}(c_\psi) n^{1/2}(\frac{\hat{z}_\psi}{\sigma} - c_\psi) \\ n^{1/2}(\frac{\hat{\sigma}^2_\psi}{\sigma^2} - 1) \end{array} \right\} = \left(\begin{array}{cc} \psi(1-\psi) & \psi(1-\psi)(\frac{c^2_\psi}{\tau_\psi} - \frac{1}{\psi}) \\ (\frac{c^2_\psi}{\tau_\psi} - \frac{1}{\psi})\psi(1-\psi) & (\frac{c^2_\psi}{\tau_\psi} - \frac{1}{\psi})^2\psi(1-\psi) + \frac{1}{\tau^2_\psi}(\varkappa_\psi - \frac{\tau^2_\psi}{\psi}) \end{array} \right).$$

The asymptotic variance in Theorem 3 is

$$\mathsf{asVar}\{2\mathsf{f}(c_\psi) n^{1/2}(\frac{\hat{z}_\psi}{\hat{\sigma}_\psi} - c_\psi)\} = \{1 - \frac{c_\psi \mathsf{f}(c_\psi)}{\tau_\psi}(c^2_\psi - \frac{\tau_\psi}{\psi})\}^2 \psi(1-\psi) + \{\frac{c_\psi \mathsf{f}(c_\psi)}{\tau_\psi}\}^2 (\varkappa_\psi - \frac{\tau^2_\psi}{\psi}).$$

With stationary regressors the process $\mathbb{K}_n(\psi) = N' \sum_{i=1}^n x_i \varepsilon_i 1_{(|\varepsilon_i \sigma| \leq \sigma c_\psi)}$ *converges to a continuous Gaussian process* $\mathbb{K}$ *on* $D[0,1]$*, which is independent of* $\mathbb{A}, \mathbb{B}$ *and has variance* $\tau_\psi \sigma^2 \Sigma$*.*

Proof of Lemma 2. The expansions (42), (43) in Theorem 5 can be expressed as

$$\begin{aligned} 2f(c_\psi) n^{1/2}(\frac{\hat{z}_\psi}{\sigma} - c_\psi) &= -\mathbb{A}_n(\psi) + \mathsf{o_P}(\psi), \\ n^{1/2}(\frac{\hat{\sigma}^2_\psi}{\sigma^2} - 1) &= \tau^{-1}_\psi \{\mathbb{B}_n(\psi) - (c^2_\psi - \frac{\tau_\psi}{\psi})\mathbb{A}_n(\psi)\} + \mathsf{o_P}(\psi), \end{aligned}$$

while the expansion (44) in Theorem 3, follows from

$$\begin{aligned} 2\mathsf{f}(c_\psi) n^{1/2}(\frac{\hat{z}_\psi}{\hat{\sigma}_\psi} - c_\psi) &= 2\mathsf{f}(c_\psi) n^{1/2}(\frac{\hat{z}_\psi}{\sigma} - c_\psi) - c_\psi \mathsf{f}(c_\psi) n^{1/2}(\frac{\hat{\sigma}^2_\psi}{\sigma^2} - 1) + \mathsf{o_P}(\psi) \\ &= -\{1 - \frac{c_\psi \mathsf{f}(c_\psi)}{\tau_\psi}(c^2_\psi - \frac{\tau_\psi}{\psi})\}\mathbb{A}_n(\psi) - \frac{c_\psi \mathsf{f}(c_\psi)}{\tau_\psi}\mathbb{B}_n(\psi) + \mathsf{o_P}(\psi), \end{aligned}$$

which gives the variances. The convergence of the finite dimensional distributions of the processes $\mathbb{A}_n$ and $\mathbb{B}_n$ follow from the Central Limit Theorem for i.i.d. variables, and the corresponding result for $\mathbb{K}_n$ from the Central Limit Theorem for martingale difference sequences. Tightness of the processes follow from Theorem 4.4 in Johansen and Nielsen (2015a). ■

Proof of Theorem 5. The desired results follow from Lemmas D.10 and D.11 in Johansen and Nielsen (2015a), albeit with different notation. To recognize the results let $\tilde{b} = N^{-1}(\tilde{\beta} - \beta)$ to match the notation in the beginning of section D.2. The present order statistic $\tilde{\xi}_{(k)}$ is written as $\sigma \hat{c}^{\tilde{b}}_\psi$ to match (D.12), noting that $\psi = k/n$. We can then proceed to write the objects of interest in terms of the weighted and marked absolute empirical processes $\hat{\mathsf{G}}^{g,p}_n$ defined in (D.3).

