Testing the Presence of Outliers to Assess Misspecification in Regression Models

Xiyu Jiao[1] and Felix Pretis[1,2*]

[1]Department of Economics, University of Oxford

[2]Programme for Economic Modelling, INET at the Oxford Martin School

This version: November 10, 2017

Abstract

The presence of outlying observations in a regression model can be indicative of model misspecification, consequently, it is important to check for possible outlier contamination. However, algorithms used to detect outliers have a positive probability of finding outliers even when, in fact, the data generation process has no outliers. Deriving distributional results on the expected retention rate of falsely discovered outliers, we propose two set of tests for the overall presence of outliers: first, tests on whether the observed proportion and number of detected outliers deviate from their expected values. Second, 'scaling' tests on whether the number of detected outliers decreases proportionally with the level of significance used to detect outliers. We derive the asymptotic distribution of the tests based on iterated 1-step Huber-skip M-estimators. The first set of tests has power against the number of outliers present, while the second set of tests has power against both outlier magnitude and number. In applications of the tests we consider a cross-sectional macroeconomic model of economic growth, and re-visit a set of previous studies using indicator saturation. The tests are valid for stationary as well as (stochastically) trending regressors and can readily be implemented using *Autometrics* in PcGive or the R-package *gets*.

JEL Classification: C12, C52

Keywords: misspecification; outlier detection; robust estimation; indicator saturation

[*] Financial support from the Robertson Foundation, British Academy, the Oxford Martin School, and INET, is gratefully acknowledged. We thank Jennifer L. Castle, David F. Hendry, Andrew B. Martinez, Bent Nielsen, and Max Roser for helpful comments and suggestions. Corresponding author: felix.pretis@nuffield.ox.ac.uk.

1 Introduction

The presence of outlying observations in a regression model can be indicative of model misspecification. For example, estimating a linear model when the underlying data generating process (DGP) is nonlinear, may result in a high number of outliers post-estimation, consequently, it is important to check for possible outlier contamination. However, algorithms used to detect outliers have a positive probability of finding outliers even when, in fact, the data generation process has no outliers. Using the concept of a gauge – the expected retention rate of falsely discovered outliers – we propose two sets of tests of the overall presence of outliers and thus model-misspecification: first, a set of tests on whether the observed proportion and number of detected outliers deviate from their expected values in a well-specified model. Second, a set of 'scaling' tests on whether the number of detected outliers decreases proportionally with the level of significance used to detect outliers. The first set of tests has power against the number of outliers present, while the second set has power against both outlier magnitude and number.

We consider outlier detection algorithms of the form of iterated 1-step Huber-skip M-estimators, in particular we focus on robustified least-squares (RLS) and impulse indicator saturation (IIS – see Hendry et al. 2008). In these algorithms, the rate of false-discovery (or gauge) can be controlled through a cut-off threshold used to define whether an observation is an outlier or not. IIS has seen increased use as a method to asses model mis-specification (see studies listed in Table 2 in section 4), however, due to a lack of distributional results on the gauge, there was no obvious cut-off rule or threshold of observed outliers relative to the expected number of outliers that would classify a model as misspecified (see the discussion by Doornik & Hendry 2016 and Hendry & Mizon 2011a). Here we derive the asymptotic distribution of the the gauge to construct a set of simple tests for model misspecification. We assess whether the observed proportion of observations classified as outliers significantly deviates from the proportion expected when there are no outliers. For example, if outlier detection takes place at a cut-off threshold such that we expect 1% of the sample to be classified as outliers by chance (even when there are none), and subsequently 2% of the sample are classified as outliers, is the difference between the nominal level (1%) and the observed proportion (2%) sufficiently large to reject the null hypothesis of no outliers?

This first set of tests focuses exclusively on the number of outliers, not taking outlier magnitude into account. We therefore propose a second set of tests, which we refer to as 'scaling' tests. These account for both the number and magnitude of outliers by assessing whether the detected proportion of outliers scales proportionally to the level of the cut-off in outlier detection. For example, if there are no outliers in the DGP, then detecting outliers at a significance level of 5% should result in a proportionally higher number of outliers detected relative to detection at 1%. By repeated testing at different levels of detection significance, this proportional scaling can be used to test the global hypothesis of no outliers being present.

We show that when there are no outliers present, the gauge of RLS and IIS follows an asymptotic normal distribution centred around the nominal level of significance used to detect outliers. The number of outliers can be approximated by the Poisson distribution. We further show that the scaling tests based on the sum of detected outliers proportions follows an asymptotic normal distribution, with the supremum approximated by a

Gaussian process for which the critical values can be simulated. The distributional results hold for regression models with covariates that can be random variables and deterministically or stochastically trending. Simulation results show that the standard normal test on the proportion of outliers exhibits a size close to the nominal level for large ($n \geq 100$) samples. In small, samples testing on a Poisson approximation of the number is preferred. The scaling tests exhibit a size close to the nominal level and have desirable power properties under a range of alternatives – both for increasing magnitude as well as number of outliers. The tests of the presence of outliers can be straight-forwardly implemented using the R-package *gets* (Pretis et al. 2017) or the *Autometrics* algorithm in PcGive (Doornik and Hendry 2014).

1.1 Related Literature

The focus of the broader literature on outliers has remained predominantly on methods of detection, rather than the explicit testing of the overall presence of outliers. One-step estimators have been considered in Bickel (1975) and Ruppert & Carroll (1980). In particular the 1-step Huber-skip estimator was studied by Welsh & Ronchetti (2002). Asymptotic theory of 1-step M-estimators was established by Hendry, Johansen and Santos (2008) for the location and scale model, and by Johansen and Nielsen (2009) for time series regressions. Iteration of these estimators was investigated in Johansen & Nielsen (2013). Hoover and Perez (1999) originally introduced the idea of a gauge in a simulation study of general-to-specific variable selection algorithms, subsequently formally discussed by Hendry & Santos (2010) as the expected retention rate of irrelevant regressors in the context of model selection algorithms. Doornik (2009) presented a comprehensive simulation study on the gauge for the model selection algorithm Autometrics (see also Hendry and Doornik, 2014). Johansen & Nielsen (2016b) provide an asymptotic analysis for the gauge, where Jiao & Nielsen (2017) extended the point-wise asymptotics to the uniform convergence theory which allows the cut-off and the number of iterations both to increase.

