

Discussion Papers  
Department of Economics  
University of Copenhagen

No. 14-23

Outlier detection algorithms for least squares time series regression

Søren Johansen and Bent Nielsen

Øster Farimagsgade 5, Building 26, DK-1353 Copenhagen K., Denmark

Tel.: +45 35 32 30 01 – Fax: +45 35 32 30 00

<http://www.econ.ku.dk>

ISSN: 1601-2461 (E)

# Outlier detection algorithms for least squares time series regression<sup>1</sup>

Søren Johansen<sup>23</sup> & Bent Nielsen<sup>4</sup>

8 September 2014

**Summary:** We review recent asymptotic results on some robust methods for multiple regression. The regressors include stationary and non-stationary time series as well as polynomial terms. The methods include the Huber-skip M-estimator, 1-step Huber-skip M-estimators, in particular the Impulse Indicator Saturation, iterated 1-step Huber-skip M-estimators and the Forward Search. These methods classify observations as outliers or not. From the asymptotic results we establish a new asymptotic theory for the gauge of these methods, which is the expected frequency of falsely detected outliers. The asymptotic theory involves normal distribution results and Poisson distribution results. The theory is applied to a time series data set.

**Keywords:** Huber-skip M-estimators, 1-step Huber-skip M-estimators, iteration, Forward Search, Impulse Indicator Saturation, Robustified Least Squares, weighted and marked empirical processes, iterated martingale inequality, gauge.

**JEL Classification:** C22, C52

## 1 Introduction

The purpose of this paper is to review recent asymptotic results on some robust methods for multiple regression and apply these to calibrate these methods. The regressors include stationary and non-stationary time series as well as quite general deterministic terms. All the reviewed methods classify observations as outliers according to hard, binary decision rules. The methods include the Huber-skip M-estimator, 1-step versions such as the robustified least squares estimator and the Impulse Indicator Saturation, iterated 1-step versions thereof, and the Forward Search. The paper falls in two parts. In the first part we give a motivating empirical example. This is followed by an overview of the methods and a review of recent asymptotic tools and properties of the estimators. For all the presented methods the outlier classification depends on a cut-off value  $c$  which is taken as given in the first part. In the second part we provide an asymptotic theory for setting the cut-off value  $c$  indirectly from the gauge, where the gauge is defined as the frequency of observations classified as outliers, when in fact there are no outliers in the data generating process.

---

<sup>1</sup>Acknowledgements: 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.

<sup>2</sup>The first author is grateful to CREATES - Center for Research in Econometric Analysis of Time Series (DNRF78), funded by the Danish National Research Foundation.

<sup>3</sup>Department of Economics, University of Copenhagen and CREATES, Department of Economics and Business, Aarhus University, DK-8000 Aarhus C. E-mail: soren.johansen@econ.ku.dk.

<sup>4</sup>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.

Robust methods can be used in many ways. Some methods reject observations that are classified as outliers, while other methods give a smooth weight to all observations. It is open to discussion which method to use, see for instance Hampel, Ronchetti, Rousseeuw and Stahel (1986, §1.4). Here, we focus on rejection methods. We consider an empirical example, where rejection methods are useful as diagnostic tools. The idea is that most observations are ‘good’ in the sense that they conform with a regression model with symmetric, if not normal, errors. Some observations may not conform with the model - they are the outliers. When building a statistical model the user can apply the 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 analysis.

In order to use the algorithms with confidence we need to understand its properties when all observations are ‘good’. Just as in hypothesis testing, where tests are constructed by controlling their properties when the hypothesis is true, we consider the outlier detection methods when, in fact, there are no outliers. The proposal is to control the cut-off values of the robust methods in terms of their gauge. The gauge is the frequency of wrongly detected outliers when there are none. It is distinct from, but related to, the size of a hypothesis test and of false discovery rate in multiple testing (Benjamini and Hochberg, 1995).

The origins of the notion of a gauge are as follows. Hoover and 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 and Krolzig (2005), Doornik (2009) and Hendry and Doornik (2014). The Autometrics algorithm also includes an impulse indicator saturation algorithm. Through extensive simulation studies the critical values of these algorithms have been calibrated in terms of the false detection rates for irrelevant regressors and irrelevant outliers. The term gauge was introduced in Hendry and Santos (2010) and Castle, Doornik and Hendry (2011).

## 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 data of the quantity of whiting sold by one wholesaler over the period 2 Dec 1991 to 8 May 1992. Figure 1(a) shows the daily aggregated quantity  $Q_t$  measured in pounds. The logarithm of the

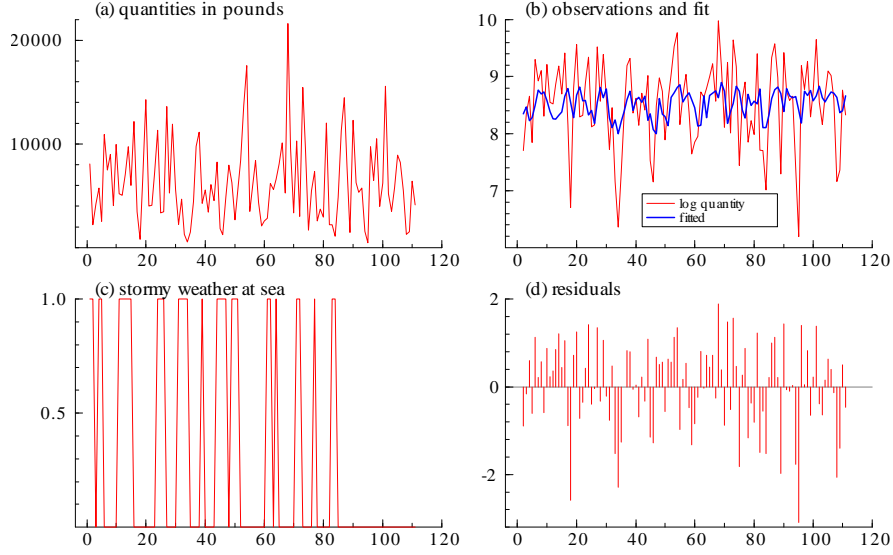


Figure 1: Data and properties of fitted model for Fulton Fish market data

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 Hendry and Nielsen (2007, §13.5).

A simple autoregressive model for log quantities  $q_t$  gives

$$\begin{aligned}
 \hat{q}_t &= 7.0 + 0.19q_{t-1} - 0.36 S_t, & (2.1) \\
 \text{(standard error)} & \quad (0.8) \quad (0.09) \quad (0.15) \\
 \text{[t-statistic]} & \quad [8.8] \quad [2.03] \quad [-2.39] \\
 \hat{\sigma} &= 0.72, \quad \hat{\ell} = -117.82, \quad R^2 = 0.090, \quad t = 2, \dots, 111, \\
 \chi_{norm}^2[2] &= 6.9 [p= 0.03], \quad \chi_{skew}^2[1] = 6.8 [p= 0.01], \quad \chi_{kurt}^2[1] = 0.04 [p= 0.84], \\
 F_{ar(1-2)}[2, 106] &= 0.9 [p= 0.40], \quad F_{arch(1)}[1, 106] = 1.4 [p= 0.24], \\
 F_{het}[3, 103] &= 2.0 [p= 0.12], \quad F_{reset}[1, 106] = 1.8 [p= 0.18].
 \end{aligned}$$

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,  $\chi_{skew}^2$ , kurtosis,  $\chi_{kurtosis}^2$  and both,  $\chi_{norm}^2 = \chi_{skew}^2 + \chi_{kurtosis}^2$ , a test  $F_{ar}$  for autoregressive temporal dependence, see Godfrey (1978), a test  $F_{arch}$  for autoregressive conditional heteroscedasticity, see Engle (1982), a test  $F_{het}$  for autoregressive conditional heteroscedasticity, see White (1980), and a test  $F_{reset}$  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 and Demiroglu (2000) and Engler and Nielsen (2009) for  $\chi_{skew}^2$ ,  $\chi_{kurtosis}^2$  and Nielsen (2006) for  $F_{ar}$ . The computations were done using OxMetrics, see Doornik and 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

$$\begin{aligned} \hat{q}_t &= \underset{(0.7)}{7.9} + \underset{(0.08)}{0.09}q_{t-1} - \underset{(0.14)}{0.36}S_t - \underset{(0.66)}{1.94}D_t^{18} - \underset{(0.66)}{1.82}D_t^{34} - \underset{(0.66)}{2.38}D_t^{95}, & (2.2) \\ &\underset{[10.8]}{\quad} \underset{[1.04]}{\quad} \underset{[-2.68]}{\quad} \underset{[-3.00]}{\quad} \underset{[-2.75]}{\quad} \underset{[-3.64]}{\quad} \\ \hat{\sigma} &= 0.64, \quad \hat{\ell} = -104.42, \quad R^2 = 0.287, \quad t = 2, \dots, 111. \end{aligned}$$

Specification tests, which are not reported, indicate a marked improvement in the specification. Comparing the regressions (2.1) and (2.2) 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 those observations that do not conform with the statistical model. In the following we will consider some algorithms for outlier detection that are inspired by this example. These algorithms are solely based on statistical information and we can then discuss their properties by mathematical means. In practice, outcomes should of course be assessed within the substantive context. We return to this example in §11.

### 3 Model

Throughout, we consider data  $(y_i, x_i)$ ,  $i = 1, \dots, n$  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  $(y_i, x_i)$ ,  $i = 1, \dots, n$  satisfy the multiple regression equation

$$y_i = \beta' x_i + \varepsilon_i, \quad i = 1, \dots, n. \quad (3.1)$$

The innovations,  $\varepsilon_i$ , are independent of the filtration  $\mathcal{F}_{i-1}$ , which is the sigma-field generated by  $x_1, \dots, x_i$  and  $\varepsilon_1, \dots, \varepsilon_{i-1}$ . Moreover,  $\varepsilon_i$  are identically distributed with mean zero and variance  $\sigma^2$ , so that  $\varepsilon_i/\sigma$  has known symmetric density  $f$  and distribution function  $F(c) = P(\varepsilon_i \leq \sigma c)$ . In practice, the distribution  $F$  will often be standard normal.

We will think of the outliers as pairs of observations  $(y_i, x_i)$  that do not conform with the model (3.1). In other words, a pair of observations  $(y_i, x_i)$  gives us an outlier if the scaled innovation  $\varepsilon_i/\sigma$  does not conform with 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  $(y_1, x_1), \dots, (y_n, x_n)$  are unrelated. Thus, if the innovation  $\varepsilon_i$  is classified as an outlier, then the pair of observations  $(y_i, x_i)$  is dropped. We can interpret this as an innovation 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 will include lagged dependent variables. For instance, for a first order autoregression  $x_i = y_{i-1}$ . We distinguish between innovative outliers and additive outlier. Classifying the innovation  $\varepsilon_i$  as an outlier, has the consequence that we discard the evaluation of the dynamics from  $y_{i-1}$  to  $y_i$  without discarding

the observations  $y_{i-1}$  and  $y_i$ . 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 innovative outliers  $\varepsilon_i$  and  $\varepsilon_{i+1}$ . Discarding these, the observation  $y_i$  will not appear.

We consider algorithms using absolute residuals and calculation of least squares estimators from selected observations. Both these choices implicitly assume a symmetric density: If non-outlying innovations were asymmetric then the symmetrically truncated innovations would in general be asymmetric and the least squares estimator for location would be biased.

With symmetry the absolute value errors  $|\varepsilon_i|/\sigma$  have density  $\mathbf{g}(c) = 2\mathbf{f}(c)$  and distribution function  $\mathbf{G}(c) = \mathbf{P}(|\varepsilon_1| \leq \sigma c) = 2\mathbf{F}(c) - 1$ . We define  $\psi = \mathbf{G}(c)$  so that  $c$  is the  $\psi$  quantile

$$c = \mathbf{G}^{-1}(\psi) = \mathbf{F}^{-1}\{(1 + \psi)/2\}, \quad \psi \in [0, 1[,$$

while the probability of exceeding the cut-off value  $c$  is

$$\gamma = 1 - \psi = 1 - \mathbf{G}(c).$$

Define also the truncated moments

$$\tau = \int_{-c}^c u^2 \mathbf{f}(u) du, \quad \varkappa = \int_{-c}^c u^4 \mathbf{f}(u) du, \quad \kappa = \int_{-\infty}^{\infty} u^4 \mathbf{f}(u) du, \quad (3.2)$$

and the conditional variance of  $\varepsilon_1/\sigma$  given  $\{|\varepsilon_1| \leq \sigma c\}$  as

$$\zeta^2 = \tau/\psi = \int_{-c}^c u^2 \mathbf{f}(u) du / \mathbf{P}(|\varepsilon_1/\sigma| \leq c), \quad (3.3)$$

which will serve as a bias correction for the variance estimators based on the truncated sample. Define also the quantity

$$\zeta = 2c(c^2 - \zeta^2)\mathbf{f}(c). \quad (3.4)$$

In this paper we focus on the normal reference distribution. The truncated moments then simplify as follows

$$\tau = \psi - 2c\mathbf{f}(c), \quad \varkappa = 3\psi - 2c(c^2 + 3)\mathbf{f}(c), \quad \kappa = 3. \quad (3.5)$$

## 4 Some outlier detection algorithms

Least squares estimators are known to be fragile with respect to outliers. A number of robust methods have been developed over the years. We study a variety of estimators with the common property that outlying observations are skipped.

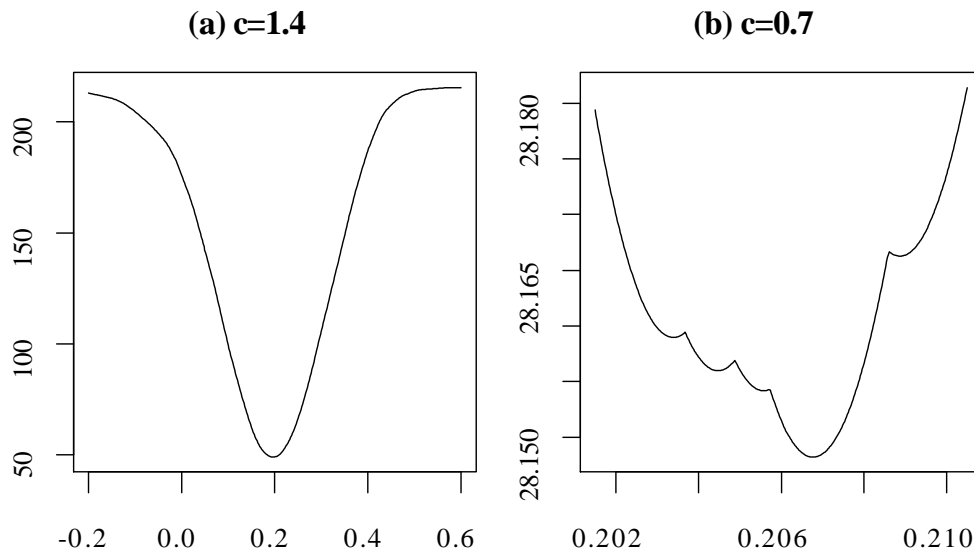


Figure 2: Huber-skip objective function for Fulton fish data.

## 4.1 M-estimators

Huber (1964) introduced M-estimators as a class of maximum likelihood type estimators for location. The M-estimator for the regression model (3.1) is defined as the minimizer of

$$R_n(\beta) = n^{-1} \sum_{i=1}^n \rho(y_i - x_i' \beta). \quad (4.1)$$

for some absolutely continuous and non-negative criterion function  $\rho$ . In particular, the least squares estimator arises when  $\rho(u) = u^2$  while the median or least absolute deviation estimator arises for  $\rho(u) = |u|$ . We will pursue the idea of hard rejection of outliers through the non-convex Huber-skip criterion function  $\rho(u) = u^2 1_{(|u| \leq \sigma c)} + c^2 1_{(|u| > \sigma c)}$  for some cut-off  $c > 0$  and known scale  $\sigma$ .

The objective function of the Huber-skip M-estimator is non-convex. Figure 2 illustrates the objective function for the Fish data.<sup>5</sup> The specification is as in equation (2.1). All parameters apart from that on  $q_{t-1}$  are held fixed at the values in (2.1). Panel (a) shows that when the cut-off  $c$  is large the Huber-skip is quadratic in the central part. Panel (b) shows that when the cut-off  $c$  is smaller the objective function is non-differentiable in a finite number of points. Subsequently, we consider estimators that are easier to compute and apply for unknown scale, while hopefully preserving some useful robustness properties.