1. *Expansion* (41). For the regression estimator we have

$$N^{-1}(\hat{\beta} - \beta) = \{\hat{\mathsf{G}}^{xx,0}_n(\tilde{b}, \hat{c}^{\tilde{b}}_\psi)\}^{-1} n^{1/2} \hat{\mathsf{G}}^{x,1}_n(\tilde{b}, \hat{c}^{\tilde{b}}_\psi).$$

Lemma D.10(c) gives the desired expansion as long as $\tilde{b}$ is tight, or even slowly diverging.

2. *Expansion* (43). This is the Bahadur expansion in Lemma D.2(b).
3. *Expansion* (42). Since $\sum_{i=1}^n v_i = k = \psi n$ for the 1-step LTS, we have that

$$n^{1/2}(\hat{\sigma}^2 - \sigma^2) = n^{1/2}(\frac{1}{\tau}[\hat{\mathsf{G}}_n^{1,2}(\tilde{b}, \hat{c}_\psi^{\tilde{b}}) - \{\hat{\mathsf{G}}_n^{x,1}(\tilde{b}, \hat{c}_\psi^{\tilde{b}})\}'\{\hat{\mathsf{G}}_n^{xx,0}(\tilde{b}, \hat{c}_\psi^{\tilde{b}})\}^{-1}\hat{\mathsf{G}}_n^{x,1}(\tilde{b}, \hat{c}_\psi^{\tilde{b}})] - \sigma^2).$$

Lemma D.11(b) shows that $n^{1/2}(\hat{\mathsf{G}}_n^{x,1})'(\hat{\mathsf{G}}_n^{xx,0})^{-1}\hat{\mathsf{G}}_n^{x,1} = \mathrm{o}_{\mathsf{P}}(1)$. Lemma D.11(a) expands the remaining components in terms of the processes

$$\mathbb{G}_n^{1,0}(c_\psi) = \mathbb{A}_n(\psi), \qquad \mathbb{G}_n^{1,2}(c_\psi) = \sigma^2\mathbb{B}_n(\psi) + \sigma^2\frac{\tau}{\psi}\mathbb{A}_n(\psi),$$

where $\mathbb{A}_n, \mathbb{B}_n$ were defined in (54). Thus, we get

$$n^{1/2}(\hat{\sigma}^2 - \sigma^2) = \frac{\sigma^2}{\tau}\{\mathbb{B}_n(\psi) - (c_\psi^2 - \frac{\tau}{\psi})\mathbb{A}_n(\psi)\} + \mathrm{o}_{\mathsf{P}}(1).$$

Then apply that $-\mathbb{A}_n(\psi) = 2c_\psi \mathsf{f}(c_\psi)\sigma^{-1}(\hat{z}_\psi/c_\psi - \sigma)$ according to (43) along with the definition of ρ_σ in (34). ■

Proof of Theorem 7. Write the gauge as $\hat{\gamma} = n^{-1}\sum_{i=1}^n(1 - v_i)$ and apply the expansion (21) for $n^{-1}\sum_{i=1}^n v_i$, see also Lemma D5 of Johansen and Nielsen (2015a). The expansion implies $\hat{\gamma} \to 1 - \psi$ in probability. Note that convergence in probability is equivalent to convergence in mean since the empirical gauge is bounded, see Theorem 5.4 in Billingsley (1968). ■

Proof of Corollary 4. Insert $\hat{\sigma}^2 = \sigma^2$ in the expansion in Theorem 7 and apply the Central Limit Theorem for i.i.d. variables to the binomial term, see Lemma 2. ■

Proof of Corollary 5. Iterate the expansion (36) for $(\hat{\sigma}^{(1)})^2$ in Theorem 3 to get

$$n^{1/2}\{(\sigma^{(m-1)})^2 - \sigma^2\} = \rho_\sigma^{m-1}n^{1/2}(\tilde{\sigma}^2 - \sigma^2) + \frac{1 - \rho_\sigma^{m-1}}{(1 - \rho_\sigma)\tau}\sigma^2 n^{-1/2}\sum_{i=1}^n(\frac{\varepsilon_i^2}{\sigma^2} - \frac{\tau}{\psi})1_{(|\varepsilon_i| \le \sigma c)} + \mathrm{o}_{\mathsf{P}}(1).$$