Comprehensive overviews of outlier detection approaches beyond iterated 1-step estimators are provided in the earlier surveys in Hodge & Austin (2004) and Chandola et al. (2009). On the testing side, Grubbs (1969) provides one of the first formal tests for the presence of outliers by iteratively testing observations in a single series. In the context of regression modelling, Tietjen et al. (1973) propose a test for a single outlier in simple linear regression based on standardised residuals. This approach was extended by Prescott (1975). A concern with post-estimation tests on standardized residuals is that the residuals themselves are determined during estimation which may be affected by the presence of outliers. More recently Srivastava & von Rosen (1998) provide likelihood ratio tests for single outliers.

In contrast to the existing literature, here we construct overall tests for the presence of outliers based on the proportion of observations classified as outliers. Even though the focus in the indicator saturation and RLS literature has remained predominantly on time series, our proposed tests can be readily applied to cross-sectional or panel data. Further use extends to analysis of model output where the underlying model is unknown or analytically intractable such as large-scale simulated models in economics or climate research. The remainder of the paper is structured as follows. Section 2.1 introduces the underlying model with section 2.2 describing the algorithms used to detect outliers. Sections 2.3 and 2.4 derive the distribution of the gauge for the two algorithms under consideration: RLS and IIS. Section 2.5 then derives the distribution of the scaling tests. The simulation performance of the tests is considered in section 3, section 4 considers applications to a cross-sectional macroeconomic model of economic growth and re-visits a set of published studies using IIS. Section 5 concludes.

2 Testing the Presence of Outliers using Iterated Estimators

2.1 Model

We consider data (y_i, x_i), $i = 1, 2, \dots, n$, where y_i is univariate and x_i may be multivariate with dimension $\dim x$. Assume the data satisfies the regression equation in (2.1):

$$y_i = x_i'\beta + \varepsilon_i, \quad i = 1, 2, \dots, n. \tag{2.1}$$

This setting can represent both classical regression and time series models. Moreover, regressors x_i can be deterministic or stochastically trending. We assume innovations ε_i are independent of the filtration $\mathcal{F}_{i-1}$ generated by $(x_1, \dots, x_i, \varepsilon_1, \dots, \varepsilon_{i-1})$, and are identically distributed with scale σ so that ε_i/σ has the known density f and distribution $\mathsf{F}(c) = \mathsf{P}(\varepsilon_i/\sigma \le c)$. In practice, the innovation distribution, characterised by f, F, will often be assumed to be standard normal or at least symmetric. For the absolute error $|\varepsilon_i|/\sigma$ denote a density by g and a distribution function by $\mathsf{G}(c) = \mathsf{P}(|\varepsilon_i|/\sigma \le c)$ for $c > 0$. Now with a symmetry assumption, $\mathsf{G}(c) = 2\mathsf{F}(c) - 1$ and $\mathsf{g}(c) = 2\mathsf{f}(c)$. Define $\psi_c = \mathsf{G}(c)$ so the probability of exceeding the cut-off c is $\gamma_c = 1 - \psi_c$. Suppose the k-th moment of the density f exists, then we introduce

$$\tau_k = \int_{-\infty}^{\infty} u^k \mathsf{f}(u) du, \qquad \tau_k^c = \int_{-c}^{c} u^k \mathsf{f}(u) du. \tag{2.2}$$

Thus $\tau_0^c = \psi_c$, $\tau_2 = 1$ while $\tau_k = \tau_k^c = 0$ for odd k when assuming symmetry. We define the conditional variance of ε_i/σ given scaled innovations fall below a cutoff-threshold $(|\varepsilon_i|/\sigma \le c)$ as

$$\varsigma_c^2 = \frac{\tau_2^c}{\psi_c} = \frac{\int_{-c}^{c} u^2 \mathsf{f}(u) du}{\mathsf{P}(|\varepsilon_i| \le \sigma c)}. \tag{2.3}$$

This will be used as a bias correction factor for the variance estimate computed from the selected non-outlying sample. For a standard normal reference distribution, we have $\tau_2^c = \psi_c - 2c\mathsf{f}(c)$, $\tau_4^c = 3\psi_c - 2c(c^2 + 3)\mathsf{f}(c)$ and $\tau_4 = 3$.

A model of y_i may be misspecified in a number of ways (from incorrect functional form, changing variance, to omitting variables, non-normality of the residuals, and mis-measured data) potentially resulting in a series of outlying observations in the estimated mis-specified model. The aim of the proposed tests here is to test the above DGP (2.1) for the presence of outliers identified by indicator variables. Outlier detection algorithms considered here use absolute residuals and then calculate robust least squares estimators from the non-outlying sample. This implicitly assumes symmetry, while non-symmetry leads to bias. All our asymptotic analysis is carried out under the null hypothesis that there are no outliers contained in the model matching the DGP (2.1).

2.2 Algorithms to Detect Outliers: Iterated 1-step Huber-skip M-estimators

We consider two algorithms to detect outliers in regression models, RLS and IIS. Both algorithms can be analysed as iterated 1-step Huber-skip M-estimators which mimic the Huber (1964) skip estimator, which has criterion function $\rho(t) = \min(t^2, c^2)/2$ as opposed to the Huber estimator with criterion function $\rho(t) = t^2/2$ for $|t| \leq c$ and $\rho(t) = c|t| - c^2/2$ otherwise, see Hampel et al. 1986 (p. 104) and Jurečková et al. 2013 (p. 175). We define the iterated 1-step Huber-skip M-estimator in algorithm 2.1.