The asymptotic theory of M-estimators has been studied in some detail for the situation without outliers. Huber (1964) proposed a theory for location models and convex criterion functions  $\rho$ . Jurečková and Sen (1996, p. 215f) analyzed the regression problem with convex criterion functions. Non-convex criterion functions were considered for location models in Jurečková and Sen (1996, p. 197f), see also Jurečková, Sen, and Pícek (2012). Chen and Wu (1988) showed strong consistency of M-estimators for general criterion functions with i.i.d. or

<sup>5</sup>Graphics were done using R 3.1.1, see R Development Core Team (2014).

deterministic regressors, while time series regression is analyzed in Johansen and Nielsen (2014b). We review the latter theory in §7.1.

## 4.2 Huber-skip estimators

We consider some estimators that involve skipping data points, but are not necessarily M-estimators. The objective functions have binary stochastic weights  $v_i$  for each observation. These weights are defined in various ways below. In all cases the objective function is

$$R_n(\beta) = n^{-1} \sum_{i=1}^n \{(y_i - x_i' \beta)^2 v_i + c^2(1 - v_i)\}. \quad (4.2)$$

The weights  $v_i$  may depend on  $\beta$ . The first example is the *Huber-skip M-estimator* which depends on a cut-off point  $c$ , where

$$v_i = 1_{(|y_i - x_i' \beta| \leq c\sigma)}. \quad (4.3)$$

Another example is the *Least Trimmed Squares* estimator of Rousseeuw (1984) which depends on an integer  $k \leq n$ , where

$$v_i = 1_{(|y_i - x_i' \beta| \leq \xi_{(k)})}, \quad (4.4)$$

for  $\xi_{(k)}$  chosen as the  $k$ -th smallest order statistic of absolute residuals  $\xi_i = |y_i - x_i' \beta|$  for  $i = 1, \dots, n$ . Given an integer  $k \leq n$  we can find  $\psi$  and  $c$  so  $k/n = \psi = \mathbf{G}^{-1}(c)$ , and  $\psi, c, k$  are different ways of calibrating the methods. In either case, once the regression estimator  $\hat{\beta}$  has been determined the scale can be estimated by

$$\hat{\sigma}^2 = \zeta^{-2} (\sum_{i=1}^n v_i)^{-1} \{ \sum_{i=1}^n v_i (y_i - x_i' \hat{\beta})^2 \}, \quad (4.5)$$

where  $\zeta^2 = \tau/\psi$  is the consistency correction factor defined in (3.3).

For the Least Trimmed Squares estimator it holds that  $\sum_{i=1}^n (1 - v_i) = n - k$ . Thus, the last term in the objective function (4.2) does not depend on  $\beta$ , so that it is equivalent to optimize

$$R_{n,LTS}(\beta) = n^{-1} \sum_{i=1}^n v_i (y_i - x_i' \beta)^2. \quad (4.6)$$

The Least Trimmed Squares weight (4.4) is scale invariant in contrast to the Huber-skip M-estimator. It is known to have breakdown point of  $\gamma = 1 - \psi = 1 - k/n$  for  $\psi < 1/2$ , see Rousseeuw and Leroy (1987, §3.4). An asymptotic theory is provided by Vřsek (2006a,b,c). The estimator is computed through a binomial search algorithm which is uncomputable in most practical situations, see Maronna, Martin and Yohai (2006, §5.7) for a discussion. A number of iterative approximations have been suggested such as the Fast LTS algorithm by Rousseeuw and van Driessen (1998). This leaves additional questions with respect to the properties of the approximating algorithms.

If the weights  $v_i$  do not depend on  $\beta$ , the objective function has a least squares solution

$$\hat{\beta} = (\sum_{i=1}^n v_i x_i x_i')^{-1} (\sum_{i=1}^n v_i x_i y_i). \quad (4.7)$$



From this the variance estimator (4.5) can be computed. Examples include *1-step Huber-skip M-estimators* based on initial estimators  $\tilde{\beta}, \tilde{\sigma}^2$ , where

$$v_i = 1_{(|y_i - x_i' \tilde{\beta}| \leq c \tilde{\sigma})}, \quad (4.8)$$

and *1-step Huber-skip L-estimators* based on an initial estimator  $\tilde{\beta}$  and a cut-off  $k < n$ , which defines the  $k$ -th smallest order statistic  $\tilde{\xi}_{(k)}$  of absolute residuals  $\xi_i = |y_i - x_i' \tilde{\beta}|$ , where

$$v_i = 1_{(|y_i - x_i' \tilde{\beta}| \leq \tilde{\xi}_{(k)})}. \quad (4.9)$$

These estimators are computationally attractive, but require a good starting point. They can also be iterated. As before, we see that the 1-step L-estimator does not require an initial scale estimator in contrast to the 1-step M-estimator.

Robustified least squares arises if the initial estimators  $\tilde{\beta}, \tilde{\sigma}^2$  are the full-sample least squares estimators. This relates to the estimation procedure for the Fulton Fish Market data in §2. This approach can be fragile, especially when there are more than a few outliers, see Welsh and Ronchetti (2002) for a discussion.

The 1-step estimators relate to the 1-step M-estimators of Bickel (1975), although he was primarily concerned with smooths weights  $v_i$ . His idea was to apply preliminary estimators  $\hat{\beta}^{(0)}, (\hat{\sigma}^{(0)})^2$  and then define the 1-step estimator  $\hat{\beta}^{(1)}$  by linearising the first order condition. He also suggested iteration, but no results were given.

Ruppert and Carroll (1980) studied a related 1-step L-estimator for which fixed proportions of negative and positive residuals are skipped. Following their suggestion we refer to the estimator with weights (4.9) as a 1-step Huber-skip L-estimator, because the objective function is defined by weights involving an order statistics. We note that there is a mismatch in the nomenclature of L and M-estimators. Jaeckel (1971) defined L-estimators for location problems in terms of the estimator, whereas Huber (1964) defined M-estimators in terms of the objective function. Thus, the Least Trimmed Squares estimator is not classified as an L-estimator, although its objective function is a quadratic combination of order statistics.

### 4.3 Some statistical algorithms

We give three statistical algorithms involving iteration of 1-step Huber-skip estimators. These are the iterated 1-step Huber-skip M-estimator, the Impulse Indicator Saturation, and the Forward Search.

The *1-step Huber-skip estimators* are amenable to iteration. Here we consider iterated Huber-skip M-estimators.

**Algorithm 4.1** *Iterated 1-step Huber-skip M-estimator.* Choose a cut-off  $c > 0$ .

1. Choose initial estimators  $\hat{\beta}^{(0)}, (\hat{\sigma}^{(0)})^2$  and let  $m = 0$ .
2. Define indicator variables  $v_i^{(m)}$  as in (4.8), replacing  $\tilde{\beta}, \tilde{\sigma}^2$  by  $\hat{\beta}^{(m)}, (\hat{\sigma}^{(m)})^2$ .
3. Compute least squares estimators  $\hat{\beta}^{(m+1)}, (\hat{\sigma}^{(m+1)})^2$  as in (4.7), (4.5) replacing  $v_i$  by  $v_i^{(m)}$ .
4. Let  $m = m + 1$  and repeat 2 and 3.

The Iteration Algorithm 4.1 does not have a stopping rule. This leaves the questions whether the algorithm convergences with increasing  $m$  and  $n$  and in which sense it approximates the Huber-skip estimator.

The *Impulse Indicator Saturation* algorithm has its roots in the empirical work of Hendry (1999) and Hendry, Johansen and Santos (2008). It is a 1-step M-estimator, where the initial estimator is formed by exploiting in a simple way the assumption, that a subset of observations is free of outliers. The idea is to divide the sample into two sub-samples. Then run a regression on each sub-sample and use this to find outliers in the other sub-sample.

**Algorithm 4.2 *Impulse Indicator Saturation.*** Choose a cut-off  $c > 0$ .

1.1. Split indices in sets  $\mathcal{I}_j$ , for  $j = 1, 2$ , of  $n_j$  observations.

1.2. Calculate the least squares estimators for  $(\beta, \sigma^2)$  based upon sample  $\mathcal{I}_j$  as

$$\hat{\beta}_j = (\sum_{i \in \mathcal{I}_j} x_i x_i')^{-1} (\sum_{i \in \mathcal{I}_j} x_i y_i), \quad \hat{\sigma}_j^2 = \frac{1}{n_j} \sum_{i \in \mathcal{I}_j} (y_i - x_i' \hat{\beta}_j)^2.$$

1.3. Define indicator variables for each observation

$$\hat{v}_i^{(-1)} = 1_{(i \in \mathcal{I}_1)} 1_{(|y_i - x_i' \hat{\beta}_2| \leq c \hat{\sigma}_2)} + 1_{(i \in \mathcal{I}_2)} 1_{(|y_i - x_i' \hat{\beta}_1| \leq c \hat{\sigma}_1)}, \quad (4.10)$$

1.4. Compute least squares estimators  $\hat{\beta}^{(0)}, (\hat{\sigma}^{(0)})^2$  using (4.7), (4.5), replacing  $v_i$  by  $\hat{v}_i^{(-1)}$  and let  $m = 0$ .

2. Define indicator variables  $v_i^{(m)} = 1_{(|y_i - x_i' \hat{\beta}^{(m)}| \leq c \hat{\sigma}^{(m)})}$  as in (4.8).

3. Compute least squares estimators  $\hat{\beta}^{(m+1)}, (\hat{\sigma}^{(m+1)})^2$  as in (4.7), (4.5), replacing  $v_i$  by  $v_i^{(m)}$ .

4. Let  $m = m + 1$  and repeat 2 and 3.

Due to its split half approach to the initial estimation, the Impulse Indicator Saturation may be more robust than robustified least squares. The Impulse Indicator Saturation estimator will work best when the outliers are known to be in a particular subset of the observations. For instance, consider the split half case where index sets  $\mathcal{I}_1, \mathcal{I}_2$  are chosen as the first half and the second half of the observations, respectively. Then the algorithm has a good ability to detect for instance a level shift half way through the second sample, while it is poor at detecting outliers scattered throughout both samples, because both sample halves are contaminated. If the location of the contamination is unknown, one will have to iterate over the choice of the initial sets  $\mathcal{I}_1, \mathcal{I}_2$ . This is what the more widely used Autometrics algorithm does, see Doornik (2009) and Doornik and Hendry (2014).

The *Forward Search* algorithm is an iterated 1-step Huber-skip L-estimator suggested for the multivariate location model by Hadi (1992) and for multiple regression by Hadi and Simonoff (1993) and developed further by Atkinson and Riani (2000), see also Atkinson, Riani and Cerioli (2010). The algorithm starts with a robust estimate of the regression parameters. This is used to construct the set of observations with the smallest  $m_0$  absolute residuals. We then run a regression on those  $m_0$  observations and compute absolute residuals of all  $n$  observations. The observations with  $m_0 + 1$  smallest residuals are then selected, and a new regression is performed on these  $m_0 + 1$  observations. This is then iterated. Since the estimator based on the  $m_0 + 1$  observation is computed in terms of the order statistic based on the estimator for the  $m_0$  observation, it is a 1-step Huber-skip L-estimator. When iterating the order of the order statistics is gradually expanding.

**Algorithm 4.3 *Forward Search.***

1. Choose an integer  $m_0 < n$  and an initial estimators  $\hat{\beta}^{(m_0)}$ , and let  $m = m_0$ .

- 2.1. Compute absolute residuals  $\widehat{\xi}_i^{(m)} = |y_i - x_i' \widehat{\beta}^{(m)}|$  for  $i = 1, \dots, n$ .
- 2.2. Find the  $(m + 1)$ th smallest order statistic  $\widehat{z}^{(m)} = \widehat{\xi}_{(m+1)}^{(m)}$ .
- 2.3. Define indicator variables  $v_i^{(m)} = 1_{(|y_i - x_i' \widehat{\beta}^{(m)}| \leq \widehat{z}^{(m)})}$  as in (4.9).
3. Compute least squares estimators  $\widehat{\beta}^{(m+1)}, (\widehat{\sigma}^{(m+1)})^2$  as in (4.7), (4.5) replacing  $v_i$  by  $v_i^{(m)}$ .
4. If  $m < n$  let  $m = m + 1$  and repeat 2 and 3.

The Forward Search Algorithm 4.3 has a finite number of steps. It terminates when  $m = n - 1$  and  $\widehat{\beta}^{(n)}, (\widehat{\sigma}^{(n)})^2$  are the full sample least squares estimators. Applying the algorithm for  $m = m_0, \dots, n - 1$ , results in sequences of least squares estimators  $\widehat{\beta}^{(m)}, (\widehat{\sigma}^{(m)})^2$  and order statistics  $\widehat{z}^{(m)} = \widehat{\xi}_{(m+1)}^{(m)}$ .

The idea of the Forward Search is to monitor the plot of scaled forward residuals  $\widehat{z}^{(m)}/\widehat{\sigma}^{(m)}$ . For each  $m$  we can find the asymptotic distribution of  $\widehat{z}^{(m)}/\widehat{\sigma}^{(m)}$  and add a curve of point-wise  $p$ -quantiles as a function of  $m$  for some  $p$ . The first  $m$  for which  $\widehat{z}^{(m)}/\widehat{\sigma}^{(m)}$  exceeds the quantile curve is the estimate  $\widehat{m}$  of the number of non-outliers. Asymptotic theory for the forward residuals  $\widehat{z}^{(m)}/\widehat{\sigma}^{(m)}$  is reviewed in §8.3. A theory for the estimator  $\widehat{m}$  is given in §10.

A variant of the Forward Search advocated by Atkinson and Riani (2000) is to use the minimum deletion residuals  $\widehat{d}^{(m)} = \min_{i \notin S^{(m)}} \widehat{\xi}_i^{(m)}$  instead of the forward residuals  $\widehat{z}^{(m)}$ .

## 5 Overview of the results for the location case

We give an overview of the asymptotic theory for the M-type Huber-skip estimators for the location problem, where  $x_i' \beta$  reduces to a location parameter  $\mu$ . The theory evolves around two asymptotic results. The first is an asymptotic distribution for the M-estimator. The second is an asymptotic expansion for 1-step M-estimators. The iterated 1-step M-estimators are found to converge to the M-estimator.

Huber (1964) proposed an asymptotic theory for M-estimators with convex objective function in location models. His proof did not extend to the Huber-skip M-estimator based on the weights (4.3). Instead he assumed consistency and conjectured that the asymptotic distribution would be, for symmetric  $f$ ,

$$n^{1/2}(\widehat{\mu} - \mu) = \frac{1}{\psi - 2cf(c)} n^{-1/2} \sum_{i=1}^n \varepsilon_i 1_{(|\varepsilon_i| \leq \sigma c)} + o_P(1) \xrightarrow{D} \mathbf{N}\left[0, \frac{\tau \sigma^2}{\{\psi - 2cf(c)\}^2}\right]. \quad (5.1)$$

This result is generalized to time series regression in Theorem 7.1.

In the situation with normal errors  $\tau = \psi - 2cf(c)$ , see (3.5), the asymptotic variance in (5.1) reduces to  $\sigma^2/\tau$ . The efficiency relative to the sample average, which is the least squares estimator, is therefore  $\tau$ . The bottom curve plotted in Figure 3 shows the efficiency as a function of  $\psi$ . The least trimmed squares estimator has the same asymptotic distribution.

Next, consider the 1-step Huber-skip M-estimator based on weights (4.8). It has asymptotic expansion linking the updated estimator  $\widehat{\mu}^{(1)}$  with the initial estimator  $\widehat{\mu}^{(0)}$  through

$$n^{1/2}(\widehat{\mu}^{(1)} - \mu) = \frac{1}{\psi} n^{-1/2} \sum_{i=1}^n \varepsilon_i 1_{(|\varepsilon_i| \leq \sigma c)} + \frac{2cf(c)}{\psi} n^{1/2}(\widehat{\mu}^{(0)} - \mu) + o_P(1), \quad (5.2)$$

see Theorem 7.2 for regression.

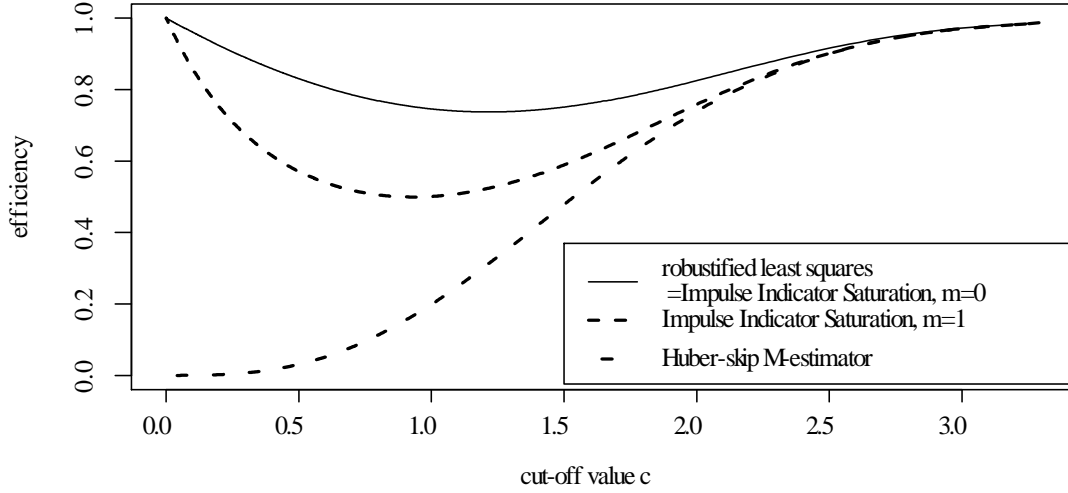


Figure 3: The efficiency of robustified least squares, the Impulse Indicator Saturation, and the Huber-skip M-estimator relative to full sample least squares when the reference distribution is normal.