Recall that $n^{1/2}(\tilde{\sigma}^2 - \sigma^2) = n^{-1/2}\sum_{i=1}^n(\varepsilon_i^2 - \sigma^2)$ and $n^{1/2}(\sigma^{(m-1)} - \sigma) = n^{1/2}\{(\sigma^{(m-1)})^2 - \sigma^2\}/(2\sigma)$. Insert this into expression (46) for $\gamma^{(m)}$ in Theorem 7 to get

$$\mathsf{asVar}(\hat{\gamma}^{(m)}) = \mathsf{Var}[-1_{(|\varepsilon_i| \le \sigma c)} - c\mathsf{f}(c)\{\rho_\sigma^{m-1}(\varepsilon_i^2/\sigma^2 - 1) + \frac{1 - \rho_\sigma^{m-1}}{(1 - \rho_\sigma)\tau}(\frac{\varepsilon_i^2}{\sigma^2} - \frac{\tau}{\psi})1_{(|\varepsilon_i| \le \sigma c)}\}].$$

Use Lemma 1 and the definition of $\eta_\sigma^{(m-1)}$ in Corollary 1. ■

Remark 1 *Assumption 1(i'a) implies that $c_n = \mathrm{O}(n^{1/r})$ where $1/r < 1/4$. To see this, combine the definition $\mathsf{P}(|\varepsilon_i| > \sigma c_n) = \lambda/n$ with the Markov inequality $\mathsf{P}(|\varepsilon_i| > \sigma c_n) \le (\sigma c_n)^{-r}\mathsf{E}|\varepsilon_i|^r$ so that $c_n \le \sigma^{-1}(\mathsf{E}|\varepsilon_i|^r)^{1/r}\lambda^{-1/r}n^{1/r} = \mathrm{O}(n^{1/r})$.*

Remark 2 *Assumption 1(i') holds if $\mathsf{f} = \varphi$ is standard normal. For (b) use the Mill's ratio result $\{(4 + c^2)^{1/2} - c\}/2 < \{1 - \Phi(c)\}/\varphi(c)$; see Sampford (1953). For (c) note that $2\log\{\mathsf{f}(c_n - n^{-1/4}A)/\mathsf{f}(c_n)\} = c_n^2 - (c_n - n^{-1/4}A)^2 = 2c_n n^{-1/4}A - n^{-1/2}A^2$, use Remark 1.*

Proof of Theorem 8. 1. *A bound on the sample space.* Since $N^{-1}(\tilde\beta_n-\beta)$ and $n^{1/2}(\tilde\sigma_n^2-\sigma^2)$ are $\mathrm{O_P}(1)$, then $n^{1/2}(\tilde\sigma_n-\sigma)$ is $\mathrm{O_P}(1)$ and in light of Assumption $1(ii,d)$, there exists for all $\epsilon>0$ a constant $A_0>1$ such that the set

$$\mathcal{B}_n=\{|N^{-1}(\tilde\beta_n-\beta)|+n^{1/2}|\tilde\sigma_n-\sigma|+n^{1/4}\max_{1\le i\le n}|N'x_i|\le A_0\} \tag{55}$$

has probability larger than $1-\epsilon$ for all n.

2. *A bound on the indicators.* Introduce the quantity

$$s_i=\tilde\sigma c_n-y_i+x_i'\tilde\beta_n+\varepsilon_i=\sigma c_n+n^{-1/2}n^{1/2}(\tilde\sigma_n-\sigma)c_n+x_i'NN^{-1}(\tilde\beta_n-\beta).$$

On the set $\mathcal{B}_n$, using $c_n=\mathrm{o}(n^{1/4})$, by Remark 1 the quantity s_i satisfies, for some $A_1>0$,

$$\begin{aligned}s_i&\le\sigma c_n+n^{-1/2}A_0c_n+n^{-1/4}A_0^2\le\sigma(c_n+n^{-1/4}A_1),\\ s_i&\ge\sigma c_n-n^{-1/2}A_0c_n-n^{-1/4}A_0^2\ge\sigma(c_n-n^{-1/4}A_1).\end{aligned}$$

It therefore holds that

$$1_{(\varepsilon_i/\sigma>c_n+n^{-1/4}A_1)}\le 1_{(y_i-x_i'\tilde\beta_n>\tilde\sigma_nc_n)}=1_{(\varepsilon_i>s_i)}\le 1_{(\varepsilon_i/\sigma>c_n-n^{-1/4}A_1)}.$$