Algorithm 2.1. ***Iterated 1-step Huber-skip M-estimator***. *Choose a cut-off* $c > 0$.
1. Choose initial estimators $\widehat{\beta}_c^{(0)}$, $(\widehat{\sigma}_c^{(0)})^2$ *and let* $m = 0$.
2. Define indicator variables for selecting non-outlying observations

$$v_{i,c}^{(m)} = 1_{(|y_i - x_i'\widehat{\beta}_c^{(m)}| \leq \widehat{\sigma}_c^{(m)} c)}. \tag{2.4}$$

3. Compute least squares estimators

$$\widehat{\beta}_c^{(m+1)} = (\sum_{i=1}^{n} x_i x_i' v_{i,c}^{(m)})^{-1} (\sum_{i=1}^{n} x_i y_i v_{i,c}^{(m)}), \tag{2.5}$$

$$(\widehat{\sigma}_c^{(m+1)})^2 = \varsigma_c^{-2} (\sum_{i=1}^{n} v_{i,c}^{(m)})^{-1} \{\sum_{i=1}^{n} (y_i - x_i'\widehat{\beta}_c^{(m+1)})^2 v_{i,c}^{(m)}\}. \tag{2.6}$$

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

The algorithm could start with a robust estimator, while RLS is initiated using the full sample least squares. The latter is not robust with respect to high leverage points in cross sectional data.[1] Another example is IIS (initially proposed by Hendry 1999) and first studied by Hendry, Johansen and Santos (2008). The basic idea in IIS is to divide full sample into two sub-samples and use regression estimates calculated from each sub-sample to detect outliers in the other sub-sample. This 'split-half' approach is used to derive the theory of IIS (see algorithm 2.2), however, as the location of contaminated observations is unknown in most practical situations, the initial sets $\mathcal{I}_1$ and $\mathcal{I}_2$ should be iterated as is currently implemented in IIS in the R-package *gets* (Pretis et al. 2017) as well as *Autometrics* (Doornik, 2009).

Algorithm 2.2. ***Impulse Indicator Saturation***. *Choose a cut-off* $c > 0$.
1.1. Split full sample into two sets $\mathcal{I}_j$, $j = 1, 2$ *of* n_j *observations where* $\sum_{j=1}^{2} n_j = n$.
1.2. Calculate least squares estimators based upon each sub-sample $\mathcal{I}_j$ *for* $j = 1, 2$

$$\widehat{\beta}_j = (\sum_{i \in \mathcal{I}_j} x_i x_i')^{-1} (\sum_{i \in \mathcal{I}_j} x_i y_i), \qquad \widehat{\sigma}_j^2 = \frac{1}{n_j} \sum_{i \in \mathcal{I}_j} (y_i - x_i'\widehat{\beta}_j)^2. \tag{2.7}$$

[1] Leverage points seem to be less of a problem in time series models when lagged variables are included as regressors.

1.3. Define the initial indicator variables for selecting non-outlying observations

$$v_{i,c}^{(-1)} = 1_{(i\in\mathcal{I}_1)}1_{(|y_i - x_i'\widehat{\beta}_2| \leq \widehat{\sigma}_2 c)} + 1_{(i\in\mathcal{I}_2)}1_{(|y_i - x_i'\widehat{\beta}_1| \leq \widehat{\sigma}_1 c)}. \tag{2.8}$$

1.4. Compute $\widehat{\beta}_c^{(0)}$, $(\widehat{\sigma}_c^{(0)})^2$ *using* (2.5), (2.6) *with* $m = -1$, *and then let* $m = 0$.
2. Follow the step 2,3,4 in Algorithm 2.1.

IIS is possibly more robust than RLS when we have prior knowledge that outliers are located in a particular subset of the whole sample.

2.3 The Rate of False-Detection: Gauge

Outlier detection algorithms have a positive probability of finding outliers even when, in fact, the data generation process has no outliers. We evaluate the performance of such algorithms by the concept of a gauge, which is the expected retention rate of falsely discovered outliers. This is a measure of type I error and it gives us an indirect way of choosing the cut-off c. It is defined as follows. The algorithms assign stochastic indicators $v_{i,c}^{(m)}$ to all observations such as in (2.4) so that $v_{i,c}^{(m)} = 0$ when observation i is declared as an outlier, otherwise $v_{i,c}^{(m)} = 1$. When the model has no contamination, the sample and population gauge are the proportion spuriously classified as outliers:

$$\widehat{\gamma}_c^{(m)} = \frac{1}{n}\sum_{i=1}^{n}(1 - v_{i,c}^{(m)}) \text{ and } \mathsf{E}\widehat{\gamma}_c^{(m)} = \frac{1}{n}\sum_{i=1}^{n}\mathsf{E}(1 - v_{i,c}^{(m)}). \tag{2.9}$$

Deriving distributional results of the gauge, we propose two sets of sets for the overall presence of outliers. First, tests on the proportion and number of outliers in section 2.4. Second, we propose tests on whether the proportion (or number) of outliers scales proportionally to the level of significance of outlier detection in section 2.5.

2.4 Testing on the Gauge: Proportion and Count Tests

Here we derive the asymptotic distribution of the gauge $\widehat{\gamma}_c^{(m)}$ under the null of no outliers to test whether the detected proportion of outliers significantly deviates from the expected proportion of outliers when there are no outliers present. The main result of asymptotic normality of the gauge is presented in theorems 2.3 and 2.4. In particular, theorem 2.3 shows the asymptotic normal distribution of the gauge for a robustified least squares estimator, and theorem 2.4 shows that the distribution of the gauge under IIS matches that of the robustified implementation. A Poisson approximation to the number of outliers is shown in theorem 2.6.