Robustified least squares arises, if we choose the initial estimator  $\hat{\mu}^{(0)}$  as the least squares estimator. In that case we get the expansion

$$n^{1/2}(\hat{\mu}^{(1)} - \mu) = \frac{1}{\psi} n^{-1/2} \sum_{i=1}^n \varepsilon_i 1_{(|\varepsilon_i| \leq \sigma c)} + \frac{2\text{cf}(c)}{\psi} n^{-1/2} \sum_{i=1}^n \varepsilon_i + o_{\mathbb{P}}(1). \quad (5.3)$$

We can use the Central Limit Theorem to show asymptotic normality of the estimator. The asymptotic variance follows in Theorem 7.3. The efficiency relative to least squares estimation is shown as the top curve in Figure 3.

Starting with other estimators give different asymptotic variances. An example is the Impulse Indicator Saturation Algorithm 4.2. Theorem 7.4 shows that the initial split-half estimator  $\hat{\mu}^{(0)}$  has the same asymptotic distribution as the robustified least squares estimator. The updated 1-step estimator  $\hat{\mu}^{(1)}$  is slightly less efficient, as shown by the middle curve in Figure 3, but hopefully more robust.

The 1-step M-estimator can be iterated along the lines of Algorithm 4.1. This iteration has a fixed point  $\hat{\mu}^*$  solving the equation

$$n^{1/2}(\hat{\mu}^* - \mu) = \frac{1}{\psi} n^{-1/2} \sum_{i=1}^n \varepsilon_i 1_{(|\varepsilon_i| \leq \sigma c)} + \frac{2\text{cf}(c)}{\psi} n^{1/2}(\hat{\mu}^* - \mu) + o_{\mathbb{P}}(1), \quad (5.4)$$

see Theorem 7.6. Thus, any influence of the initial estimator is lost through iteration. Solving this equation gives

$$n^{1/2}(\hat{\mu}^* - \mu) = \frac{1}{\psi - 2\text{cf}(c)} \sum_{i=1}^n \varepsilon_i 1_{(|\varepsilon_i| \leq \sigma c)} + o_{\mathbb{P}}(1), \quad (5.5)$$

with the same leading term as the Huber-skip M-estimator in (5.1).

## 6 Preliminary asymptotic results

We present the main ingredient for the asymptotic theory.

### 6.1 Assumptions on regressors and density

The innovations  $\varepsilon_i$  and regressors  $x_i$  must satisfy moment assumptions. The innovations  $\varepsilon_i$  have symmetric density with derivative satisfying boundedness and tail conditions. Related conditions on the density are often seen in the literatures on empirical processes and quantile processes. These conditions are satisfied for the normal distribution and t-distributions, see Johansen and Nielsen (2014a) for a discussion. For the iterated estimator we need an assumption of unimodality. The minimal assumptions vary for the different estimators, as explored in Johansen and Nielsen (2009, 2013, 2014a, 2014b) for 1-step Huber-skip M-estimators, for iterated 1-step Huber-skip M-estimators, for the Forward Search and for general M-estimators, respectively.

For this presentation we simply assume a normal reference distribution, which, of course, is most used in practice. With normality we avoid a somewhat tedious discussion of existence of moments of a certain order. The regressors can be temporally dependent and possibly deterministically or stochastically trending.

**Assumption 6.1** *Let  $\mathcal{F}_i$  be the filtration generated by  $x_1, \dots, x_{i+1}$  and  $\varepsilon_1, \dots, \varepsilon_i$ . Assume*

- (i) *innovations  $\varepsilon_i/\sigma$  are independent of  $\mathcal{F}_{i-1}$  and standard normal;*
- (ii) *regressors  $x_i$  satisfy, for some non-stochastic normalisation matrix  $N \rightarrow 0$  and random matrices  $V, \Sigma, \mu$ , the following joint convergence results hold*

- (a)  $V_n = N' \sum_{i=1}^n x_i \varepsilon_i \xrightarrow{D} V$ ;
- (b)  $\Sigma_n = N' \sum_{i=1}^n x_i x_i' N \xrightarrow{D} \Sigma \stackrel{a.s.}{>} 0$ ;
- (c)  $n^{-1/2} N' \sum_{i=1}^n x_i \xrightarrow{D} \mu$ ;
- (d)  $\max_{i \leq n} |n^{1/2} N' x_i| = o_{\mathbb{P}}(n^\phi)$  for all  $\phi > 0$ ;
- (e)  $n^{-1} \mathbb{E} \sum_{i=1}^n |n^{1/2} N' x_i|^q = O(1)$  for some  $q > 9$ .

The Assumption 6.1(ii) for the regressors are 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 also holds for quite general regressors including polynomial 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.

We note that Assumption 6.1 implies Assumption 3.1(i, ii) of Johansen and Nielsen (2014a) by choosing  $\eta = 1/4$ ,  $q_0 = q > 9$  and a  $\kappa = \phi$  for a small  $\phi > 0$  so that  $0 < \kappa$  is bounded by the minimum of  $\eta$ ,  $1/(1 + \dim x)$  and  $(q - 9)/(q - 1)$ .

### 6.2 Weighted and marked empirical processes

The asymptotic analysis of Huber-skip estimators is concerned with a class of weighted and marked empirical processes. The 1-step estimators for  $\beta$  and  $\sigma^2$  have estimation errors that can be expressed in terms of statistics of the form

$$\sum_{i=1}^n v_i, \quad \sum_{i=1}^n v_i x_i \varepsilon_i, \quad \sum_{i=1}^n v_i x_i x_i', \quad \sum_{i=1}^n v_i \varepsilon_i^2, \quad (6.1)$$

where  $v_i$  are indicator functions for small residuals. Such sums of indicator functions are the basis for empirical process. 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  $\varepsilon_i$  and  $\varepsilon_i^2$  are said to be marks. For M-type estimators, the indicator functions have the form

$$v_{M,i} = 1_{\{|\varepsilon_i - x'_i(\tilde{\beta} - \beta)| \leq \sigma c + (\tilde{\sigma} - \sigma)c\}} = 1_{\{|\varepsilon_i - x'_i N \tilde{b}| \leq \sigma c + n^{-1/2} \tilde{a} c\}}, \quad (6.2)$$

which allows for estimation uncertainty  $\tilde{b} = N^{-1}(\tilde{\beta} - \beta)$  and  $\tilde{a} = n^{1/2}(\tilde{\sigma} - \sigma)$  in the regression coefficient  $\beta$  and in the scale  $\sigma$ . For L-type estimators the indicators are

$$v_{L,i} = 1_{\{|\varepsilon_i - x'_i(\tilde{\beta} - \beta)| \leq \tilde{\xi}_{(k)}\}} = 1_{\{|\varepsilon_i - x'_i N \tilde{b}| \leq \sigma(c + n^{-1/2} \tilde{d})\}}, \quad (6.3)$$

which allows for estimation uncertainty  $\tilde{b} = N^{-1}(\tilde{\beta} - \beta)$  and  $\tilde{d} = n^{1/2}(\tilde{\xi}/\sigma - c)$  in the regression coefficient  $\beta$  and in the quantile  $\xi$ .

We will need an asymptotic linearization of the statistics (6.1) with respect to the estimation uncertainty. For this purpose we start by considering weights

$$v_i^{b,c,d} = 1_{\{|\varepsilon_i - x'_i N b| \leq \sigma(c + n^{-1/2} d)\}}, \quad (6.4)$$

where the estimation uncertainty is replaced by bounded, deterministic terms  $b, d$ . Subsequently, we apply the result to M-type and L-type estimators, by replacing  $b$  by  $\tilde{b}$  and  $d$  by  $\tilde{a}c/\sigma$  and  $\tilde{d}$ , respectively.

The following asymptotic expansion is a version of Lemma D.5 of Johansen and Nielsen (2014a) formulated under the present simplified Assumption 6.1.

**Theorem 6.1** (Johansen and Nielsen, 2014a, Lemma D.5) *Suppose Assumption 6.1 holds. Consider the product moments (6.1) with weights  $v_i^{b,c,d}$  given by (6.4) and expansions*

$$\begin{aligned} n^{-1/2} \sum_{i=1}^n v_i^{b,c,d} &= n^{-1/2} \sum_{i=1}^n 1_{\{|\varepsilon_i| \leq \sigma c\}} + 2f(c)d + R_v(b, c, d), \\ n^{-1/2} \sum_{i=1}^n v_i^{b,c,d} \varepsilon_i^2 &= n^{-1/2} \sum_{i=1}^n \varepsilon_i^2 1_{\{|\varepsilon_i| \leq \sigma c\}} + 2\sigma^2 c^2 f(c)d + R_{v\varepsilon\varepsilon}(b, c, d), \\ N' \sum_{i=1}^n v_i^{b,c,d} x_i \varepsilon_i &= N' \sum_{i=1}^n x_i \varepsilon_i 1_{\{|\varepsilon_i| \leq \sigma c\}} + 2cf(c)N' \sum_{i=1}^n x_i x'_i N b + R_{vx\varepsilon}(b, c, d), \\ N' \sum_{i=1}^n v_i^{b,c,d} x_i x'_i N &= \psi N' \sum_{i=1}^n x_i x'_i N + R_{vxx}(b, c, d). \end{aligned}$$

Let

$$R(b, c, d) = |R_v(b, c, d)| + |R_{v\varepsilon\varepsilon}(b, c, d)| + |R_{vx\varepsilon}(b, c, d)| + |R_{vxx}(b, c, d)|$$

Then it holds for all (large)  $B > 0$ , all (small)  $\eta > 0$  and  $n \rightarrow \infty$  that

$$\sup_{|b|, |d| \leq n^{1/4 - \eta} B} \sup_{0 < c < \infty} R(b, c, d) = o_{\mathbb{P}}(1). \quad (6.5)$$

In particular, for bounded  $c$ , then

$$\sup_{|a|, |b| \leq n^{1/4 - \eta} B} \sup_{0 < c \leq B} R(b, c, ac/\sigma) = o_{\mathbb{P}}(1). \quad (6.6)$$

Theorem 6.1 is proved by a chaining argument. The idea is to cover the domain of  $b, d$  with a finite number of balls. The supremum over the large compact set can then be replaced by considering the maximum value over the centers of the balls and the maximum of the variation within balls. By subtracting the compensators of the product moments we turn them into martingales. The argument will therefore be a consideration of the tail behaviour of the maximum of a family of martingales using the iterated martingale inequality presented in §6.3 and Taylor expansions of the compensators.

Related results of Theorem 6.1 are considered in the literature. Koul and Ossiander (1994) considered weighted empirical processes without marks and with  $\eta > 1/4$ . Johansen and Nielsen (2009) considered the situation (6.6) for fixed  $c$  and with  $\eta > 1/4$ .

### 6.3 An iterated martingale inequality

We present an iterated martingale inequality, which can be used to assess the tail behaviour of the maximum of a family of martingales. It builds on an exponential martingale inequality by Bercu and Touati (2008).

**Theorem 6.2** (Bercu and Touati, 2008, Theorem 2.1) *For  $i = 1, \dots, n$  let  $(m_i, \mathcal{F}_i)$  be a locally square integrable martingale difference. Then, for all  $x, y > 0$ ,*

$$\mathbf{P}\left[\left|\sum_{i=1}^n m_i\right| \geq x, \sum_{i=1}^n \{m_i^2 + \mathbf{E}(m_i^2 | \mathcal{F}_{i-1})\} \leq y\right] \leq 2 \exp\left(-\frac{x^2}{2y}\right).$$

In order to bound a family of martingales it is useful to iterate this martingale inequality to get the following iterated martingale inequality.

**Theorem 6.3** (Johansen and Nielsen, 2014a, Theorem 5.2.) *For  $\ell = 1, \dots, L$  let  $z_{\ell,i}$  be  $\mathcal{F}_i$ -adapted so  $\mathbf{E}z_{\ell,i}^{2\bar{r}} < \infty$  for some  $\bar{r} \in \mathbb{N}$ . Let  $D_r = \max_{1 \leq \ell \leq L} \sum_{i=1}^n \mathbf{E}(z_{\ell,i}^{2r} | \mathcal{F}_{i-1})$  for  $1 \leq r \leq \bar{r}$ . Then, for all  $\kappa_0, \kappa_1, \dots, \kappa_{\bar{r}} > 0$ , it holds*

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

Theorem 6.3 contains parameters  $\kappa_0, \kappa_1, \dots, \kappa_{\bar{r}}$ , which can be chosen in various ways. We give two examples taken from Theorems 5.3, 5.4 of Johansen and Nielsen (2014a)

The first example is to show that the remainder terms in Theorem 6.1 are uniformly small. In the proof we consider a family of size  $L = O(n^\lambda)$  where  $\lambda > 0$  depends on the dimension of the regressor and seek to prove that the maximum of the family of martingales is of order  $o_{\mathbf{P}}(n^{1/2})$ . Choosing  $\kappa_q = (\kappa n^{1/2})^{2q} (28\lambda \log n)^{1-2q}$ , so that  $\kappa_q^2 / \kappa_{q+1} = 28\lambda \log n$  and  $\kappa_0 = \kappa n^{1/2}$ , a result of that type follows.

The second example is to show that the empirical processes (6.1) are tight. In this case the family is of fixed size  $L$ , and now the probability that the maximum of the family of martingales is larger than  $\kappa n^{1/2}$  has to be bounded by a small number. Choosing  $\kappa_q = \kappa n^{2q-1} \theta^{1-2q}$  so  $\kappa_q^2 / \kappa_{q+1} = \kappa \theta$  and  $\kappa_0 = \kappa n^{1/2}$  a result of that type follows.

## 7 Asymptotic results for Huber-skip M-estimators

We consider recent results on the Huber-skip M-estimator as well as for 1-step Huber-skip M-estimators and iterations thereof.

### 7.1 Huber-skip M-estimators

The Huber-skip M-estimator is the solution to the optimization problem (4.2) with weights (4.3). Since this problem is non-convex we need an additional assumption that bounds the frequency of small regressors. That bound involves a function that is an approximate inverse of the function  $\lambda_n(\alpha)$  appearing in the analysis of S-estimators by Davies (1990), see also Chen and Wu (1988). The bound can be satisfied for 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 (2014b).

**Theorem 7.1** (*Johansen and Nielsen, 2014b, Theorems 1,2,3*) *Consider the Huber-skip M-estimator defined from (4.2), (4.3). Suppose Assumption 6.1 holds and that the frequency of small regressors is bounded as outlined above. Then any minimizer of the objective function (4.2) has a measurable version and satisfies*

$$N^{-1}(\widehat{\beta} - \beta) = \frac{1}{\psi - 2\text{cf}(c)} \Sigma_n^{-1} N' \sum_{i=1}^n x_i \varepsilon_i 1_{(|\varepsilon_i| \leq \sigma c)} + o_{\mathbb{P}}(1).$$

If, in addition the regressors are stationary then

$$n^{1/2}(\widehat{\beta} - \beta) \xrightarrow{\mathbb{D}} \mathbf{N}\{0, \Sigma^{-1} \sigma^2 / \tau\}.$$

Theorem 7.1 proves the conjecture (5.1) of Huber (1964) for time series regression. The regularity conditions on the regressors are much weaker than those normally considered in for instance Chen and Wu (1988), Liese and Vajda (1994), Maronna, Martin, and Yohai (2006), Huber and Ronchetti (2009), and Jurečková, Sen, and Picek (2012). Theorem 7.1 extends to non-normal, but symmetric densities and even to non-symmetric densities and objective function, by introducing a bias correction.

Theorem 7.1 is proved in three steps. First, it is shown that  $\widehat{\beta}$  is tight, that is  $N^{-1}(\widehat{\beta} - \beta) = O_{\mathbb{P}}(n^{1/2})$ , through a geometric argument that requires the assumption to the frequency of small regressors. Secondly, it is shown that  $\widehat{\beta}$  is consistent, in the sense that  $N^{-1}(\widehat{\beta} - \beta) = O_{\mathbb{P}}(n^{1/2-\tau})$  for any  $\tau < 1/4$ , using the iterated martingale inequality of Theorem 6.3. Finally, the presented expansion of Theorem 7.1 is proved, again using Theorem 6.3.