With a similar inequality for $1_{(y_i-x_i'\hat\beta<-\hat\sigma c_n)}$ we find

$$1_{(|\varepsilon_i/\sigma|>c_n+n^{-1/4}A_1)}\le 1-v_i=1_{(|y_i-x_i'\tilde\beta_n|>\tilde\sigma_nc_n)}\le 1_{(|\varepsilon_i/\sigma|>c_n-n^{-1/4}A_1)}. \tag{56}$$

3. *Expectation of indicator bounds.* It will be argued that

$$n\mathsf{E}1_{(|\varepsilon_i/\sigma|>c_n+n^{-1/4}A_1)}\to\lambda,\qquad n\mathsf{E}1_{(|\varepsilon_i/\sigma|>c_n-n^{-1/4}A_1)}\to\lambda. \tag{57}$$

Since $n\mathsf{E}1_{(|\varepsilon_i/\sigma|>c_n)}=\lambda$ it suffices to argue that

$$\mathcal{E}_n=n\mathsf{E}\{1_{(|\varepsilon_i/\sigma|>c_n-n^{-1/4}A_1)}-1_{(|\varepsilon_i/\sigma|>c_n+n^{-1/4}A_1)}\}\to 0.$$

The mean value theorem and the identity $2\{1-\mathsf{F}(c_n)\}=\lambda/n$ give for $|c^*-c_n|\le n^{-1/4}A_1$,

$$\begin{aligned}\mathcal{E}_n&=n\int_{c_n-n^{-1/4}A_1}^{c_n+n^{-1/4}A_1}2\mathsf{f}(x)dx=n4n^{-1/4}A_1\mathsf{f}(c^*)=\frac{4\lambda n^{-1/4}A_1\mathsf{f}(c^*)}{2\{1-\mathsf{F}(c_n)\}},\\ &=2\lambda n^{-1/4}A_1\{\frac{\mathsf{f}(c^*)}{\mathsf{f}(c_n-n^{-1/4}A_1)}\}\{\frac{\mathsf{f}(c_n-n^{-1/4}A_1)}{\mathsf{f}(c_n)}\}[\frac{\mathsf{f}(c_n)}{c_n\{1-\mathsf{F}(c_n)\}}]c_n.\end{aligned}$$

The first ratio is bounded by 1 since f has decreasing tails. The second and third ratios are bounded by Assumption 1 $(i'b,i'c)$. Then use that $n^{-1/4}c_n=\mathrm{o}(1)$ by Remark 1.

4. *Comparison of* $\hat\beta_n^{(1)},\hat\sigma_n^{(1)}$ *and* $\hat\beta_{LS},\hat\sigma_{LS}$. The estimation errors of the 1-step Huber-skip, $N^{-1}(\hat\beta_n-\beta)$, $n^{1/2}(\hat\sigma_n^2-\sigma^2)$, are based on product moments of the form, see (20),

$$\sum_{i=1}^n v_ig_i=\sum_{i=1}^n g_i-\sum_{i=1}^n(1-v_i)g_i,$$

where $v_i=1_{(|y_i-x_i'\tilde\beta|\le c_n\tilde\sigma)}$ and g_i is $N'x_ix_i'N$, $N'x_i\varepsilon_i$, $n^{-1/2}(\varepsilon_i^2-\sigma^2)$ or n^{-1}. The first terms give the least squares estimators, and (49) follows for $m=1$ if the second terms are $\mathrm{o_P}(1)$.

5. *Bound on the second terms.* On the set $\mathcal{B}_n$ given in (55), we use the bound on $1-v_i$ in (56). When g_i is $n^{-1/2}(\varepsilon_i^2-\sigma^2)$ or n^{-1} we can consider, for $s=0,2$,

$$\begin{aligned}\mathsf{E}|n^{-1/2}\sum_{i=1}^n(1-v_i)\varepsilon_i^s 1_{\mathcal{B}_n}| &\le \mathsf{E}n^{-1/2}\sum_{i=1}^n\varepsilon_i^s 1_{(|\varepsilon_i/\sigma|>c_n-n^{-1/4}A_1)} \le n^{1/2}\mathsf{E}\varepsilon_i^s 1_{(|\varepsilon_i/\sigma|>c_n-n^{-1/4}A_1)} \\ &\le n^{1/2}\mathsf{E}^{1/2}1_{(|\varepsilon_1/\sigma|>c_n-n^{-1/4}A_1)}\mathsf{E}^{1/2}\varepsilon_1^{2s}1_{(|\varepsilon_1/\sigma|>c_n-n^{-1/4}A_1)} = \mathrm{o}(1),\end{aligned} \tag{58}$$