We then extend these results to a set of scaling tests, where the cutoff threshold c is varied, and we assess whether the detected proportion of outliers scales proportionally with c as expected under the null of no outliers. Two scaling tests are derived based on the sum and supremum of the detected proportion of outliers, with the asymptotic normal distributions established in theorems 2.7 and 2.8. The result is uniform in the cut-off value, which generalises the results in Johansen and Nielsen (2016b) where either the threshold or number of iterations is fixed. This allows us to analyse the scaling test when the cut-off value is drifting. We consider a third implementation of a scaling test using corrections for multiple testing in section 2.5.3. Proofs of theorems are provided in the appendix.

2.4.1 Assumptions

Innovations ε_i and regressors x_i must satisfy moment conditions in Assumption 2.1 so as to carry out asymptotic analysis. Regressors x_i can be temporally dependent and deterministically or stochastically trending. We therefore require a normalisation matrix N that allows for different behaviour of the components of the regressor vector x_i. In the case of a stationary regressor we need a standard $n^{-1/2}$ normalisation so that N must be proportional to the identity matrix of the same dimension as x_i, that is $N = n^{-1/2} I_{\dim x}$. Likewise, if x_i is a random walk we have $N = n^{-1} I_{\dim x}$. If the regressors are unbalanced as in $x_i = (1, i)'$ we can choose $N = \mathrm{diag}(n^{-1/2}, n^{-3/2})$.

Assumption 2.1. *Let $\mathcal{F}_i$ be an increasing sequence of σ-fields so ε_{i-1} and x_i are $\mathcal{F}_{i-1}$ measurable and ε_i is independent of $\mathcal{F}_{i-1}$. Let ε_i/σ have a symmetric, continuously differentiable density* f *which is positive on the real line $\mathbb{R}$. For some values of κ, η such that $0 \leq \kappa < \eta \leq 1/4$, choose an integer $r \geq 2$ so $2^{r-1} \geq 1 + (1/4 + \kappa - \eta)(1 + \dim x)$. Let $q = 1 + 2^{r+1}$. Suppose*
(i) *the density* f *satisfies*
(a) $u^q \mathsf{f}(u)$, $|u^{q+1}\dot{\mathsf{f}}(u)|$ *are decreasing for large u;*
(b) $\mathsf{f}(u_n - n^{-1/4}A)/\mathsf{f}(u_n) = \mathrm{O}(1)$ *as $n \to \infty$ for some $A > 0$ and all sequences $u_n \to \infty$ so $u_n = \mathrm{o}(n^{1/4})$;*
(c) $\mathsf{f}(u)/[u\{1 - \mathsf{F}(u)\}] = \mathrm{O}(1)$ *for $u \to \infty$;*
(ii) *the regressors x_i satisfy*
(a) $\Sigma_n = \sum_{i=1}^n N' x_i x_i' N \overset{\mathsf{P}}{\to} \Sigma \overset{a.s.}{>} 0$;
(b) $\max_{1 \leq i \leq n} |n^{1/2-\kappa} N' x_i| = \mathrm{O}_{\mathsf{P}}(1)$;
(c) $n^{-1} \mathsf{E} \sum_{i=1}^n |n^{1/2} N' x_i|^q = \mathrm{O}(1)$;
(iii) *the initial estimator* $(\widetilde{\beta}, \widetilde{\sigma}^2)$ *satisfies*
(a) $N^{-1}(\widetilde{\beta} - \beta) = \mathrm{O}_{\mathsf{P}}(n^{1/4-\eta})$;
(b) $n^{1/2}(\widetilde{\sigma}^2 - \sigma^2) = \mathrm{O}_{\mathsf{P}}(n^{1/4-\eta})$.

There is a trade-off between κ, η, the dimension $\dim x$ and the required number of moments r, see Johansen & Nielsen 2016a (Remark 3.1). Conditions (i), (ii) are satisfied in a range of situations. In particular, the condition (ia) is satisfied by the normal and t distribution, see Johansen and Nielsen (2016a, Example 3.1), while the condition (ib, ic) is satisfied by the normal distribution, see Johansen and Nielsen (2016b, Remark 2). Condition (ii) is satisfied by stationary, random walk and deterministically trending regressors, see Johansen and Nielsen (2016a, Example 3.2). Condition (iii) allows the standardised estimation errors to diverge at a rate of $n^{1/4-\eta}$ rather than being bounded in probability. In particular, $\eta = 1/4$ can be chosen for estimators with standard convergence rates.

2.4.2 Gaussian and Poisson approximation to the Gauge

The distributional results are structured as follows: initially a stochastic expansion and tightness result is given for the gauge of iterated estimators defined in Algorithm 2.1 and 2.2, and a weak convergence theory follows. A Gaussian process theory arises when the proportion of detected outliers is controlled by choosing the cut-off value c as n increases, whereas a Poisson exceedence theory is established when the c is set to allow a fixed number of outliers regardless of the sample size n. The proof involves the empirical process theory developed by Jiao and Nielsen (2017). Thus we first present the asymptotic expansion of the following empirical process.

Lemma 2.1. *Suppose Assumption 2.1(ia, ii) holds. Let $v_i^{a,b,c} = 1_{(|\varepsilon_i - x_{in}'b| \leq \sigma c + n^{-1/2}ac)}$. Then we have an expansion*

$$n^{-1/2}\sum_{i=1}^{n} v_i^{a,b,c} = n^{-1/2}\sum_{i=1}^{n} 1_{(|\varepsilon_i| \leq \sigma c)} + 2\mathsf{f}(c)\frac{ac}{\sigma} + R(a,b,c).$$

Then for any $B > 0$ and as $n \to \infty$

$$\sup_{0<c<\infty}\ \sup_{|a|,|b| \leq n^{1/4-\eta}B} |R(a,b,c)| = \mathrm{o}_{\mathsf{P}}(1).$$

The proof is given in the appendix.