### 7.2 1-step Huber-skip M-estimators

The asymptotic theory of the 1-step Huber-skip M-estimator for regression is given in Johansen and Nielsen (2009). The main result is a stochastic expansion of the updated estimation error in terms of a kernel and the original estimation error. It follows from a direct application of Theorem 6.1.



**Theorem 7.2** (Johansen and Nielsen, 2009, Corollary 1.2) Consider the 1-step Huber-skip M-estimators  $\widehat{\beta}^{(1)}$ ,  $\widehat{\sigma}^{(1)}$  defined by (4.7), (4.5) with weights (4.8). Suppose Assumption 6.1 holds and that  $N^{-1}(\widehat{\beta}^{(0)} - \beta)$  and  $n^{1/2}(\widehat{\sigma}^{(0)} - \sigma)$  are  $O_P(1)$ . Then

$$N^{-1}(\widehat{\beta}^{(1)} - \beta) = \frac{1}{\psi} \Sigma_n^{-1} N' \sum_{i=1}^n x_i \varepsilon_i 1_{(|\varepsilon_i| \leq \sigma c)} + \frac{2\text{cf}(c)}{\psi} N^{-1}(\widehat{\beta}^{(0)} - \beta) + o_P(1). \quad (7.1)$$

$$n^{1/2}(\widehat{\sigma}^{(1)} - \sigma) = \frac{1}{2\tau\sigma} n^{-1/2} \sum_{i=1}^n (\varepsilon_i^2 - \sigma^2 \frac{\tau}{\psi}) 1_{(|\varepsilon_i| \leq \sigma c)} + \frac{\zeta}{2\tau} n^{1/2}(\widehat{\sigma}^{(0)} - \sigma) + o_P(1) \quad (7.2)$$

Theorem 7.2 generalises the statement (5.2) for the location problem. Theorem 7.2 shows that the updated regression estimator  $\widehat{\beta}^{(1)}$  only depends on the initial regression estimator  $\widehat{\beta}^{(0)}$  and not on the initial scale estimator  $\widehat{\sigma}^{(0)}$ . 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 cut-off is made in a matching non-symmetric way. In that situation both expansions involve the initial estimation uncertainty for  $\beta$  and  $\sigma^2$ .

We can immediately use Theorem 7.2 for an  $m$ -fold iteration of (7.1), (7.2). Results for infinite iterations follow in §7.5.

### 7.3 Robustified Least Squares

Robustified least squares arises when the initial estimators are the full-sample least squares estimator. We can analyze this 1-step Huber-skip M-estimator using Theorem 7.2. The product moment properties in Assumption 6.1 imply that the initial estimators satisfy

$$N^{-1}(\widetilde{\beta} - \beta) = O_P(1), \quad n^{1/2}(\widetilde{\sigma}^2 - \sigma^2) = n^{-1/2} \sum_{i=1}^n (\varepsilon_i^2 - \sigma^2) + O_P(n^{-1/2}). \quad (7.3)$$

Thus, the conditions of Theorem 7.2 are satisfied so that the robustified least squares estimators can be expanded as in (7.1), (7.2). The asymptotic distribution of estimator for  $\beta$  will depend on the properties of the regressors. For simplicity the regressors are assumed stationary in the following result.

**Theorem 7.3** (Johansen and Nielsen, 2009, Corollary 1.4) Consider the 1-step Huber-skip M-estimator defined with the weights (4.8) and where the initial estimators  $\widetilde{\beta}$ ,  $\widetilde{\sigma}^2$  are the full-sample least squares estimators. Suppose Assumption 6.1 holds and that the regressors are stationary. Then

$$n^{1/2} \begin{pmatrix} \widehat{\beta} - \beta \\ \widehat{\sigma}^2 - \sigma^2 \end{pmatrix} \xrightarrow{D} \mathbf{N} \left\{ 0, \begin{pmatrix} \sigma^2 \eta_\beta \Sigma^{-1} & 0 \\ 0 & 2\sigma^4 \eta_\sigma \end{pmatrix} \right\},$$

where, using the coefficients  $(\tau, \varkappa, \zeta)$  from (3.2) and (3.4), the efficiency factors  $\eta_\beta, \eta_\sigma$  are

$$\psi^2 \eta_\beta = \tau \{1 + 4\text{cf}(c)\} + \{2\text{cf}(c)\}^2, \quad 2\tau^2 \eta_\sigma = (\varkappa - \tau^2/\psi)(1 + \zeta) + \frac{\zeta^2}{4}(\kappa - 1). \quad (7.4)$$

The result generalises the statement (5.3) for the location problem. The efficiency factor  $\eta_\beta$  is plotted as the top curve in Figure 3. A plot of the efficiency for the variance,  $\eta_\sigma$  can be found in Johansen and Nielsen (2009, Figure 1.1). Situations with non-stationary regressors are also discussed in that paper.

## 7.4 Impulse Indicator Saturation

Impulse Indicator Saturation is a second example of a 1-step Huber-skip M-estimator. This requires the choice of sub-sample  $\mathcal{I}_j$ , each with  $n_j$  observations. If the product moment properties of Assumption 6.1 hold for each sub-sample and  $n_j/n \rightarrow \lambda_j > 0$  then the initial estimators satisfy

$$N^{-1}(\widehat{\beta}_j - \beta) = O_{\mathbb{P}}(1), \quad n_j^{1/2}(\widehat{\sigma}^2 - \sigma^2) = n_j^{-1/2} \sum_{i \in \mathcal{I}_j} (\varepsilon_i^2 - \sigma^2) + O_{\mathbb{P}}(n^{-1/2}). \quad (7.5)$$

The asymptotic distribution theory will depend on the choice of sub-samples and regressors. For simplicity we only report the split-half case with subsets  $\mathcal{I}_1 = (i \leq n/2)$  and  $\mathcal{I}_2 = (i > n/2)$  and stationary regressors.

**Theorem 7.4** (*Johansen and Nielsen, 2009, Theorems 1.5, 1.7*) *Consider the split-half Impulse Indicator Saturation estimator of Algorithm 4.2. Suppose Assumption 6.1 holds with stationary regressors. Recall the efficiency factors  $\eta_\beta, \eta_\sigma$  from (7.4). Then the initial estimators satisfy*

$$n^{1/2} \begin{Bmatrix} \widehat{\beta}^{(0)} - \beta \\ (\widehat{\sigma}^{(0)})^2 - \sigma^2 \end{Bmatrix} \xrightarrow{\mathbb{D}} \mathbf{N} \left\{ 0, \begin{pmatrix} \sigma^2 \eta_\beta \Sigma^{-1} & 0 \\ 0 & 2\sigma^4 \eta_\sigma \end{pmatrix} \right\},$$

Moreover, the updated Impulse Indicator Saturation estimator satisfies

$$n^{1/2}(\widehat{\beta}^{(1)} - \beta) \xrightarrow{\mathbb{D}} \mathbf{N}(0, \sigma^2 \eta_\beta^{\text{iis}} \Sigma^{-1}),$$

where

$$\psi^4 \eta_\beta^{\text{iis}} = \{\psi + 2\text{cf}(c)\} \tau [\psi + 2\text{cf}(c) + 2\{2\text{cf}(c)\}^2] + \frac{1}{2} \{2\text{cf}(c)\}^4.$$

The efficiency factors  $\eta_\beta$  and  $\eta_\beta^{\text{iis}}$  for the split-half case are plotted as the top and the middle curve, respectively, in Figure 3. Johansen and Nielsen (2009) also discuss situations with general index sets  $\mathcal{I}_1, \mathcal{I}_2$  and where the regressors are non-stationary.

## 7.5 Iterated 1-step Huber-skip M-estimators

The asymptotic theory of the iterated 1-step Huber-skip M-estimator for regression is given in Johansen and Nielsen (2013). This includes iteration of the robustified least squares estimator and of the Impulse Indicator Saturation estimator with general index sets and general regressors. In each step the asymptotic theory is governed by Theorem 7.2. But what does it take to control the iteration and establish a fixed point result?

We start by showing that the sequence of normalised estimators  $\widehat{\beta}^{(m)}, \widehat{\sigma}^{(m)}$  is tight.

**Theorem 7.5** (*Johansen and Nielsen, 2013, Theorem 3.3*) *Consider the iterated 1-step Huber-skip M-estimator in Algorithm 4.1. Suppose Assumption 6.1 holds and that  $N^{-1}(\widehat{\beta}^{(0)} - \beta)$  and  $n^{1/2}(\widehat{\sigma}^{(0)} - \sigma)$  are  $O_{\mathbb{P}}(1)$ . Then*

$$\sup_{0 \leq m < \infty} |N^{-1}(\widehat{\beta}^{(m)} - \beta)| + |n^{1/2}(\widehat{\sigma}^{(m)} - \sigma)| = O_{\mathbb{P}}(1).$$

Theorem 7.5 is proved by showing that the expansions (7.1), (7.2) are contractions. Necessary conditions are that  $2cf(c)/\psi < 1$  and  $\zeta/(2\tau) < 1$ . This holds for normal or  $t$ -distributed innovations, see Johansen and Nielsen (2013, Theorem 3.6).

In turn, Theorem 7.5 leads to a fixed point result for infinitely iterated estimators.

**Theorem 7.6** (Johansen and Nielsen, 2013, Theorem 3.3) *Consider the iterated 1-step Huber-skip M-estimator in Algorithm 4.1. Suppose Assumption 6.1 holds and that  $N^{-1}(\widehat{\beta}^{(0)} - \beta)$  and  $n^{1/2}(\widehat{\sigma}^{(0)} - \sigma)$  are  $O_{\mathbf{P}}(1)$ . Then, for all  $\epsilon, \eta > 0$  a pair  $m_0, n_0 > 0$  exists so for all  $m > m_0$  and  $n > n_0$  it holds*

$$\mathbf{P}\{|N^{-1}(\widehat{\beta}^{(m)} - \widehat{\beta}^*)| + n^{1/2}|\widehat{\sigma}^{(m)} - \widehat{\sigma}^*| > \eta\} < \epsilon,$$

where

$$N^{-1}(\widehat{\beta}^* - \beta) = \frac{1}{\psi - 2cf(c)} \Sigma_n^{-1} N' \sum_{i=1}^n x_i \varepsilon_i 1_{(|\varepsilon_i| \leq \sigma c)}, \quad (7.6)$$

$$n^{1/2}\{(\widehat{\sigma}^*)^2 - \sigma^2\} = \frac{2}{2\tau - \zeta} n^{-1/2} \sum_{i=1}^n (\varepsilon_i^2 - \sigma^2 \tau / \psi) 1_{(|\varepsilon_i| \leq \sigma c)}. \quad (7.7)$$

Recently Cavaliere and Georgiev (2013) made a similar analysis of a sequence of Huber-skip M-estimators for the parameter of a first order autoregression with infinite variance errors and an autoregressive coefficient of unity.

Iterated 1-step Huber-skip M-estimators can be viewed as iteratively reweighted least squares with binary weights. Dollinger and Staudte (1981) gave conditions for convergence of iteratively reweighted least squares for smooth weights. Their argument was cast in terms of influence functions. While Theorem 7.6 is similar in spirit, the employed tightness argument is different because of the binary weights.

An issue of interest in the literature is, whether a slow initial convergence rate can be improved upon through iteration. This would open up for using robust estimators converging for instance at an  $n^{1/3}$  rate as initial estimator. An example would be the Least Median Squares estimator of Rousseeuw (1984). Such a result would complement the result of He and Portnoy (1992), who find that the convergence rate cannot be improved in a single step of the iteration, as well as Theorem 8.3 below showing that the Forward Search can improve the rate of a slowly converging initial estimator.

## 8 Asymptotic results for Huber-skip L-type estimators

The difference between the Huber-skip estimators of the M-type and the L-type is that the former have a fixed cut-off, whereas the latter have a cut-off determined from the order statistics of the absolute residuals. The asymptotic results appear to be the same, but the argument to get there is a bit more convoluted for the L-type estimators because of the quantiles involved. We give an overview of the results for Least Trimmed Squares estimators, 1-step Huber-skip L-estimators as well as the Forward Search.

## 8.1 Least Trimmed Squares

The Least Trimmed Squares estimator has the same asymptotic expansion as the Huber-skip M-estimator. Vášek (2006a,b,c) proved this for the case of fixed regressors.

**Theorem 8.1** (*Vášek, 2006c, Theorem 1*) *Consider the Least Trimmed Squares estimator  $\widehat{\beta}_{LTS}$  defined as minimizer of (4.6). Suppose Assumption 6.1 holds. Suppose the regressors are fixed and that their empirical distribution can be suitably approximated by a continuous distribution function, see Vášek (2006c) for details. Then*

$$N^{-1}(\widehat{\beta}_{LTS} - \beta) = \frac{1}{\psi - 2\text{cf}(c)} \Sigma_n^{-1} N' \sum_{i=1}^n x_i \varepsilon_i 1_{(|\varepsilon_i| \leq \sigma c)} + o_{\mathbf{P}}(1).$$

## 8.2 1-step Huber-skip L-estimators

The 1-step Huber-skip L-estimator has the following expansion.

**Theorem 8.2** *Consider the 1-step Huber-skip L-estimators  $\widehat{\beta}^{(1)}$ ,  $\widehat{\sigma}^{(1)}$  defined by (4.7), (4.5) with weights (4.9). Suppose Assumption 6.1 holds and that  $N^{-1}(\widehat{\beta}^{(0)} - \beta)$  is  $O_{\mathbf{P}}(1)$ . Then*

$$N^{-1}(\widehat{\beta}^{(1)} - \beta) = \frac{1}{\psi} \Sigma_n^{-1} N' \sum_{i=1}^n x_i \varepsilon_i 1_{(|\varepsilon_i| \leq \sigma c)} + \frac{2\text{cf}(c)}{\psi} N^{-1}(\widehat{\beta}^{(0)} - \beta) + o_{\mathbf{P}}(1). \quad (8.1)$$

$$\begin{aligned} n^{1/2}(\widehat{\sigma}^{(1)} - \sigma) &= \frac{1}{2\tau\sigma} n^{-1/2} \sum_{i=1}^n (\varepsilon_i^2 - \sigma^2 \frac{\tau}{\psi}) 1_{(|\varepsilon_i| \leq \sigma c)} + \frac{\zeta}{2\tau} n^{1/2} \left( \frac{\widehat{\xi}^{(k)}}{c} - \sigma \right) + o_{\mathbf{P}}(1) \\ &= \frac{1}{2\tau\sigma} n^{-1/2} \sum_{i=1}^n (\varepsilon_i^2 - \sigma^2 \frac{\tau}{\psi}) 1_{(|\varepsilon_i| \leq \sigma c)} + \frac{\sigma\zeta}{4c\tau} n^{-1/2} \sum_{i=1}^n \{1_{(|\varepsilon_i| \leq \sigma c)} - \psi\} + o_{\mathbf{P}}(1) \end{aligned} \quad (8.2)$$

$$(8.3)$$

**Proof.** Equations (8.1), (8.2) follow from Theorem 6.1. Equation (8.3) with its expansion of the quantile  $\widehat{\xi}^{(k)}$  follows from Johansen and Nielsen (2014a, Lemma D.11). ■

Ruppert and Carroll (1980) state a similar result for a related 1-step L-estimator, but omit the details of the proof. It is interesting to note that the expansions of the one-step regression estimator of L-type in (8.1) is the same as for the M-type in (7.1). In contrast, the variance estimators have different expansions. In particular, the L-estimator does not use the initial variance estimator and, consequently, the expansion does not involve uncertainty from the initial estimation.