because $n\mathsf{E}1_{(|\varepsilon_1/\sigma|>c_n-n^{-1/4}A_1)}\to\lambda$, (57), $\mathsf{E}\varepsilon_1^{2s}<\infty, s\le 2$, Assumption $1(i'a)$, and $c_n - n^{-1/4}A_1\to\infty$. When g_i is $N'x_ix_i'N$ or $N'x_i\varepsilon_i$ we can consider, for $s=0,1$,

$$\mathcal{E} = \mathsf{E}\sum_{i=1}^n(1-v_i)|g_i|1_{\mathcal{B}_n} \le \mathsf{E}\sum_{i=1}^n|N'x_i|^{2-s}|\varepsilon_i|^s 1_{(|\varepsilon_i/\sigma|>c_n-n^{-1/4}A_1)}.$$

Taking iterated expectations with respect to $\mathcal{F}_{i-1}$ shows that

$$\mathcal{E} \le n^{s/2}\{\mathsf{E}n^{-1}\sum_{i=1}^n(|n^{1/2}N'x_i|^{2-s})\}\mathsf{E}\{|\varepsilon_i|^s 1_{(|\varepsilon_i/\sigma|>c_n-n^{-1/4}A_1)}\}.$$

The first expectation is $\mathrm{O}(1)$ by Assumption $1(iie)$. The second expectation is $\mathrm{o}(n^{-1/2})$ by (58). Overall we get $\mathcal{E}=\mathrm{o}(1)$.

6. *Proof (49)*: In item 5 we have proved (49) for $m=1$. A consequence is that the estimation errors for $\hat\beta_n^{(1)},\hat\sigma_n^{(1)}$, the initial estimators for $m=2$, are bounded in probability and hence the same conclusion holds for $\hat\beta_n^{(2)},\hat\sigma_n^{(2)}$, etc.

7. *Proof of (50)*: Using the bounds in item 2, (56), it holds on the set $\mathcal{B}_n$ that

$$\frac{1}{n}\sum_{i=1}^n 1_{(|\varepsilon_i/\sigma|>c_n+n^{-1/4}A_1)} \le \hat\gamma = \frac{1}{n}\sum_{i=1}^n 1_{(y_i-x_i'\tilde\beta_n>\tilde\sigma_n c_n)} \le \frac{1}{n}\sum_{i=1}^n 1_{(|\varepsilon_i/\sigma|>c_n-n^{-1/4}A_1)}.$$

Using (57), the Poisson limit theorem shows that the upper and lower bounds have Poisson limits with mean λ. ■

Proof of Theorem 9. Corollary 3 shows that $\mathbb{Z}_n$ converges to a Gaussian process $\mathbb{Z}$ on $D[\psi_0,1]$. The variance of $\mathbb{Z}(c_\psi)$ vanishes for $\psi\to 1$ so a truncation argument is needed to deal with the ratio $\mathbb{X}_n(c_\psi)=\mathbb{Z}_n(c_\psi)/\mathsf{sdv}\{\mathbb{Z}(c_\psi)\}$. Approximate the sample gauge by

$$\hat\gamma_v = \frac{n-\hat m}{n}1_{(\hat m\le vn)} = \frac{1}{n}\sum_{j=m_1}^{\mathrm{int}(nv)-1}1_{(\hat m\le j)},$$

for some $v<1$ and using (51). Then the sample gauge is $\hat\gamma=\hat\gamma_1$, and

$$0\le\hat\gamma-\hat\gamma_v = \frac{n-\hat m}{n}1_{(\hat m>vn)} < \frac{n-nv}{n} = 1-v.$$