Notice this empirical process result generalises Johansen and Nielsen (2009, 2016a) which did either allow variation in a, b but fixed c or variation in b, c but fixed a. Thus, built on their restricted empirical process result, Johansen and Nielsen (2016b) only considered pointwise convergence theory in c for the gauge. One contribution of the present paper is to extend pointwise asymptotics to convergence that is uniform in c. A stochastic expansion is given for the gauge of iterated 1-step Huber-skip M-estimator algorithms.

Theorem 2.1. *Consider the iterated 1-step Huber-skip M-estimator in Algorithm 2.1. Suppose Assumption 2.1(ia, ii) holds, and that $N^{-1}(\widehat{\beta}_c^{(m)} - \beta)$, $n^{1/2}(\widehat{\sigma}_c^{(m)} - \sigma)$ are $\mathrm{O}_{\mathsf{P}}(1)$. Then uniformly in $m \in [0, \infty)$, $c \in \mathbb{R}_+$ and as $n \to \infty$*

$$n^{1/2}(\widehat{\gamma}_c^{(m)} - \gamma_c) = n^{-1/2}\sum_{i=1}^{n}\{1_{(|\varepsilon_i| > \sigma c)} - \gamma_c\} - 2c\mathsf{f}(c)n^{1/2}(\frac{\widehat{\sigma}_c^{(m)}}{\sigma} - 1) + \mathrm{o}_{\mathsf{P}}(1).$$

The proof is given in the appendix.

The next result shows the gauge is tight uniformly in iteration $m \in [0, \infty)$ and in the cut-off value $c \in [c_0, \infty)$. Note $c_0 > 0$ is a finite number.

Theorem 2.2. *Consider the iterated 1-step Huber-skip M-estimator in Algorithm 2.1. Suppose Assumption 2.1(ia, ii, iii) holds with $\eta = 1/4$. Then as $n \to \infty$*

$$\sup_{0 \leq m < \infty}\ \sup_{c_0 \leq c < \infty} |n^{1/2}(\widehat{\gamma}_c^{(m)} - \gamma_c)| = \mathrm{O}_{\mathsf{P}}(1).$$

We control the proportion of discovered outliers by selecting the cut-off value c as n increases, and obtain a process of the gauge by varying c. A weak convergence theory follows from a finite dimensional convergence result derived by the expansion in Theorem 2.1 and tightness in Theorem 2.2, see Billingsley 2013 (Theorem 13.1). Here we analyse RLS and IIS by asymptotically approximating the process of their gauge as a Gaussian process (see the definition of Gaussian processes in Adler & Taylor 2009, p. 27).

Theorem 2.3. ***Gaussian approximation to the gauge for Robustified Least Squares****: Consider the Robustified Least Squares, where the initial estimators $\widehat{\beta}_c^{(0)} = \widetilde{\beta}$, $(\widehat{\sigma}_c^{(0)})^2 = \widetilde{\sigma}^2$ are the full sample least squares estimators. Suppose Assumption 2.1(ia, ii) holds. Let $\mathbb{G}_n(c) = n^{1/2}(\widehat{\gamma}_c^{(0)} - \gamma_c)$ for $c \in [c_0, \infty)$. Then as $n \to \infty$*

$$\mathbb{G}_n \rightsquigarrow \mathsf{GP}(0, \Sigma),$$

where $\mathsf{GP}(c)$ for $c \in [c_0, \infty)$ is the Gaussian process with mean $\mathsf{E}\{\mathsf{GP}(t)\} = 0$ and variance-covariance structure $\Sigma_{st} = \mathsf{Cov}\{\mathsf{GP}(s), \mathsf{GP}(t)\}$ so

$$\Sigma_{st} = \gamma_t(1-\gamma_s) + s\mathsf{f}(s)(\tau_2^t + \gamma_t - 1) + t\mathsf{f}(t)(\tau_2^s + \gamma_s - 1) + st\mathsf{f}(s)\mathsf{f}(t)(\tau_4 - 1),$$

for any $c_0 \leq s \leq t < \infty$. Finite dimensional convergence follows by the CLT so for any $c \in [c_0, \infty)$ then

$$\mathbb{G}_n(c) = n^{1/2}(\widehat{\gamma}_c^{(0)} - \gamma_c) \overset{\mathsf{D}}{\to} \mathsf{N}\{0, \gamma_c(1-\gamma_c) + 2c\mathsf{f}(c)(\tau_2^c + \gamma_c - 1) + c^2\mathsf{f}^2(c)(\tau_4 - 1)\}. \tag{2.10}$$

The proof is given in the appendix.

Theorem 2.3 establishes the asymptotic normality of the gauge when detecting outliers using RLS. The following theorem 2.4 shows that the same results apply when detecting outliers using IIS.

Theorem 2.4. ***Gaussian approximation to the gauge for Impulse Indicator Saturation****: Consider the split-half Impulse Indicator Saturation in Algorithm 2.2 where $n_1 = n_2 = n/2$. Suppose Assumption 2.1(ia, ii) holds for each subsample set $\mathcal{I}_1$, $\mathcal{I}_2$. The sample gauge is*

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

Then as $n \to \infty$ the process of the normalized gauge $n^{1/2}(\widehat{\gamma}_c^{(-1)} - \gamma_c)$ for $c \in [c_0, \infty)$ weakly converges to the same Gaussian process as for RLS reported in Theorem 2.3. The proof is given in the appendix.