## 8.3 Forward Search

The Forward Search is an iterated 1-step Huber-skip L-estimator, where the cut-off changes slightly in each step. We highlight asymptotic expansions for the forward regression estimators  $\widehat{\beta}^{(m)}$  and for the scaled forward residuals  $\widehat{z}^{(m)}/\widehat{\sigma}^{(m)}$ . The results are formulated in terms of embeddings of the time series  $\widehat{\beta}^{(m)}$ ,  $\widehat{\sigma}^{(m)}$ ,  $\widehat{z}^{(m)}$  for  $m = m_0 + 1, \dots, n$  into the space  $D[0, 1]$  of right continuous functions with limits from the left, for instance,

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

**Theorem 8.3** (Johansen and Nielsen, 2014a, Theorems 3.1, 3.2, 3.5) Consider the Forward Search estimator in Algorithm 4.3. Suppose Assumption 6.1 holds and that  $N^{-1}(\widehat{\beta}^{(m_0)} - \beta)$  is  $O_{\mathbb{P}}(n^{1/4-\eta})$  for some  $\eta > 0$ . Let  $\psi_1 > \psi_0$ . Then, it holds

$$\begin{aligned} \sup_{\psi_1 \leq \psi \leq 1} |N^{-1}(\widehat{\beta}_{\psi} - \beta) - \frac{1}{\psi - 2c_{\psi}f(c_{\psi})} \Sigma_n^{-1} N' \sum_{i=1}^n x_i \varepsilon_i 1_{(|\varepsilon_i/\sigma| \leq c_{\psi})}| &= o_{\mathbb{P}}(1), \\ \sup_{\psi_0 \leq \psi \leq n/(n+1)} |\tau n^{1/2} (\frac{\widehat{\sigma}_{\psi}^2}{\sigma^2} - 1) - n^{-1/2} \sum_{i=1}^n \{(\frac{\varepsilon_i}{\sigma})^2 - \frac{\tau}{\psi}\} 1_{(|\varepsilon_i/\sigma| \leq c_{\psi})} & \\ + (c_{\psi}^2 - \frac{\tau}{\psi}) n^{-1/2} \sum_{i=1}^n (1_{(|\varepsilon_i/\sigma| \leq c_{\psi})} - \psi) &| = o_{\mathbb{P}}(1), \\ \sup_{\psi_0 \leq \psi \leq n/(n+1)} |2f(c_{\psi}) n^{1/2} (\widehat{z}_{\psi}/\sigma - c_{\psi}) + n^{-1/2} \sum_{i=1}^n \{1_{(|\varepsilon_i/\sigma| \leq c_{\psi})} - \psi\} &| = o_{\mathbb{P}}(1). \end{aligned}$$

The asymptotic variances and covariances are given in Theorem A.1.

The proof uses the theory of weighted and marked empirical processes outlined in §6.2 combined with the theory of quantile processes discussed in Csörgö (1983). A single step of the algorithm was previously analyzed in Johansen and Nielsen (2010).

Comparing Theorem 8.3 with Theorems, 7.6, 8.1, we recognise the asymptotic result for the estimator for  $\beta$ . The efficiency relative to the least squares estimator is shown as the bottom curve in Figure 3. The asymptotic expansion for the variance estimator  $\widehat{\sigma}_{\psi}^2$  is, however, different from the expression for the iterated 1-step Huber-skip M-estimator in Theorem 7.6, reflecting the different handling of the scale. The Bahadur (1966) representation linking the empirical distribution of the scaled innovations  $\varepsilon_i/\sigma$  with their order statistics,  $\widehat{c}_{\psi}$  say, implies that  $2f(c_{\psi}) n^{1/2} (\widehat{z}_{\psi}/\sigma - \widehat{c}_{\psi})$  vanishes. Moreover, the minimum deletion residual  $\widehat{d}^{(m)} = \min_{i \notin S^{(m)}} \widehat{\xi}_i^{(m)}$  has the same asymptotic expansion as  $\widehat{z}^{(m)} = \widehat{\xi}_{(m+1)}^{(m)}$  after a burn-in period. See Johansen and Nielsen (2014a, Theorem 3.4) for details.

The idea of the Forward Search is to monitor the plot of the sequence of scaled forward residuals. Combining the expansions for  $\widehat{\sigma}_{\psi}$  and  $\widehat{z}_{\psi}$  in Theorem 8.3 gives the next result.

**Theorem 8.4** (Johansen and Nielsen 2014a, Theorem 3.3). Consider the Forward Search estimator in Algorithm 4.3. Suppose Assumption 6.1 holds and that  $N^{-1}(\widehat{\beta}^{(m_0)} - \beta)$  is  $O_{\mathbb{P}}(n^{1/4-\eta})$  for some  $\eta > 0$ . Let  $\psi_1 > \psi_0$ . Then

$$\sup_{\psi_0 \leq \psi \leq n/(n+1)} |2f(c_{\psi}) n^{1/2} (\frac{\widehat{z}_{\psi}}{\widehat{\sigma}_{\psi}} - c_{\psi}) + \mathbb{Z}_n(c_{\psi})| = o_{\mathbb{P}}(1),$$

where the process  $\mathbb{Z}_n(c)$  given by

$$\left\{1 - \frac{c_{\psi}f(c_{\psi})}{\tau} (c_{\psi}^2 - \frac{\tau}{\psi})\right\} n^{-1/2} \sum_{i=1}^n \{1_{(|\varepsilon_i/\sigma| \leq c)} - \psi\} + \frac{c_{\psi}f(c_{\psi})}{\tau} n^{-1/2} \sum_{i=1}^n \left\{(\frac{\varepsilon_i}{\sigma})^2 - \frac{\tau}{\psi}\right\} 1_{(|\varepsilon_i/\sigma| \leq c)} \quad (8.4)$$

converges to a Gaussian process  $\mathbb{Z}$ . The covariance of  $\mathbb{Z}$  is given in Theorem A.1.

## Part II

# Gauge as a measure of false detection

We now present some new results for the outlier detection algorithms. Outlier detection algorithms will detect outliers with a positive probability when in fact there are no outliers. xxx We analyze this in terms of the gauge, which is the expected frequency of falsely detected outliers when, in fact, the data generating process has no outliers. The idea of a gauge originates in the work of Hoover and Perez (1999) and is formally introduced in Hendry and Santos (2010), see also Castle, Doornik and Hendry (2011).

The gauge concept is related to, but also distinct from the concept of a size of a statistical test, which is the probability of falsely rejecting a true hypothesis. For a statistical test we choose the critical value indirectly from the size we are willing to tolerate. In the same way, for an outlier detection algorithm, we can choose the cut-off for outliers indirectly from the gauge we are willing to tolerate.

The detection algorithms assign binary weights  $\hat{v}_i$  to each observation, so that  $\hat{v}_i = 0$  for outliers and  $\hat{v}_i = 1$  otherwise. We define the empirical or sample gauge as the frequency of falsely detected outliers

$$\hat{\gamma} = \frac{1}{n} \sum_{i=1}^n (1 - \hat{v}_i). \quad (8.5)$$

In turn, the population gauge is the expected frequency of falsely detected outliers, when in fact the model has no contamination, that is

$$\mathbb{E}\hat{\gamma} = \mathbb{E} \frac{1}{n} \sum_{i=1}^n (1 - \hat{v}_i).$$

To see how the gauge of an outlier detection algorithm relates to the size of a statistical test, consider an outlier detection algorithm which classify observations as outliers if the absolute residuals  $|y_i - x_i' \hat{\beta}| / \hat{\sigma}$  is large for some estimator  $(\hat{\beta}, \hat{\sigma})$ . That algorithm has gauge

$$\hat{\gamma} = \frac{1}{n} \sum_{i=1}^n (1 - \hat{v}_i) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}_{(|y_i - x_i' \hat{\beta}| > \hat{\sigma} c)}. \quad (8.6)$$

Suppose the parameters  $\beta, \sigma$  where known so that we could choose  $\hat{\beta}, \hat{\sigma}$  as  $\beta, \sigma$ . Then the population gauge reduces to the size of a test that a single observation is an outlier, that is

$$\mathbb{E} \frac{1}{n} \sum_{i=1}^n \mathbf{1}_{(|y_i - x_i' \beta| > \sigma c)} = \mathbb{E} \frac{1}{n} \sum_{i=1}^n \mathbf{1}_{(|\varepsilon_i| > \sigma c)} = \mathbb{P}(|\varepsilon_1| > \sigma c) = \gamma.$$

In general, the population gauge will, however, be different from the size of such a test because of the estimation error. In §9, §10 we analyze the gauge implicit in the definition of a variety of estimators of type M and L, respectively. Proofs follow in the appendix.

## 9 The gauge of Huber-skip M-estimators

Initially a consistency result is given for the gauge of Huber-skip M-estimators and a distribution theory follows. A normal theory arises when the proportion of falsely detected outliers is controlled by fixing the cut-off  $c$  as  $n$  increases whereas a Poisson exceedence theory arises when  $n\gamma$  is held fixed as  $n$  increases.

## 9.1 Asymptotic analysis of the gauge

We give an asymptotic expansion of the sample gauge of the type (8.6).

**Theorem 9.1** *Consider a sample gauge  $\hat{\gamma}$  of the form (8.6). Suppose Assumption 6.1 holds and that  $N^{-1}(\hat{\beta} - \beta)$ ,  $n^{1/2}(\hat{\sigma}^2 - \sigma^2)$  are  $O_P(1)$ . Then, for fixed  $c$ ,*

$$n^{1/2}(\hat{\gamma} - \gamma) = n^{-1/2} \sum_{i=1}^n \{1_{(|\varepsilon_i| > \sigma c)} - \gamma\} + 2\text{cf}(c)n^{1/2}\left(\frac{\hat{\sigma}}{\sigma} - 1\right) + o_P(1). \quad (9.1)$$

It follows that  $E\hat{\gamma} \rightarrow \gamma$ .

Note that convergence in mean is equivalent to convergence in probability since the gauge takes values in the interval  $[0, 1]$ , see Billingsley (1968, Theorem 5.4).

Theorem 9.1 applies to various Huber-skip M-estimators. For the Huber-skip M-estimator the estimators  $\hat{\beta}, \hat{\sigma}^2$  are the Huber-skip estimator and corresponding variance estimator. For 1-step Huber-skip M-estimator the estimators  $\hat{\beta}, \hat{\sigma}^2$  are the initial estimators. For Impulse Indicator Saturation or 1-step Huber-skip M-estimator iterated  $m$  times the estimators  $\hat{\beta}, \hat{\sigma}^2$  are the estimators from step  $m - 1$ .

## 9.2 Normal approximations to gauge

We control the *proportion* of falsely discovered outliers by fixing the cut-off  $c$ . In that case an asymptotically normal distribution theory follows from the expansion in Theorem 9.1. The asymptotic variance is analyzed case by case since the expansion in Theorem 9.1 depends on the variance estimator  $\hat{\sigma}^2$ .

*Huber-skip M-estimator:* Theorem 7.1 shows that  $N^{-1}(\hat{\beta} - \beta)$  is tight. This is the simplest case to analyse since the variance is assumed known so that  $\hat{\sigma}^2 = \sigma^2$ . Therefore only the first binomial term in Theorem 9.1 matters.

**Theorem 9.2** *Consider the Huber-skip M-estimator  $\hat{\beta}$  defined from (4.2), (4.3) with known  $\sigma$ , fixed  $c$  and sample gauge  $\hat{\gamma} = n^{-1} \sum_{i=1}^n 1_{(|y_i - x_i' \hat{\beta}| > \sigma c)}$ . Suppose Assumption 6.1 holds. Then*

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

*The robustified least squares estimator:* This is the 1-step Huber-skip M-estimator  $\hat{\beta}$  defined in (4.7), (4.8), where the initial estimators  $\tilde{\beta}, \tilde{\sigma}^2$  are the full-sample least squares estimators. The binomial term in Theorem 9.1 is now combined with a term from the initial variance estimator  $\tilde{\sigma}^2$ .

**Theorem 9.3** *Consider the robustified least squares estimator  $\hat{\beta}$  defined from (4.7), (4.8), and the initial estimators  $\tilde{\beta}$  and  $\tilde{\sigma}^2$  are the full sample least squares estimators, fixed  $c$  and sample gauge is  $\tilde{\gamma} = n^{-1} \sum_{i=1}^n 1_{(|y_i - x_i' \tilde{\beta}| > \tilde{\sigma} c)}$ . Suppose Assumption 6.1 holds. Then*

$$n^{1/2}(\tilde{\gamma} - \gamma) \xrightarrow{D} \mathbf{N}[0, \gamma(1 - \gamma) + 2\text{cf}(c)(\tau - \psi) + \{\text{cf}(c)\}^2(\kappa - 1)].$$

The variance in Theorem 9.3 is larger than the binomial variance for a normal reference distribution and any choice of  $\gamma$ . This is seen through differentiation with respect to  $c$ .

The split-half *Impulse Indicator Saturation estimator*: The estimator is defined in Algorithm 4.2. Initially, the outliers are defined using the indicator  $\widehat{v}_i^{(-1)}$  based on the split-sample estimators  $\widehat{\beta}_1, \widehat{\sigma}_1^2$  and  $\widehat{\beta}_2, \widehat{\sigma}_2^2$ , see (4.10). The outliers are reassessed using the updated estimators  $\widehat{\beta}^{(0)}, \widehat{\sigma}^{(0)}$ . Thus, the algorithm gives rise to two sample gauges

$$\widehat{\gamma}^{(-1)} = n^{-1} \sum_{i \in \mathcal{I}_1} \mathbf{1}_{(|y_i - x_i' \widehat{\beta}_2| > \widehat{\sigma}_2 c)} + n^{-1} \sum_{i \in \mathcal{I}_2} \mathbf{1}_{(|y_i - x_i' \widehat{\beta}_1| > \widehat{\sigma}_1 c)}, \quad (9.2)$$

$$\widehat{\gamma}^{(0)} = n^{-1} \sum_{i=1}^n \mathbf{1}_{(|y_i - x_i' \widehat{\beta}^{(0)}| > \widehat{\sigma}^{(0)} c)}. \quad (9.3)$$

For simplicity we only report the result for the initial gauge  $\widehat{\gamma}^{(-1)}$ . The updated gauge  $\widehat{\gamma}^{(0)}$  has a different asymptotic variance.

**Theorem 9.4** *Consider the Impulse Indicator Saturation. Suppose Assumption 6.1 holds for each set  $\mathcal{I}_1, \mathcal{I}_2$ . Then, for fixed  $c$ , the initial sample gauge  $\widehat{\gamma}^{(-1)}$  has the same asymptotic distribution as the sample gauge for robustified least squares reported in Theorem 9.3.*

*The iterated 1-step Huber-skip M-estimator*: The estimator is defined in Algorithm 4.1. Special cases are the iterated robustified least squares estimator and the Impulse Indicator Saturation. If the algorithm is stopped after  $m + 1$  steps the sample gauge is

$$\widehat{\gamma}^{(m)} = n^{-1} \sum_{i=1}^n \mathbf{1}_{(|y_i - x_i' \widehat{\beta}^{(m)}| > \widehat{\sigma}^{(m)} c)} \quad \text{for } m = 0, 1, 2, \dots$$

Because the estimation errors  $N^{-1}(\widehat{\beta}^{(m)} - \beta), n^{1/2}(\widehat{\sigma}^{(m)} - \sigma)$  are tight by Theorem 7.5, the sequence of sample gauges will also be tight. Theorem 9.1 then generalises as follows.

**Theorem 9.5** *Consider the iterated 1-step Huber-skip estimator. Suppose Assumption 6.1 holds and that the initial estimators satisfy that  $N^{-1}(\widehat{\beta}^{(0)} - \beta)$  and  $n^{1/2}(\widehat{\sigma}^{(0)} - \sigma)$  are  $\text{O}_{\mathbb{P}}(1)$ . Then, for fixed  $c$ , the sequence of sample gauges  $\widehat{\gamma}^{(m)}$  satisfies*

$$\sup_{0 \leq m < \infty} |\mathbb{E} \widehat{\gamma}^{(m)} - \gamma| \rightarrow 0, n \rightarrow \infty.$$

A fixed point result can also be derived for the gauge.

**Theorem 9.6** *Consider the iterated 1-step Huber-skip estimator  $\beta^*, \sigma^*$ , see (7.6) and (7.7) and the  $v_i^*, \gamma^*$  defined from these. Suppose Assumption 6.1 holds and that the initial estimators satisfy that  $N^{-1}(\widehat{\beta}^{(0)} - \beta)$  and  $n^{1/2}(\widehat{\sigma}^{(0)} - \sigma)$  are  $\text{O}_{\mathbb{P}}(1)$ . Then, for all  $\epsilon, \eta > 0$  a pair  $n_0, m_0 > 0$  exists so that for all  $n, m$  so  $n \geq n_0$  and  $m \geq m_0$  it holds, for fixed  $c$ ,*

$$\mathbb{P}(n^{1/2} |\widehat{\gamma}^{(m)} - \gamma^*| > \eta) < \epsilon.$$

where, for  $(\tau, \varkappa, \zeta)$  from (3.2) and (3.4),

$$n^{1/2}(\widehat{\gamma}^* - \gamma) = n^{-1/2} \sum_{i=1}^n \{ \mathbf{1}_{(|\varepsilon_i| > \sigma c)} - \gamma \} - \frac{2\text{cf}(c)}{2\tau - \zeta} n^{-1/2} \sum_{i=1}^n \left( \frac{\varepsilon_i^2}{\sigma^2} - \frac{\tau}{\psi} \right) \mathbf{1}_{(|\varepsilon_i| \leq \sigma c)} + \text{o}_{\mathbb{P}}(1).$$

Moreover, the two sums are asymptotically independent and it holds that

$$n^{1/2}(\widehat{\gamma}^* - \gamma) \xrightarrow{\text{D}} \mathbf{N}\left[0, \gamma(1 - \gamma) + \left\{ \frac{2\text{cf}(c)}{2\tau - \zeta} \right\}^2 \left( \varkappa - \frac{\tau}{\psi} \right) \right].$$



$\gamma$	0.05	0.01	0.005	0.0025	0.001
$c$	1.960	2.576	2.807	3.023	3.291
sdv for Huber-skip M	0.218	0.0995	0.0705	0.0499	0.0316
sdv for RLS	0.146	0.0844	0.0634	0.0467	0.0305
sdv for iterated 1-step Huber-skip M	0.314	0.117	0.0783	0.0534	0.0327

Table 1: Asymptotic standard deviations of the empirical gauge.