The process $\mathbb{X}_n(c_\psi)$ converges on $D[\psi_1,v]$. The Continuous Mapping Theorem 5.1 of Billingsley (1968) then shows that $\sup_{\psi_1\le\psi\le u}\mathbb{X}_n(c_\psi)$ converges as a process in u on $D[\psi_1,v]$. In turn, for a given q, the deterministic function $\mathsf{P}(\hat m\le nu)=\mathsf{P}\{\sup_{\psi_1\le\psi\le u}\mathbb{X}_n(c_\psi)>q\}$ in $\psi_1\le u\le v$ converges to a continuous increasing function $\mathsf{p}(u)$ on $[\psi_1,v]$, which is bounded by unity. In particular it holds that

$$\mathsf{E}\hat\gamma_v = \frac{1}{n}\sum_{j=m_1}^{\mathrm{int}(nv)-1}\mathsf{E}1_{(\hat m\le j)} = \frac{1}{n}\sum_{j=m_1}^{\mathrm{int}(nv)-1}\mathsf{P}(\hat m\le j)\to\gamma_v = \int_{\psi_1}^v \mathsf{p}(u)du \le v-\psi_1\le 1-\psi_1,$$

and

$$\gamma_v = \int_{\psi_1}^{v} \mathsf{p}(u)du \nearrow \gamma = \int_{\psi_1}^{1} \mathsf{p}(u)du = \int_{\psi_1}^{1} \mathsf{P}[\sup_{\psi_1 \le \psi \le u} \frac{\mathbb{Z}(c_\psi)}{\mathsf{sdv}\{\mathbb{Z}(c_\psi)\}} > c]du$$

regardless of the behaviour of the process $\mathbb{X}_n(c)$ for ψ close to unity.

Now return to the sample gauge $\hat{\gamma}$, and rewrite it as

$$\hat{\gamma} - \gamma = (\gamma_v - \gamma) + (\hat{\gamma}_v - \gamma_v) + (\hat{\gamma} - \hat{\gamma}_v)$$

for some fixed v. Then

$$|\hat{\gamma} - \gamma| \le 1 - v + |\hat{\gamma}_v - \gamma_v| + 1 - v.$$

Choose an $\epsilon > 0$ and v such that $1 - v \le \epsilon$, and then n so large that $|\hat{\gamma}_v - \gamma_v| \le \epsilon$ with large probability, then $|\hat{\gamma} - \gamma| \le 3\epsilon$ with large probability, which completes the proof. ■