The above weak convergence result for the gauge does not depend on the stochastic properties of regressors, which can be stationary, or deterministically or stochastically trending, since only the variance estimator appears in the asymptotic expansion in Theorem 2.1. For the 1-step Huber-skip M-estimator either starting from the full sample or split-half sample least squares, the gauge has the same limiting Gaussian process. Denote $\widetilde{\gamma}_c$ as $\widehat{\gamma}_c^{(0)}$ in RLS or as $\widehat{\gamma}_c^{(-1)}$ in IIS. The weak convergence in Theorems 2.3, 2.4 implies the pointwise asymptotic result in Johansen and Nielsen (2016b) so that for any $c \in [c_0, \infty)$

$$n^{1/2}(\widetilde{\gamma}_c - \gamma_c) \overset{\mathsf{D}}{\to} \mathsf{N}\{0, \gamma_c(1-\gamma_c) + 2c\mathsf{f}(c)(\tau_2^c + \gamma_c - 1) + c^2\mathsf{f}^2(c)(\tau_4 - 1)\}. \tag{2.11}$$

We have thus established the asymptotic distribution of the sample gauge $\widetilde{\gamma}_c$ under the null of no outliers. The sample gauge using RLS or IIS follows an asymptotic normal distribution given by equation (2.11) centred around γ_c defined by the chosen cutoff c (e.g. $\gamma_c = 0.05$ for $c = 1.96$). A fixed point can also be derived for the gauge as in theorem 2.5.

Theorem 2.5. *Consider the iterated 1-step Huber-skip M-estimator in Algorithm 2.1. Suppose Assumption 2.1(ia, ii, iii) holds with $\eta = 1/4$. Then for all $\epsilon, \delta > 0$ a pair $m_0, n_0 > 0$ exists, so for all $m > m_0$ and $n > n_0$*

$$\mathsf{P}\{\sup_{c_0 \le c < \infty} |n^{1/2}(\widehat{\gamma}_c^{(m)} - \widehat{\gamma}_c^*)| > \delta\} < \epsilon,$$

where

$$n^{1/2}(\widehat{\gamma}_c^* - \gamma_c) = n^{-1/2}\sum_{i=1}^n \{1_{(|\varepsilon_i| > \sigma c)} - \gamma_c\}$$
$$- \frac{c\mathsf{f}(c)}{\tau_2^c - c(c^2 - \varsigma_c^2)\mathsf{f}(c)} n^{-1/2} \sum_{i=1}^n (\frac{\varepsilon_i^2}{\sigma^2} - \varsigma_c^2) 1_{(|\varepsilon_i| \le \sigma c)} + \mathsf{o}_{\mathsf{P}}(1),$$

uniformly in $c \in [c_0, \infty)$. Let $\mathbb{G}_n^*(c) = n^{1/2}(\widehat{\gamma}_c^* - \gamma_c)$ for $c \in [c_0, \infty)$. Then as $n \to \infty$

$$\mathbb{G}_n^* \rightsquigarrow \mathsf{GP}(0, \Sigma),$$

where $\mathsf{GP}(c)$ for $c \in [c_0, \infty)$ is the Gaussian process with mean $\mathsf{E}\{\mathsf{GP}(t)\} = 0$ and variance-covariance structure $\Sigma_{st} = \mathsf{Cov}\{\mathsf{GP}(s), \mathsf{GP}(t)\}$ so

$$\Sigma_{st} = \gamma_t(1-\gamma_s) - \frac{t\mathsf{f}(t)}{\tau_2^t - t(t^2 - \varsigma_t^2)\mathsf{f}(t)}\{\varsigma_t^2(1-\gamma_s) - \tau_2^s\}$$
$$+ \frac{s\mathsf{f}(s)}{\tau_2^s - s(s^2 - \varsigma_s^2)\mathsf{f}(s)} \frac{t\mathsf{f}(t)}{\tau_2^t - t(t^2 - \varsigma_t^2)\mathsf{f}(t)}\{\tau_4^s - \frac{(\tau_2^s)^2}{1-\gamma_s}\},$$

for any $c_0 \le s \le t < \infty$. The proof is given in the appendix.

While the gauge can be approximated by a normal distribution, the number of outliers (given by $\widehat{\gamma}_c^{(m)} n$) can be approximated by a Poisson distribution. Johansen and Nielsen (2016b) proved the Poisson approximation to the gauge for the finite step Huber-skip M-estimator, while the iterated result was established by Jiao and Nielsen (2017). A Poisson exceedence theory arises in the scenario where the cut-off value c is set to allow the fixed number λ of outliers regardless of the sample size n. For some $\lambda > 0$, the cut-off value c_n is set so as to let

$$n\mathsf{P}(|\varepsilon_i| > \sigma c_n) = \lambda. \tag{2.12}$$

Notice that $c_n \to \infty$ as $n \to \infty$. Define $v_{i,c_n}^{(m)}$, $\widehat{\beta}_{c_n}^{(m+1)}$, $(\widehat{\sigma}_{c_n}^{(m+1)})^2$ by replacing c by c_n in expressions (2.4), (2.5), (2.6). The corresponding sample gauge is

$$\widehat{\gamma}_{c_n}^{(m)} = \frac{1}{n}\sum_{i=1}^{n}(1 - v_{i,c_n}^{(m)}) = \frac{1}{n}\sum_{i=1}^{n} 1_{(|y_i - x_i'\widehat{\beta}_{c_n}^{(m)}| > \widehat{\sigma}_{c_n}^{(m)} c_n)}. \tag{2.13}$$

Theorem 2.6. *(Jiao and Nielsen, 2017, Theorem 4).* ***Poisson Approximation****: Consider the iterated 1-step Huber-skip M-estimator in Algorithm 2.1. Let c_n be defined from (2.12). Suppose Assumption 2.1 holds with $\eta = 1/4$. Then for all $0 \leq m < \infty$ and as $n \to \infty$*

$$n\widehat{\gamma}_{c_n}^{(m)} \overset{\mathsf{D}}{\to} \mathsf{Poisson}(\lambda).$$

The Poisson approximation is uniform in the iteration step m, thus more robust than the Gaussian approximation since the gauge in the different step m has the distinct variance-covariance structure in the limiting Gaussian process. RLS and IIS are special versions of iterated 1-step Huber-skip M-estimators with different starting points. Their initial points do not depend on the cut-off, and thus satisfy the tightness property. Therefore, the above theorems apply for these algorithms.