Table 1 shows the asymptotic variances for the Huber-skip M-estimator, the Robustified Least Squares and for the fully iterated 1-step Huber-skip estimators. The latter include iterated Robustified Least Squares and iterated Impulse Indicator Saturation. The results are taken from Theorems 9.2, 9.3, 9.6, respectively. 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 Hendry and Doornik (2014, §7.6).

### 9.3 Poisson approximation to gauge

If we set the cut-off so as to accept the same fixed *number*  $\lambda$  of falsely discovered outliers regardless of the sample size, then a Poisson exceedence theory arises.

The idea is to choose the cut-off  $c_n$  so that, for some  $\lambda > 0$ ,

$$P(|\varepsilon_i| > \sigma c_n) = \lambda/n. \quad (9.4)$$

The cut-off  $c_n$  appears both in the definition of the gauge and in the definition of the estimators, so some care is needed. We build the argument around the 1-step M-estimator. Let  $\hat{\beta}_n$  and  $\hat{\sigma}_n$  be sequences of estimators that may depend on  $c_n$ , hence the subscript  $n$  in the notation for the estimators. Given these estimators, the sample gauge is

$$\hat{\gamma}_n = n^{-1} \sum_{i=1}^n 1_{(|y_i - x_i' \hat{\beta}_n| > \hat{\sigma}_n c_n)}. \quad (9.5)$$

In the first result we assume that estimation errors  $N^{-1}(\hat{\beta}_n - \beta)$  and  $n^{1/2}(\hat{\sigma}_n - \sigma)$  are tight. Thus, the result immediately applies to robustified least squares, where the initial estimators  $\hat{\beta}_n$  and  $\hat{\sigma}_n$  are the full sample least squares estimators, which do not depend on the cut-off  $c_n$ . But, in general we need to check this tightness condition.

**Theorem 9.7** *Consider the 1-step Huber-skip M-estimator, where  $nP(|\varepsilon_1| \geq \sigma c_n) = \lambda$ . Suppose Assumption 6.1 holds, and that  $N^{-1}(\hat{\beta}_n - \beta)$  and  $n^{1/2}(\hat{\sigma}_n^2 - \sigma^2)$  are  $O_P(1)$ . Then the sample gauge  $\hat{\gamma}_n$  in (9.5) satisfies*

$$n\hat{\gamma}_n \xrightarrow{D} \text{Poisson}(\lambda).$$

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

We next discuss this result for particular initial estimators.

*Robustified least squares estimator:* The initial estimators  $\tilde{\beta}$  and  $\tilde{\sigma}^2$  are the full sample least squares estimators. These do not depend on  $c_n$  so Theorem 9.7 trivially applies.

**Theorem 9.8** Consider the robustified least squares estimator  $\hat{\beta}$  defined from (4.7), (4.8), where the initial estimators  $\tilde{\beta}$  and  $\tilde{\sigma}^2$  are the full sample least squares estimators, while  $c_n$  is defined from (9.4). Suppose Assumption 6.1 holds. Then the sample gauge  $\tilde{\gamma}_n = n^{-1}\sum_{i=1}^n 1_{(|y_i - x'_i \tilde{\beta}| > \tilde{\sigma} c_n)}$  satisfies

$$n\tilde{\gamma}_n \xrightarrow{D} \text{Poisson}(\lambda).$$

*Impulse Indicator Saturation:* Let  $\hat{\beta}_j$  and  $\hat{\sigma}_j^2$  be the split sample least squares estimators. These do not depend on  $c_n$  so Theorem 9.7 trivially applies for the split sample gauge based on

$$\hat{v}_{i,n}^{(-1)} = 1_{(i \in \mathcal{I}_1)} 1_{(|y_i - x'_i \hat{\beta}_2| > \hat{\sigma}_2 c_n)} + 1_{(i \in \mathcal{I}_2)} 1_{(|y_i - x'_i \hat{\beta}_1| > \hat{\sigma}_1 c_n)}.$$

The updated estimators  $\hat{\beta}_n^{(0)}$  and  $(\hat{\sigma}_n^{(0)})^2$  do, however, depend on the cut-off. Thus, an additional argument is needed, when considering the gauge based on the combined initial estimator as in

$$\hat{v}_{i,n}^{(0)} = 1_{(|y_i - x'_i \hat{\beta}_n^{(0)}| > \hat{\sigma}_n^{(0)} c_n)}.$$

**Theorem 9.9** Consider the Impulse Indicator Saturation Algorithm 4.2. Let  $c_n$  be defined from (9.4). Suppose Assumption 6.1 holds for each set  $\mathcal{I}_1, \mathcal{I}_2$ . Let the estimators  $\hat{\beta}_n^{(0)}$  and  $(\hat{\sigma}_n^{(0)})^2$  be defined from (4.7), (4.5) replacing  $v_i$  by  $\hat{v}_{i,n}^{(-1)}$ . Then  $N^{-1}(\hat{\beta}_n^{(0)} - \beta)$  and  $n^{1/2}(\hat{\sigma}_n^{(0)} - \sigma)$  are  $O_{\mathbb{P}}(1)$ , and

$$n\hat{\gamma}_n^{(m)} = \sum_{i=1}^n (1 - \hat{v}_{i,n}^{(m)}) \xrightarrow{D} \text{Poisson}(\lambda) \quad \text{for } m = -1, 0.$$

Table 1 shows the Poisson approximation to the probability of finding at most  $x$  outliers for different values of  $\lambda$ . For small  $\lambda$  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  $\lambda$  decreasing from 5 to 0.1. In particular, for  $\lambda = 1$  and  $n = 100$  so the cut-off 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.

## 10 The gauge of Huber-skip L-type estimators

We now consider the gauge for the L-type estimators. The results and their consequences are somewhat different from the results for M-type estimators. For the Least Trimmed Squares estimator the gauge is trivially  $\hat{\gamma} = \gamma$ , because the purpose of the estimator is to keep the trimming proportion fixed. For the Forward Search the idea is to stop the algorithm once the forward residuals  $\hat{z}^{(m)}/\hat{\sigma}^{(m)}$  become too big. We develop a stopping rule from the gauge.

### 10.1 Gauge for the Forward Search

The forward plot of forward residuals consists of the scaled forward residuals  $\hat{z}^{(m)}/\hat{\sigma}^{(m)}$  for  $m = m_0, \dots, n-1$ . Along with this we plot point-wise confidence bands derived from Theorem 8.4. Suppose we define some stopping time  $\hat{m}$  based on this information, so that  $\hat{m}$  is the number of non-outlying observations while  $n - \hat{m}$  is the number of the outliers. This stopping time can then be calibrated in terms of the sample gauge (8.5), which simplifies as

$$\hat{\gamma} = \frac{n - \hat{m}}{n} = \frac{1}{n} \sum_{m=m_0}^{n-1} (n - m) 1_{(\hat{m}=m)}.$$

Rewrite this by substituting  $n - m = \sum_{j=m}^{n-1} 1$  and change order of summation to get

$$\hat{\gamma} = \frac{1}{n} \sum_{j=m_0}^{n-1} 1_{(\hat{m} \leq j)}. \quad (10.1)$$

If the stopping time is an exit time, then the event  $(\hat{m} \leq j)$  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 8.4 shows that

$$\tilde{\mathbb{Z}}_n(c_\psi) = 2f(c)n^{1/2} \left( \frac{\hat{z}_\psi}{\hat{\sigma}_\psi} - c_\psi \right) = \mathbb{Z}_n(c_\psi) + o_P(1) \quad (10.2)$$

uniformly in  $\psi_0 \leq \psi \leq n/(n+1)$ , where  $\mathbb{Z}_n$  converges to a Gaussian process  $\mathbb{Z}$ . We now choose the stopping time as the first time greater than or equal to  $m_1 (\geq m_0)$ ,  $\hat{z}^{(m)}/\hat{\sigma}^{(m)}$  exceeds some constant level  $q$  times its pointwise asymptotic standard deviation, that is,

$$\hat{m} = \arg \min_{m_1 \leq m < n} [\tilde{\mathbb{Z}}_n(c_{m/n}) > q \text{sdv}\{\tilde{\mathbb{Z}}_n(c_{m/n})\}]. \quad (10.3)$$

To analyze the stopping time (10.3) we consider the event  $(\hat{m} \leq j)$ . This event satisfies

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

Inserting this expression into (10.1) and then using expansion (10.2) we arrive at the following result, with details given in the appendix.

$\gamma$ vs $\psi_1$	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

Table 3: Cut-off values  $q$  for Forward Search as a function of gauge  $\gamma$  and lower point  $\psi_1$  of range for the stopping time.

**Theorem 10.1** *Consider the Forward Search. Suppose Assumption 6.1 holds. Let  $m_0 = \text{int}(\psi_0 n)$  and  $m_1 = \text{int}(\psi_1 n)$  for some  $\psi_1 \geq \psi_0 > 0$ . Consider the stopping time  $\widehat{m}$  in (10.3) for some  $c \geq 0$ . Then*

$$\mathbf{E}\widehat{\gamma} = \mathbf{E}\frac{n - \widehat{m}}{n} \rightarrow \gamma = \int_{\psi_1}^1 \mathbf{P}\left[\sup_{\psi_1 \leq \psi \leq u} \frac{\mathbb{Z}(c_\psi)}{\text{sdv}\{\mathbb{Z}(c_\psi)\}} > q\right] du.$$

If  $\psi_1 > \psi_0$ , the same limit holds for the forward search when replacing  $\widehat{z}^{(m)}$  by the deletion residual  $\widehat{d}^{(m)}$  in the definition of  $\widehat{m}$  in (10.3).

The integral in Theorem 10.1 cannot be computed analytically in an obvious way. Instead we simulated it using Ox 7, see Doornik (2007). For a given  $n$ , draws of normal  $\varepsilon_i$  can be made. From this, the process  $\mathbb{Z}_n$  in (8.4) can be computed. The maximum of  $\mathbb{Z}_n(c_{m/n})/\text{sdv}\{\mathbb{Z}(c_{m/n})\}$  over  $m_1 \leq m \leq j$  can then be computed for any  $m_1 \leq j \leq n$ . Repeating this  $n_{rep}$  times the probability appearing as the integrand can be estimated for a given value of  $q$  and  $u$ . From this the integral  $\gamma$  can be computed. This expresses  $\gamma = \gamma(\psi_1, q)$  as a function of  $q$  and  $\psi_1$ . Inverting this for fixed  $\psi_1$  expresses  $q = q(\psi_1, \gamma)$  as a function of  $\gamma$  and  $\psi_1$ . Results are reported in the Table 3 for  $n_{rep} = 10^5$  and  $n = 1600$ .

## 11 Application to fish data

### 11.1 Impulse Indicator Saturation

The Impulse Indicator Saturation of Algorithm 4.2 is an iterative procedure. Assuming innovations are normal cut-offs can be chosen according to a standard normal distribution. For a finite iteration, where the number of steps is chosen apriori, this follows from Theorem 9.1. For an infinite iteration, this follows from Theorem 9.6. Thus, the cut-off is 2.58 for a 1% gauge. When applying the procedure we split the sample in the first and last half.

The estimated model for the first sample half is

$$\widehat{q}_t^{(1^{st} \text{ half})} = \underset{(1.1)}{6.5} + \underset{(0.12)}{0.26}q_{t-1} - \underset{(0.18)}{0.51}S_t, \quad \widehat{\sigma} = 0.66, \quad t = 2, \dots, 56.$$

The preliminary second half outliers are in observations 95, 108, 68, 75, 94 with residuals  $-4.66$ ,  $-3.11$ ,  $-2.85$ ,  $-2.74$ ,  $-2.66$ . The estimated model for the second sample half is

$$\widehat{q}_t^{(2^{nd} \text{ half})} = \underset{(1.2)}{7.5} + \underset{(0.14)}{0.13}q_{t-1} - \underset{(0.30)}{0.21}S_t, \quad \widehat{\sigma} = 0.77, \quad t = 57, \dots, 111.$$

The preliminary first half outliers are in observations 18, 34 with residuals  $-3.78$ ,  $-2.95$ .

In step  $m = 0$  we estimate a model with dummies for the preliminary outliers and get the full sample model

$$\begin{aligned} \hat{q}_t^{(0)} = & - \underset{\substack{(0.60) \\ [-3.16]}}{1.98} D_t^{18} - \underset{\substack{(0.61) \\ [-2.93]}}{1.80} D_t^{34} - \underset{\substack{(0.60) \\ [-2.10]}}{1.26} D_t^{68} - \underset{\substack{(0.60) \\ [-2.23]}}{1.34} D_t^{75} - \underset{\substack{(0.60) \\ [-2.25]}}{1.35} D_t^{94} - \underset{\substack{(0.61) \\ [-3.96]}}{2.40} D_t^{95} \\ & - \underset{\substack{(0.60) \\ [-2.61]}}{1.56} D_t^{108} + \underset{(0.7)}{7.8} + \underset{(0.08)}{0.11} q_{t-1} - \underset{(0.13)}{0.41} S_t, \quad \hat{\sigma} = 0.60, \end{aligned}$$

with standardised coefficients reported in square brackets. The observations 18, 34, 95, 108 remain outliers. All residuals - for observations without indicators - are now smaller than the cut-off value. Thus we conclude that the observations 18, 34, 95, 108 are outliers.

In step  $m = 1$  we get the Impulse Indicator Saturation model

$$\begin{aligned} \hat{q}_t^{(1)} = & - \underset{\substack{(0.63) \\ [-3.10]}}{1.96} D_t^{18} - \underset{\substack{(0.65) \\ [-2.81]}}{1.82} D_t^{34} - \underset{\substack{(0.64) \\ [-3.76]}}{2.40} D_t^{95} - \underset{\substack{(0.63) \\ [-2.44]}}{1.55} D_t^{108} \\ & + \underset{(0.7)}{7.9} + \underset{(0.08)}{0.09} q_{t-1} - \underset{(0.14)}{0.39} S_t, \quad \hat{\sigma} = 0.63. \end{aligned}$$

The observations 18, 34, 95 remain outliers, while all residuals are small.

In step  $m = 2$  the estimated model is identical to the model (2.2). In that model the observations 18, 34, 95 remain outliers, while all residuals are smaller. Thus, the algorithm has reached a fixed point.

If the gauge is chosen as 0.5% or 0.25% so the cut-off is 2.81 or 3.02, respectively, the algorithm will converge to a solution taking 18, 95 or 95 as outliers, respectively.

## 11.2 Forward Search

We need to choose the initial estimator, the fractions  $\psi_0, \psi_1$  and the gauge. As initial estimator we chose the fast LTS estimator by Rousseeuw and van Driessen (1998) as implemented in the `ltsReg` function of the R-package `robustbase`. We chose to use it with breakdown point  $1 - \psi_0$ . 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 8.1 based on Věček (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 the forward residuals  $\hat{\xi}_{(m+1)}^{(m)}/\varsigma_{m/n}\hat{\sigma}^{(m+1)}$ , where the scaling is chosen in line with Atkinson, Riani and Cerioli (2010). Consider panel (a) where  $\psi_0 = \psi_1 = 0.95$ . 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 §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.