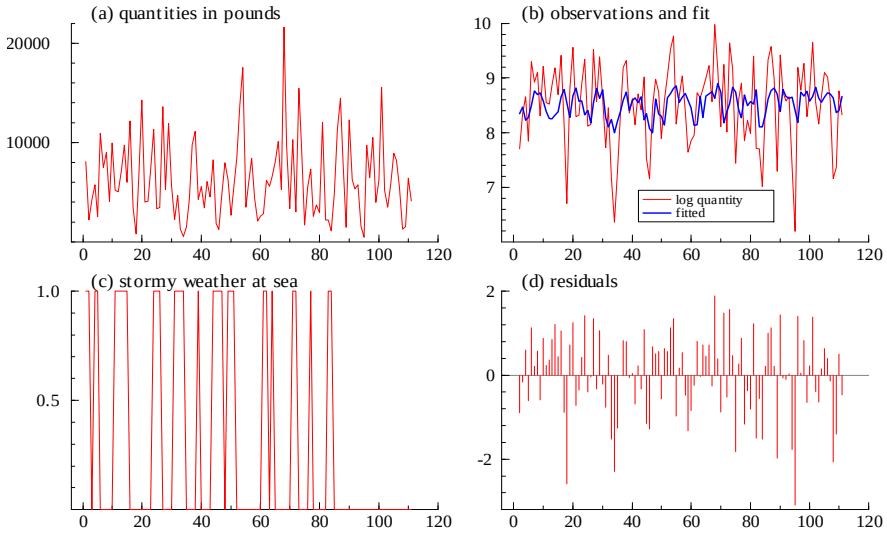

Figure 1: Data and properties of fitted model for Fulton fish market data

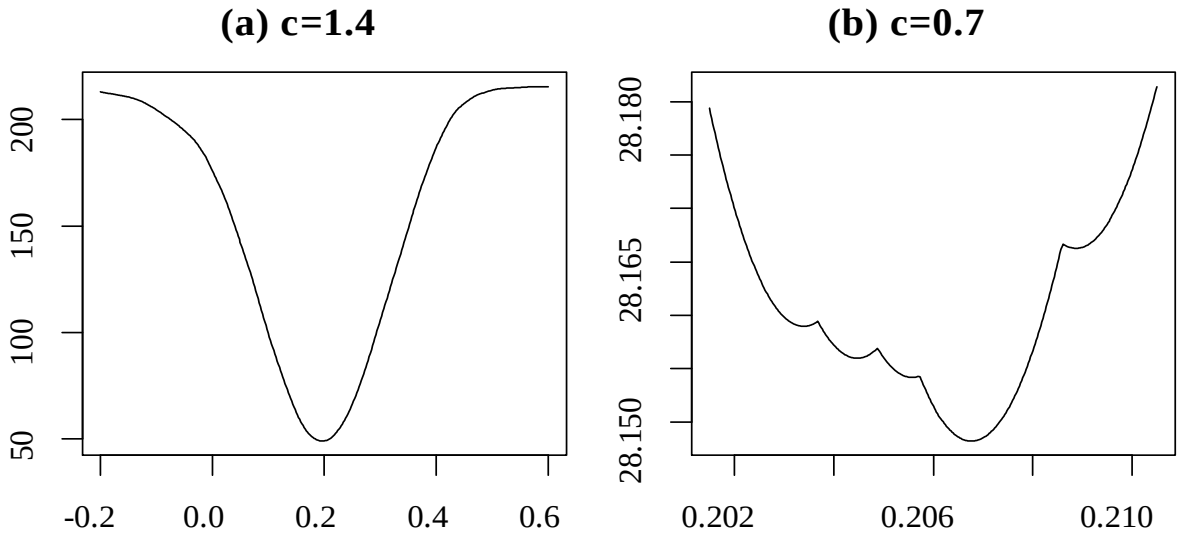

Figure 2: Huber-skip objective function R_n plotted against the parameter for the lagged dependent variable for the Fulton fish data for two values of c.

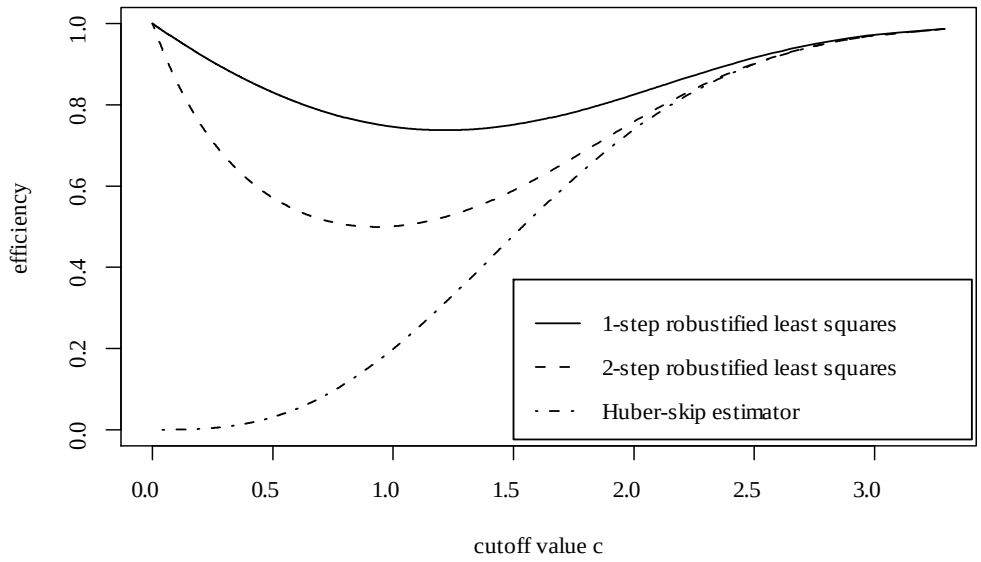

Figure 3: Efficiency as a function of c. The two top curves (—,- -) are 1 and 2-step robustified least squares, $1/\eta_\beta$, Corollary 6.3. The lowest curve $(\cdot\ -\ \cdot)$ is for Huber-skip, τ, Theorem 6.1. All measured relative to full sample least squares for a normal reference distribution.