2.4.3 *Outlier Proportion Test*: Standard Normal Test on the Proportion of Outliers

Using the distributional result in Theorems 2.3, 2.4 and equation (2.11), a simple test for whether the proportion of outliers is different from zero can be constructed. If there are no outliers then the proposed test statistic z_γ in (2.14) follows an asymptotic standard normal distribution:

$$z_\gamma = \frac{\widetilde{\gamma}_c - \gamma_c}{\sqrt{\omega_{\gamma_c}}} \underset{H_0}{\overset{a}{\sim}} N(0,1) \tag{2.14}$$

where the variance term ω_{γ_c} is given by[2]

$$\omega_{\gamma_c} = n^{-1}\left[\gamma_c(1-\gamma_c) + 2c\mathsf{f}(c)(\tau_2^c + \gamma_c - 1) + c^2\mathsf{f}^2(c)(\tau_4 - 1)\right] \tag{2.15}$$

where for a standard normal reference distribution with density f, we have $\tau_2^c = \psi_c - 2c\mathsf{f}(c)$, $\tau_4^c = 3\psi_c - 2c(c^2+3)\mathsf{f}(c)$, $\tau_4 = 3$, and $\psi_c = 1 - \gamma_c$.

The proposed test procedure, referred to here as the 'proportion test', is then to estimate a model using RLS or IIS at a chosen selection significance level γ_c (e.g. 0.05 for $c = 1.96$), record the observed proportion of outliers $\widetilde{\gamma}_c$ by dividing the number of retained impulse indicators by the total number of indicators selected over, construct and compare z_γ to the critical values of a standard normal distribution. It is worth emphasizing that there are two significance levels at play when testing on the proportion of outliers: First, selection in RLS

[2] Note that the variance expression (2.15) uses γ_c rather than $\widetilde{\gamma}_c$. While using $\widetilde{\gamma}_c$ may provide a better approximation for the variance of $\widetilde{\gamma}_c$ under the null, the power under the alternative would be drastically lower – consider the extreme case of selecting at $\gamma_c = 0.05$ and detecting $\hat{\gamma} = 0.3$ of the sample as outliers, then $\widetilde{\gamma}_c(1-\widetilde{\gamma}_c) = 0.3 \times 0.7 = 0.21$ while $\gamma_c(1-\gamma_c) = 0.0475$.

or IIS at γ_c determines the proportion of outliers expected under the null hypothesis ($\mathsf{E}[\widetilde{\gamma}_c] = \gamma_c$), second the choice of the significance level p for then testing on z_γ determines the null rejection frequency of the outlier proportion test.

For example, consider selecting in IIS at $\gamma_c = 0.01$ (with $c = 2.576$) in a sample of $n = 100$ observations, resulting in the retention of three impulse indicators corresponding to three outlying observations. Then the observed proportion of outliers is equal to $\widetilde{\gamma}_c = 3/100 = 0.03$ with approximate standard deviation of $\sqrt{\omega}_{\gamma_c=0.01}$ $= 100^{-1/2}\ [0.01(1-0.01) - 0.0028]^{1/2} = 0.00844$. Comparing $z_{\gamma=0.01} = (0.03 - 0.01)/0.00844 = 2.37$ to the critical values of a standard normal distribution, the null hypothesis of no outliers is rejected for $p = 0.05$ ($z_{\gamma=0.01} >$ 1.96) while for $p = 0.01$ ($z_{\gamma=0.01} <$ 2.576) it is not.

2.4.4 *Outlier Count Test*: Poisson Test on the Number of Outliers

An alternative is to consider choosing the critical value of selection to permit an acceptable number λ_γ of spuriously retained outliers rather than a proportion of the sample: in this case a Poisson approximation for the number of retained outliers under the null hypothesis of no outliers can be used (Theorem 2.6). Referred to here as the 'count test', the estimated number of outliers $n\widetilde{\gamma}_c$ can be evaluated against the critical values of a $\mathsf{Poisson}(\lambda_\gamma)$ distribution (where $\lambda_\gamma = n\gamma_c$):

$$\hat{k}_\gamma = n\widetilde{\gamma}_c \underset{H_0}{\overset{a}{\sim}} \text{Poisson}(\lambda_\gamma) \tag{2.16}$$

Continuing the above example of having detected $n\widetilde{\gamma}_c = 3$ outliers in a sample of $N = 100$ observations, testing against a Poisson($\lambda_{\gamma=0.01}$) distribution (where $\lambda_{\gamma=0.01} = 0.01 \times 100 = 1$) yields a p-value of $p = 0.08$ for both a two-sided and one-sided test of the sample number of outliers exceeding the hypothesized number.

2.5 Testing on the Gauge: Scaling Test Statistics

A simple test for the null hypothesis H_0 that there are no outliers in the DGP can be constructed from pointwise normal and Poisson approximation to the sample gauge by Theorem 2.3, 2.4, 2.6. However, testing on the proportion or number of outliers by simple normal or Poisson distributions ignores information about magnitudes of discovered outlying observations. A small number of extremely large outliers might be an indicator of model misspecification, however, the simple normal or Poisson tests lose the power in this scenario. For example, suppose the true proportion of large outliers in the DGP is 5%, then using the proportion outlier test at $\gamma_c = 0.05$ may lead to non-rejection of the null-hypothesis of no outliers because by chance the expected proportion of outliers coincides to the true proportion of outliers. A complete assessment ideally should not only take the proportion or number of detected outliers but also their magnitudes into account.