Consider now panel (b) where  $\psi_0 = \psi_1 = 0.80$ . 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

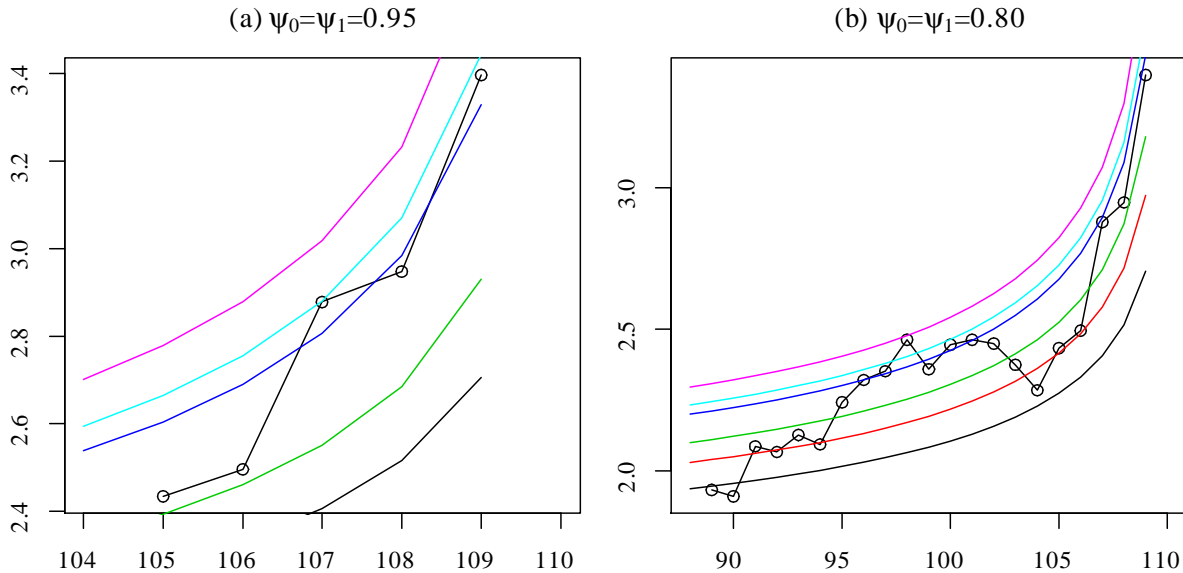


Figure 4: Forward Plots of forward residuals for fish data. Here  $\psi_0 = \psi_1$  is chosen either as 0.95 or 0.80. The bottom curve shows the pointwise median. The top curves show the exit bands for gauges chosen as, from top, 0.001, 0.005, 0.01, 0.05, respectively. Panel (b) also includes an exit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

11 other observations. This leaves some uncertainty about the best choice of the number of outliers. The present analysis is based on asymptotics and could be distorted in finite samples.

## 12 Conclusion and further work

The results presented concern the asymptotic properties of a variety of Huber-skip estimators in the situation where there are no outliers, and the reference distribution is symmetric if not normal. Combined with the concept of the gauge, these results are used for calibrating the cut-off values of the estimators.

In further 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. The probability of finding particular outliers is called potency in Hendry and Santos (2010). It will then be possible to compare the potency of two different outlier detection algorithms, that are calibrated to have the same gauge.

The approach presented is different from the traditional approaches of robust statistics. It would be of interest to compare the approach with the traditional idea of analyzing robust estimators in terms of their breakdown point, see Hampel (1971), or the influence function, see Hampel, Ronchetti, Rousseeuw and Stahel (1986) or Maronna, Martin and Yohai (2006).

First order asymptotic theory is known to be fragile in some situations. A comprehensive simulation study of the results presented would therefore be useful, possibly building on Atkinson and Riani (2006) and Hendry and Doornik (2014).

It would be of interest to extend this research to variable selection algorithms such as

Autometrics, see Hendry and Doornik (2014). The Impulse Indicator Saturation is a stylized version of Autometrics. It should work well, if the researcher can identify a part of the data, that is free from outliers. If this is not the case, one will have to iterate over the choice of sub-samples. In Autometrics potential outliers are coded as dummy variables and the algorithm then searches over these dummy variables along with the other regressors.

## A Proofs

For the asymptotic normality results for the gauge some covariance matrices have to be computed. The results are collected in

**Theorem A.1** *Suppose Assumption 6.1 holds and that  $c = \mathbf{G}(\psi)$ . Then the processes*

$$\begin{aligned}\mathbb{A}_n(c) &= n^{-1/2} \sum_{i=1}^n \{1_{(|\varepsilon_i| > \sigma c)} - \gamma\}, \\ \mathbb{B}_n(c) &= n^{-1/2} \sum_{i=1}^n \left( \frac{\varepsilon_i^2}{\sigma^2} - \frac{\tau}{\psi} \right) 1_{(|\varepsilon_i| \leq \sigma c)}, \\ \mathbb{C}_n(c) &= n^{-1/2} \sum_{i=1}^n \left( \frac{\varepsilon_i^2}{\sigma^2} - 1 \right), \\ \mathbb{K}_n(c) &= N' \sum_{i=1}^n x_i \varepsilon_i 1_{(|\varepsilon_i| \leq \sigma c)}\end{aligned}$$

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

$$\Phi = \text{Var} \begin{Bmatrix} \mathbb{A}_n(c) \\ \mathbb{B}_n(c) \\ \mathbb{C}_n(c) \end{Bmatrix} = \begin{Bmatrix} \gamma(1-\gamma) & 0 & \tau - \psi \\ 0 & \varkappa - \tau^2/\psi & \varkappa - \tau^2/\psi \\ \tau - \psi & \varkappa - \tau^2/\psi & (\varkappa - 1)/2 \end{Bmatrix}. \quad (\text{A.1})$$

*If the regressors are stationary, then  $\mathbb{K}$  is Gaussian independent of  $\mathbb{A}, \mathbb{B}, \mathbb{C}$  with variance  $\tau\sigma^2\Sigma$ .*

*It follows that the asymptotic variance in Theorem 8.3 is given by*

$$\text{asVar} \begin{Bmatrix} 2\mathbf{f}(c_\psi)n^{1/2}(\hat{z}_\psi/\sigma - c_\psi) \\ n^{1/2}(\hat{\sigma}_\psi^2/\sigma^2 - 1) \end{Bmatrix} = \text{Var} \begin{Bmatrix} \mathbb{A}_n(c_\psi) \\ \mathbb{B}_n(c_\psi)/\tau + (c_\psi^2/\tau - 1/\psi)\mathbb{A}_n(c_\psi) \end{Bmatrix} = \omega_1' \Phi \omega_1$$

for

$$\omega_1' = \begin{pmatrix} 1 & 0 & 0 \\ (c_\psi^2/\tau - 1/\psi) & 1/\tau & 0 \end{pmatrix}.$$

*The asymptotic variance in Theorem 8.4 is given by*

$$\begin{aligned}\text{asVar}(2\mathbf{f}(c_\psi)n^{1/2}(\hat{z}_\psi/\hat{\sigma}_\psi - c_\psi)) \\ = \text{Var}(-\{1 - \frac{c_\psi \mathbf{f}(c_\psi)}{\tau}(c_\psi^2 - \frac{\tau}{\psi})\}\mathbb{A}_n(c_\psi) + \frac{c_\psi \mathbf{f}(c_\psi)}{\tau}\mathbb{B}_n(c_\psi)) = \omega_2' \Phi \omega_2\end{aligned}$$

where

$$\omega_2' = \left( -\{1 - \frac{c_\psi \mathbf{f}(c_\psi)}{\tau}(c_\psi^2 - \frac{\tau}{\psi})\}, \frac{c_\psi \mathbf{f}(c_\psi)}{\tau}, 0 \right).$$

**Proof of Theorem 9.1.** Apply the asymptotic expansion in Theorem 6.1. ■

**Proof of Theorem 9.2.** Insert  $\hat{\sigma}^2 = \sigma^2$  in the expansion in Theorem 9.1 and apply Theorem A.1 to the binomial term. ■

**Proof of Theorem 9.3.** The initial estimators satisfy  $N^{-1}(\tilde{\beta} - \beta) = O_{\mathbb{P}}(1)$  and  $n^{1/2}(\tilde{\sigma}^2 - \sigma^2) = n^{-1/2} \sum_{i=1}^n (\varepsilon_i^2 - \sigma^2) + o_{\mathbb{P}}(1)$ , see (7.3). Use Theorem 9.1 to get

$$n^{1/2}(\hat{\gamma} - \gamma) = n^{-1/2} \sum_{i=1}^n \{1_{(|\varepsilon_i| > \sigma c)} - \gamma\} + \text{cf}(c) n^{-1/2} \sum_{i=1}^n (\varepsilon_i^2 - \sigma^2) + o_{\mathbb{P}}(1), \quad (\text{A.2})$$

and apply Theorem A.1. ■

**Proof of Theorem 9.4.** The initial estimators satisfy  $N^{-1}(\hat{\beta}_j - \beta) = O_{\mathbb{P}}(1)$  and  $n^{1/2}(\hat{\sigma}_j^2 - \sigma^2) = n^{-1/2} \sum_{i=1}^n (\varepsilon_i^2 - \sigma^2) + o_{\mathbb{P}}(1)$ , see (7.5). Insert this in the expansion in Theorem 9.1 to get

$$n_j^{-1/2} \sum_{i \in \mathcal{I}_j} \{1_{(|y_i - x_i' \hat{\beta}_{3-j}| > \hat{\sigma}_{3-j} c)} - \gamma\} = n_j^{-1/2} \sum_{i \in \mathcal{I}_j} \{1_{(|\varepsilon_i| > \sigma c)} - \gamma\} + \text{cf}(c) n_j^{-1/2} \sum_{i \in \mathcal{I}_j} (\varepsilon_i^2 - \sigma^2) + o_{\mathbb{P}}(1).$$

Combine the counts of outliers for the two sub-samples to get

$$\begin{aligned} n^{1/2}(\hat{\gamma}^{(-1)} - \gamma) &= n^{-1/2} \sum_{j=1}^2 \frac{n_j^{1/2}}{n_j^{1/2}} \sum_{i \in \mathcal{I}_j} \{1_{(|y_i - x_i' \hat{\beta}_{3-j}| > \hat{\sigma}_{3-j} c)} - \gamma\} \\ &= n^{-1/2} \sum_{j=1}^2 \sum_{i \in \mathcal{I}_j} \{1_{(|\varepsilon_i| > \sigma c)} - \gamma\} + \text{cf}(c) n^{-1/2} \sum_{j=1}^2 \sum_{i \in \mathcal{I}_j} (\varepsilon_i^2 - \sigma^2) + o_{\mathbb{P}}(1). \end{aligned}$$

This reduces to the expansion (A.2) for the robustified least squares estimator. ■

**Proof of Theorem 9.5.** Theorem 7.5 shows that the normalised estimators are tight. Thus, for all  $\epsilon$  there exists an  $A > 0$  so that the set

$$\mathcal{A}_n = \bigcap_{m=0}^{\infty} \{|N^{-1}(\hat{\beta}^{(m)} - \beta)| + |n^{1/2}\{(\hat{\sigma}^{(m)})^2 - \sigma^2\}| \leq U\} \quad (\text{A.3})$$

has probability of at least  $1 - \epsilon$ . Theorem 6.1 then shows that on that set

$$\hat{\gamma}^{(m)} - \gamma = \frac{1}{n} \sum_{i=1}^n \{1_{(|\varepsilon_i| > \sigma c)} - \gamma\} - n^{-1/2} 2\text{cf}(c) n^{1/2} (\hat{\sigma}^{(m)}/\sigma - 1) + o_{\mathbb{P}}(n^{-1/2}), \quad (\text{A.4})$$

where, uniformly in  $m$ , the first term and the second term are  $O_{\mathbb{P}}(n^{-1/2})$ , while the remainder term is  $o_{\mathbb{P}}(n^{-1/2})$ . Therefore

$$\sup_{0 \leq m < \infty} |\hat{\gamma}^{(m)} - \gamma| = o_{\mathbb{P}}(1).$$

Since  $\hat{\gamma}^{(m)}$  and  $\gamma$  are bounded by one then  $\mathbb{E} \sup_{0 \leq m < \infty} |\hat{\gamma}^{(m)} - \gamma|$  vanishes as  $n \rightarrow \infty$ . Thus, by the triangle inequality

$$\sup_{0 \leq m < \infty} |\mathbb{E} \hat{\gamma}^{(m)} - \gamma| \leq \sup_{0 \leq m < \infty} \mathbb{E} |\hat{\gamma}^{(m)} - \gamma| \leq \mathbb{E} \sup_{0 \leq m < \infty} |\hat{\gamma}^{(m)} - \gamma| = o(1).$$

■



**Proof of Theorem 9.6.** On the set  $\mathcal{A}_n$  defined in the proof of Theorem 9.5, see (A.3), we consider the expansion (A.4), that is,

$$n^{1/2}(\widehat{\gamma}^{(m)} - \gamma) = n^{-1/2} \sum_{i=1}^n \{1_{(|\varepsilon_i| > \sigma c)} - \gamma\} - 2cf(c)n^{1/2}(\widehat{\sigma}^{(m)}/\sigma - 1) + o_{\mathbf{P}}(1),$$

where the remainder is uniform in  $m$ . Theorem 7.6 shows that for large  $m, n$  we have

$$n^{1/2}(\widehat{\sigma}^{(m)}/\sigma - 1) = \frac{1}{2\tau - \zeta} n^{-1/2} \sum_{i=1}^n \left( \frac{\varepsilon_i^2}{\sigma^2} - \frac{\tau}{\psi} \right) 1_{(|\varepsilon_i| \leq \sigma c)} + o_{\mathbf{P}}(1),$$

where the remainder is uniform in  $m$ . Combine to get the desired expansion. The asymptotic normality follows from Lemma A.1. ■

Theorem 9.7 is a special case of the following Lemma subjected to Remarks A.1 and A.2 below, because Assumption 6.1 assumes Gaussian errors.

**Lemma A.2** *Suppose Assumption 6.1(ii, d) holds. Let the cut-off  $c_n$  be given by (9.4) and assume that*

(i) *the density  $f$  is symmetric with decreasing tails and support on  $\mathbb{R}$  so that  $c_n \rightarrow \infty$  with*

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

(ii)  $N^{-1}(\widehat{\beta} - \beta), n^{1/2}(\widehat{\sigma}^2 - \sigma^2)$  are  $O_{\mathbf{P}}(1)$ .

*Then the sample gauge  $\widehat{\gamma}$  in (8.6) satisfies*

$$n\widehat{\gamma} \xrightarrow{D} \text{Poisson}(\lambda).$$

**Remark A.1** *Assumption (ia) implies that  $c_n = O(n^{1/r})$  where  $1/r < 1/4$ . Combine the definition  $\mathbf{P}(|\varepsilon_i| > \sigma c_n) = \lambda/n$  with the Markov inequality  $\mathbf{P}(|\varepsilon_i| > \sigma c_n) \leq (\sigma c_n)^{-r} \mathbf{E}|\varepsilon_i|^r$  so that  $c_n \leq \sigma^{-1}(\mathbf{E}|\varepsilon_i|^r)^{1/r} \lambda^{-1/r} n^{1/r} = O(n^{1/r})$ .*

**Remark A.2** *Assumption (i) of Lemma A.2 holds if  $f = \varphi$  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  $2 \log\{f(c_n - n^{-1/4}A)/f(c_n)\} = c_n^2 - (c_n - n^{-1/4}A)^2 = 2c_n n^{-1/4}A - n^{-1/2}A^2$  and use Remark A.1.*

**Proof of Lemma A.2.** 1. *A bound on the sample space.* Since  $N^{-1}(\widehat{\beta} - \beta)$  and  $n^{1/2}(\widehat{\sigma}^2 - \sigma^2)$  are  $O_{\mathbf{P}}(1)$  and in light of Assumption 6.1 (ii, d) then for all  $\epsilon > 0$  exists a constant  $A_0 > 1$  such that the set

$$\mathcal{B}_n = \{|N^{-1}(\widehat{\beta} - \beta)| + n^{1/2}|\widehat{\sigma} - \sigma| + n^{1/4} \max_{1 \leq i \leq n} |N'x_i| \leq A_0\}$$

has probability larger than  $1 - \epsilon$ . It suffices to prove the theorem on this set.

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

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

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

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

It therefore holds that

$$\mathbf{1}_{(\varepsilon_i/\sigma > c_n + n^{-1/4} A_1)} \leq \mathbf{1}_{(y_i - x_i' \hat{\beta} > \hat{\sigma} c_n)} = \mathbf{1}_{(\varepsilon_i > s_i)} \leq \mathbf{1}_{(\varepsilon_i/\sigma > c_n - n^{-1/4} A_1)}.$$

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

$$\mathbf{1}_{(|\varepsilon_i/\sigma| > c_n + n^{-1/4} A_1)} \leq \mathbf{1}_{(|y_i - x_i' \hat{\beta}| > \hat{\sigma} c_n)} \leq \mathbf{1}_{(|\varepsilon_i/\sigma| > c_n - n^{-1/4} A_1)}. \quad (\text{A.5})$$

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

$$n\mathbf{E}\mathbf{1}_{(|\varepsilon_i/\sigma| > c_n + n^{-1/4} A_1)} \rightarrow \lambda, \quad n\mathbf{E}\mathbf{1}_{(|\varepsilon_i/\sigma| > c_n - n^{-1/4} A_1)} \rightarrow \lambda. \quad (\text{A.6})$$

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

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

A first order Taylor expansion and the identity  $2\{1 - \mathbf{F}(c_n)\} = \lambda/n$  give

$$\mathcal{E}_n = n \int_{c_n - n^{-1/4} A_1}^{c_n + n^{-1/4} A_1} 2\mathbf{f}(x) dx = 4n^{-1/4} A_1 \mathbf{f}(c^*) = \frac{4\lambda n^{-1/4} A_1 \mathbf{f}(c^*)}{2\{1 - \mathbf{F}(c_n)\}},$$

for  $|c^* - c_n| \leq n^{-1/4} A_1$ . Rewrite as

$$\mathcal{E}_n = 2\lambda n^{-1/4} A_1 \left\{ \frac{\mathbf{f}(c^*)}{\mathbf{f}(c_n - n^{-1/4} A_1)} \right\} \left\{ \frac{\mathbf{f}(c_n - n^{-1/4} A_1)}{\mathbf{f}(c_n)} \right\} \left[ \frac{\mathbf{f}(c_n)}{c_n \{1 - \mathbf{F}(c_n)\}} \right] c_n.$$

The first fraction is bounded by one since  $\mathbf{f}$  has decreasing tails. Then second and the third fractions are bounded by Assumption (ib, ic). Then use that  $n^{-1/4} c_n = o(1)$  by Remark A.1.