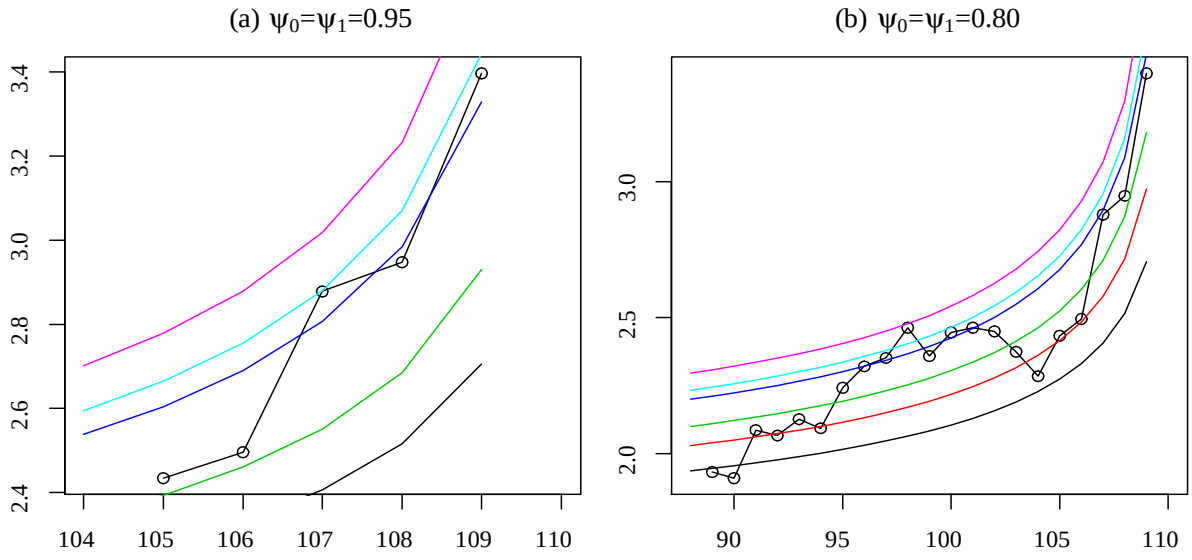

Figure 4: Forward Plots of forward residuals for fish data. Here $\psi_0 = \psi_1$ is chosen either as 0.95 or 0.80. The bottom curve shows the pointwise median. The top curves show the exit bands for gauges chosen as, from top, 0.001, 0.005, 0.01, 0.05, respectively. Panel (b) also includes an exit band for a gauge of 0.10.

Table 1: Number of errors committed when testing m hypotheses in the same model.

declared:	non-significant	significant	total
True Hypothesis	U	V	m_0
False Hypothesis	T	S	$m - m_0$
total	$m - R$	R	m

Table 2: Asymptotic standard deviations of the empirical gauge for the Huber-skip (Corollary 4) and robustified least squares (Corollary 5) and the fully iterated Huber-skip. All calculated for a normal reference distribution.

γ	0.05	0.01	0.005	0.0025	0.001
c	1.960	2.576	2.807	3.023	3.291
$\mathsf{sdv}(\hat{\gamma})$ for Huber-skip	0.218	0.0995	0.0705	0.0499	0.0316
$\mathsf{sdv}(\hat{\gamma})$ for Robustified Least Squares	0.146	0.0844	0.0634	0.0467	0.0305
$\mathsf{sdv}(\hat{\gamma})$ for fully iterated Huber-skip	0.314	0.117	0.0783	0.0534	0.0327

Table 3: Poisson approximations to the probability of finding at most x outliers for a given λ. The implied cutoff $c_n = \Phi^{-1}\{1 - \lambda/(2n)\}$ is shown for $n = 100$ and $n = 200$.

			x					
λ	$c_{n=100}$	$c_{n=200}$	0	1	2	3	4	5
5	1.960	2.241	0.01	0.04	0.12	0.27	0.44	0.62
1	2.576	2.807	0.37	0.74	0.92	0.98	1.00	
0.5	2.807	3.023	0.61	0.91	0.98	1.00		
0.25	3.023	3.227	0.78	0.97	1.00			
0.1	3.291	3.481	0.90	1.00				

Table 4: Cutoff values q for the Forward Search as a function of gauge γ and lower point ψ_1 of range for the stopping time, see Theorem 10.1.

γ vs ψ_1	0.05	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90
0.10	2.50	2.43	2.28	2.14	1.99	1.81	1.60	1.31	0.82	-
0.05	2.77	2.71	2.58	2.46	2.33	2.19	2.02	1.79	1.45	0.69
0.01	3.30	3.24	3.14	3.04	2.94	2.83	2.71	2.55	2.33	1.91
0.005	3.49	3.44	3.35	3.26	3.15	3.04	2.95	2.81	2.62	2.26
0.001	3.90	3.85	3.77	3.69	3.62	3.53	3.43	3.32	3.18	2.92