4. *Poisson distribution.* Using the bounds in item 3, it holds on the set  $\mathcal{B}_n$  that

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

Using (A.6) the Poisson limit theorem shows that the upper and lower bounds have Poisson limits with mean  $\lambda$ . ■

**Proof of Theorem 9.9.** 1. *Comparison with least squares.* The estimator  $N^{-1}(\hat{\beta}_n^{(0)} - \beta)$  is based on

$$N' \sum_{i=1}^n \hat{v}_{i,n}^{(-1)} x_i x_i' N = N' \sum_{i=1}^n x_i x_i' N - N' \sum_{i=1}^n (1 - \hat{v}_{i,n}^{(-1)}) x_i x_i' N, \quad (\text{A.7})$$

$$N' \sum_{i=1}^n \hat{v}_{i,n}^{(-1)} x_i \varepsilon_i = N' \sum_{i=1}^n x_i \varepsilon_i - N' \sum_{i=1}^n (1 - \hat{v}_{i,n}^{(-1)}) x_i \varepsilon_i, \quad (\text{A.8})$$

In each equation the first term is the full sample product moment, which converges due to Assumption 6.1, and the estimation error of the full sample least squares is bounded in

probability. It suffices to show that the second terms vanish in probability. The argument for  $n^{1/2}\{(\widehat{\sigma}_n^{(0)})^2 - \sigma^2\}$  is similar.

2. *Tightness of the initial estimators.* Because  $N^{-1}(\widehat{\beta}_j - \beta)$  and  $n^{1/2}(\widehat{\sigma}_j^2 - \sigma^2)$  are  $O_{\mathbf{P}}(1)$ , then for all  $\epsilon > 0$  there exists a constant  $A_0 > 1$  such that the set

$$\mathcal{B}_n = \left\{ \sum_{j=1}^2 |N^{-1}(\widehat{\beta}_j - \beta)| + \sum_{j=1}^2 n^{1/2} |\widehat{\sigma}_j - \sigma| + n^{1/2} \max_{1 \leq i \leq n} |N'x_i| \leq A_0 \right\}$$

has probability larger than  $1 - \epsilon$ . It suffices to prove the theorem on this set.

3. *Bounding the second terms.* The second terms of (A.7) and (A.8) are bounded by

$$\mathcal{S}_p = \sum_{i=1}^n (1 - \widehat{v}_{i,n}^{(-1)}) |N'x_i|^{2-p} |\varepsilon_i|^p \quad \text{for } p = 0, 1.$$

On the set  $\mathcal{B}_n$  we get the further bound, see (A.5) in the proof of Lemma A.2,

$$\mathcal{S}_p \mathbf{1}_{\mathcal{B}_n} \leq \sum_{i=1}^n |N'x_i|^{2-p} |\varepsilon_i|^p \mathbf{1}_{(|\varepsilon_i/\sigma| > c_n - n^{-1/4} A_1)}.$$

The expectation is bounded as

$$\mathbf{E}(\mathcal{S}_p \mathbf{1}_{\mathcal{B}_n}) \leq \mathbf{E} \sum_{i=1}^n |N'x_i|^{2-p} |\varepsilon_i|^p \mathbf{1}_{(|\varepsilon_i/\sigma| > c_n - n^{-1/4} A_1)}.$$

Now

$$|\varepsilon_i|^p \mathbf{1}_{(|\varepsilon_i/\sigma| > c_n - n^{-1/4} A_1)} \leq \mathbf{E}^{1/2} \mathbf{1}_{(|\varepsilon_i/\sigma| > c_n - n^{-1/4} A_1)} \mathbf{E}^{1/2} |\varepsilon_i|^{2p} \mathbf{1}_{(|\varepsilon_i/\sigma| > c_n - n^{-1/4} A_1)}.$$

The first factor is of order  $n^{-1/2}$ , because  $n(1 - F(c_n - n^{-1/4} A_1)) \rightarrow \lambda$ , and the second factor tends to zero because  $E\varepsilon_i^2 < \infty$ . We also have

$$E \sum_{i=1}^n |N'x_i|^{2-p} = n^{1/2(p-2)} E \sum_{i=1}^n |n^{1/2} N'x_i|^{2-p} \leq Cn^{p/2} \leq Cn^{1/2}$$

by Assumption 6.1(ii, d). Collecting these evaluations we find  $\mathcal{S}_p \xrightarrow{\mathbf{P}} 0$ . ■

**Proof of Theorem 10.1.** Theorem 8.3 implies, that  $\mathbb{Z}_n$  converges to a Gaussian process  $\mathbb{Z}$  on  $D[\psi_0, 1]$  endowed with the uniform metric. The variance of  $\mathbb{Z}(c_\psi)$  vanishes for  $\psi \rightarrow 1$  so a truncation argument is needed to deal with the ratio  $\mathbb{X}_n(c_\psi) = \mathbb{Z}_n(c_\psi) / \text{sdv}\{\mathbb{Z}(c_\psi)\}$ .

Approximate the sample gauge by

$$\widehat{\gamma}_v = \frac{n - \widehat{m}}{n} \mathbf{1}_{(\widehat{m} \leq vn)} = \frac{1}{n} \sum_{j=m_1}^{\text{int}(nv)-1} \mathbf{1}_{(\widehat{m} \leq j)},$$

for some  $v < 1$  and using (10.1). Then the sample gauge is  $\widehat{\gamma} = \widehat{\gamma}_1$ , and

$$0 \leq \widehat{\gamma} - \widehat{\gamma}_v = \frac{n - \widehat{m}}{n} \mathbf{1}_{(\widehat{m} > vn)} < \frac{n - nv}{n} = 1 - v. \quad (\text{A.9})$$

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  $\sup_{\psi_1 \leq \psi \leq u} \mathbb{X}_n(c_\psi)$  converges as a process in  $u$  on  $D[\psi_1, v]$ . In turn, for a given  $q$ , the deterministic function  $\mathbf{P}(\widehat{m} \leq nu) = \mathbf{P}\{\sup_{\psi_1 \leq \psi \leq u} \mathbb{X}_n(c_\psi) > q\}$  in  $\psi_1 \leq u \leq v$  converges to a continuous increasing function  $\mathbf{p}(u)$  on  $[\psi_1, v]$ , which is bounded by unity. In particular it holds that

$$\mathbf{E} \widehat{\gamma}_v = \frac{1}{n} \sum_{j=m_1}^{\text{int}(nv)-1} \mathbf{E} \mathbf{1}_{(\widehat{m} \leq j)} = \frac{1}{n} \sum_{j=m_1}^{\text{int}(nv)-1} \mathbf{P}(\widehat{m} \leq j) \rightarrow \gamma_v = \int_{\psi_1}^v \mathbf{p}(u) du \leq v - \psi_1 \leq 1 - \psi_1,$$

and

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

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

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

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

for some fixed  $v$ . Then

$$|\widehat{\gamma} - \gamma| \leq 1 - v + |\widehat{\gamma}_v - \gamma_v| + 1 - v.$$

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

## References

- Atkinson, A.C. and Riani, M. (2000) *Robust Diagnostic Regression Analysis*. New York: Springer.
- Atkinson, A.C. and Riani, M. (2006) Distribution theory and simulations for tests of outliers in regression. *Journal of Computational and Graphical Statistics* 15, 460–476.
- Atkinson, A.C., Riani, M. and Cerioli, A. (2010) The forward search: Theory and data analysis (with discussion). *Journal of the Korean Statistical Society* 39, 117–134.
- Bahadur, R.R. (1966) A note on quantiles in large samples. *Annals of Mathematical Statistics* 37, 577–580.
- Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B* 57, 289–300.
- Bercu, B. and Touati, A. (2008) Exponential inequalities for self-normalized martingales with applications. *Annals of Applied Probability* 18, 1848–1869.
- Bickel, P.J. (1975) One-step Huber estimates in the linear model. *Journal of the American Statistical Association* 70, 428–434.
- Billingsley, P. (1968) *Convergence of Probability Measures*. New York: Wiley.
- Castle, J.L., Doornik, J.A. and Hendry, D.F. (2011) Evaluating automatic model selection. *Journal of Time Series Econometrics* 3, Issue 1, Article 8.
- Cavaliere, G. and Georgiev, I. (2013) Exploiting infinite variance through dummy variables in nonstationary autoregressions. *Econometric Theory* 29, 1162–1195.
- Chen, X.R. and Wu, Y.H. (1988) Strong consistency of M-estimates in linear models. *Journal of Multivariate Analysis* 27, 116–130.

- 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, *The Annals of Statistics* 18, 1651–1675.
- Dollinger, M.B. and Staudte, R.G. (1991) Influence functions of iteratively reweighted least squares estimators. *Journal of the American Statistical Association* 86, 709–716.
- Doornik, J.A. (2007) *Object-Oriented Matrix Programming Using Ox*, 3rd ed. London: Timberlake Consultants Press and Oxford: www.doornik.com.
- Doornik, J.A. (2009) Autometrics. In Castle, J.L. and Shephard, N. (eds.) *The Methodology and Practice of Econometrics: A Festschrift in Honour of David F. Hendry*, pp. 88–121. Oxford: Oxford University Press.
- Doornik, J.A. and Hendry, D.F. (2013) *Empirical Econometric Modelling - PcGive 14, volume 1*. London: Timberlake Consultants.
- Engle, R.F. (1982) Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica* 50, 987–1108
- Engler, E. and Nielsen, B. (2009) The empirical process of autoregressive residuals. *Econometrics Journal* 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 Journal of Economics* 26, 75–92.
- Graddy, K. (2006) The Fulton Fish Market. *Journal of Economic Perspectives* 20, 207–220.
- Hadi, A.S. (1992) Identifying multiple outliers in multivariate data. *Journal of the Royal Statistical Society B* 54, 761–771.
- Hadi, A.S. and Simonoff, J.S. (1993) Procedures for the Identification of Multiple Outliers in Linear Models *Journal of the American Statistical Association* 88, 1264–1272.
- Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J. and Stahel, W.A. (1986) *Robust Statistics: The Approach Based on Influence Functions*. New York: John Wiley & Sons.
- Hampel, F.R. (1971) A general qualitative definition of robustness. *Annals of Mathematical Statistics* 42, 1887–1896.
- He, X. and Portnoy, S. (1992) Reweighted LS estimators converge at the same rate as the initial estimator. *Annals of Statistics* 20, 2161–2167.

- Hendry, D.F. (1999) An econometric analysis of US food expenditure, 1931–1989. In Magnus, J.R. and Morgan, M.S. *Methodology & Tacit Knowledge: Two Experiments in Econometrics*. Chichester: John Wiley & Sons, p. 341–362.
- Hendry, D.F. and Doornik, J.A. (2014) *Empirical Model Discovery and Theory Evaluation*. Cambridge MA: MIT Press.
- Hendry, D.F., Johansen, S. and Santos, C. (2008) Automatic selection of indicators in a fully saturated regression. *Computational Statistics* 23, 337–339.
- Hendry, D.F. and Krolzig, H.-M. (2005) The properties of automatic *GETS* modelling. *Economic Journal* 115, C32–61.
- Hendry, D.F. and Mizon, G.E. (2011) Econometric modelling of time series with outlying observations. *Journal of Time Series Econometrics* 3:1:6.
- Hendry, D.F. and Nielsen, B. (2007) *Econometric Modelling*. Princeton NJ: Princeton University Press.
- Hendry, D.F. and Santos, C. (2010) An automatic test of super exogeneity. In Bollerslev, T., Russell, J.R. and Watson, M.W. (eds.) *Volatility and Time Series Econometrics: Essays in Honor of Robert F. Engle*, pp. 164–193. Oxford: Oxford University Press.
- Hoover, K.D. and Perez, S.J. (1999) Data mining reconsidered: encompassing and the general-to-specific approach to specification search (with discussion). *Econometrics Journal* 2, 167–191.
- Huber, P.J. (1964) Robust estimation of a location parameter. *Annals of Mathematical Statistics* 35, 73–101.
- Huber, P.J. and Ronchetti, E.M. (2009) *Robust Statistics*. New York: Wiley.
- Jaeckel, L.A. (1971) Robust estimates of location: Symmetry and asymmetric contamination. *Annals of Mathematical Statistics* 42, 1020–1034.
- Johansen, S. and Nielsen, B. (2009) An analysis of the indicator saturation estimator as a robust regression estimator. In Castle, J.L. and Shephard, N. (eds.) *The Methodology and Practice of Econometrics: A Festschrift in Honour of David F. Hendry*, pp. 1–36. Oxford: Oxford University Press.
- Johansen, S. and Nielsen, B. (2010) Discussion: The forward search: Theory and data analysis. *Journal of the Korean Statistical Society* 39, 137–145.
- Johansen, S. and Nielsen, B. (2013) Asymptotic theory for iterated one-step Huber-skip estimators. *Econometrics* 1, 53–70.
- Johansen, S. and Nielsen, B. (2014a) Analysis of the Forward Search using some new results for martingales and empirical processes. Updated version of 2013 Discussion Paper with title Asymptotic theory of the Forward Search.

- Johansen, S. and Nielsen, B. (2014b) Asymptotic theory of M-estimators for multiple regression. Work in Progress.
- Jurečková, J. and Sen, P.K. (1996) *Robust Statistical Procedures: Asymptotics and Interrelations*. New York: John Wiley & Sons.
- Jurečková, J., Sen, P.K. and Picek, J. (2012) *Methodological Tools in Robust and Nonparametric Statistics*. London: Chapman & Hall/CRC Press.
- Kilian, L. and Demiroglu, U. (2000) Residual based tests for normality in autoregressions: asymptotic theory and simulations. *Journal of Economic Business and Control* 18, 40–50
- Koul, H.L. (2002) *Weighted Empirical Processes in Dynamic Nonlinear Models*. 2nd edition. New York: Springer.
- Koul, H.L. and Ossiander, M. (1994) Weak convergence of randomly weighted dependent residual empiricals with applications to autoregression. *Annals of Statistics*, **22**, 540–582.
- Liese, F. and Vajda, I. (1994) Consistency of M-estimates in general regression models. *Journal of Multivariate Analysis* 50, 93–110.
- Maronna, R.A., Martin, R.D. and Yohai, V.J. (2006) *Robust Statistics: Theory and Methods*. Chichester: John Wiley & Sons.
- Nielsen, B. (2006) Order determination in general vector autoregressions. In Ho, H.-C., Ing, C.-K., and Lai, T.L. (eds): *Time Series and Related Topics: In Memory of Ching-Zong Wei*. IMS Lecture Notes and Monograph Series 52, 93-112.
- 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. *Journal of the Royal Statistical Society B* 31, 350–371.
- Riani, M., Atkinson, A.C. and Cerioli, A. (2009) Finding an unknown number of multivariate outliers. *Journal of the Royal Statistical Society B*, **71**, 447–466.
- Rousseeuw, P.J. (1984) Least median of squares regression. *Journal of the American Statistical Association*, 79, 871–880.
- Rousseeuw, P.J. and van Driessen, K. (1998) A fast algorithm for the minimum covariance determinant estimator. *Technometrics* 41, 212–223.
- Rousseeuw, P.J. and Leroy, A.M. (1987) *Robust Regression and Outlier Detection*. New York: Wiley.
- Ruppert, D. and Carroll, R.J. (1980) Trimmed least squares estimation in the linear model. *Journal of the American Statistical Association*, 75, 828–838.

- Sampford, M.R. (1953) Some inequalities on Mill's ratio and related functions. *Annals of Mathematical Statistics*, 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. and Ronchetti, E. (2002) A journey in single steps: robust one-step M-estimation in linear regression *Journal of Statistical Planning and Inference*, 103, 287–310.
- White, H. (1980) A heteroskedastic-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica* 48, 817–838.