We therefore propose what we call 'scaling tests' to consider both the proportion and magnitudes of outliers by assessing whether the sample gauge scales proportionally to chosen levels of significance γ_c. Consider selection at increasingly conservative significance levels. Under the null H_0, it would be expected to observe a corresponding decrease in detecting proportion of outliers, whereas under the alternative H_1, especially in the case of a small number of large outliers, the sample gauge would maintain in a certain level and not have a large change when the level of significance is sufficiently small. We choose a set of K different levels of significance $\mathcal{S}_K = \{\gamma_{c_k}\}_{k=1}^K$ with corresponding cut-offs $\mathcal{C}_K = \{c_k\}_{k=1}^K$ through the one-to-one mapping $\gamma_c = 1 - \mathsf{G}(c)$ and $c = \mathsf{G}^{-1}(1-\gamma_c)$, where significance levels are chosen in an increasingly conservative way so $\gamma_{c_1} > \gamma_{c_2} > \cdots > \gamma_{c_K}$ and $c_1 < c_2 < \cdots < c_K$. Note c_1 is selected so $c_1 \geq c_0$ where c_0 is a small finite real value. Denote $\widetilde{\gamma}_c$ as $\widehat{\gamma}_c^{(0)}$ in RLS or $\widehat{\gamma}_c^{(-1)}$ in IIS. For each γ_{c_k} and c_k, run RLS or IIS and record sample gauges $\widetilde{\gamma}_{c_k}$ for $k = 1, 2, \ldots, K$, and subsequently test whether the sample gauge deviates from the population gauge for the full range of significance levels.

We consider three types of scaling tests. The first scaling test sums all deviations to construct the test statistic as $\sum_{c\in\mathcal{C}_K} n^{1/2}(\widetilde{\gamma}_c - \gamma_c)$. The second scaling test takes the supremum over all deviations so the test statistic is $\sup_{c\in\mathcal{C}_K} |n^{1/2}(\widetilde{\gamma}_c - \gamma_c)|$. The third type of test relies on testing the global hypothesis of no outliers based on repeated testing using the proportion or count tests, applying corrections for multiple-hypothesis testing. Intuitively, the sum test considers deviations from the expected proportion of outliers, the supremum test considers the highest deviation from the expected proportion, and the global test conducts multiple hypothesis tests corrected for repeated testing.

The next two theorems provide asymptotic distributions for the above proposed test statistics under the null H_0. Notice that the limiting distributions of two test statistics do not depend on the stochastic properties of regressors.

2.5.1 *Scaling Sum Test*

The proposed scaling sum test is to construct the test statistic S_{sum} given in equation (2.17) by summing over the entire set of deviations between the observed and expected proportion of detect outliers:

$$S_{sum} = \sum_{c\in\mathcal{C}_K} n^{1/2}(\widetilde{\gamma}_c - \gamma_c) \tag{2.17}$$

and then compare it against the critical values of a normal distribution with variance given in theorem 2.7.

Theorem 2.7. *Consider Robustified Least Squares or Impulse Indicator Saturation. Suppose Assumption 2.1(ia, ii) holds. Let $\widetilde{\gamma}_c$ be $\widehat{\gamma}_c^{(0)}$ in RLS or $\widehat{\gamma}_c^{(-1)}$ in IIS. Choose K different significance levels $\mathcal{S}_K = \{\gamma_{c_k}\}_{k=1}^K$ with corresponding cut-offs $\mathcal{C}_K = \{c_k\}_{k=1}^K$ such that $\gamma_{c_1} > \gamma_{c_2} > \cdots > \gamma_{c_K}$ and $c_1 < c_2 < \cdots < c_K$. Then under H_0 and as $n \to \infty$*

$$S_{sum} = \sum_{c\in\mathcal{C}_K} n^{1/2}(\widetilde{\gamma}_c - \gamma_c) \overset{\mathsf{D}}{\to} \mathsf{N}(0, \mathsf{Var}_{\mathcal{C}_K}), \tag{2.18}$$

where

$$\begin{aligned}\mathsf{Var}_{\mathcal{C}_K} = &\sum_{k=1}^K \gamma_{c_k}(1-\gamma_{c_k}) + 2\sum_{1\leq k<l\leq K}(\gamma_{c_l} - \gamma_{c_k}\gamma_{c_l}) + \{\sum_{k=1}^K c_k \mathsf{f}(c_k)\}^2(\tau_4 - 1)\\ &+ 2\{\sum_{k=1}^K c_k \mathsf{f}(c_k)\}\{\sum_{k=1}^K(\tau_2^{c_k} + \gamma_{c_k} - 1)\}.\end{aligned}$$

The proof is given in the appendix.

2.5.2 *Scaling Supremum Test*

The proposed scaling supremum test is to construct the test statistic S_{sup} considering the maximum over the entire set of deviations between the observed and expected proportion of detect outliers by taking the supremum as given in equation (2.19):

$$S_{sup} = \sup_{c \in \mathcal{C}_K} |n^{1/2}(\widetilde{\gamma}_c - \gamma_c)| \tag{2.19}$$

The asymptotic distribution of the scaling sup test is derived in theorem 2.8. In practise, critical values of the multivariate Gaussian process can be obtained via simulation.

Theorem 2.8. *Consider Robustified Least Squares or Impulse Indicator Saturation. Suppose Assumption 2.1(ia, ii) holds. Let $\widetilde{\gamma}_c$ be $\widehat{\gamma}_c^{(0)}$ in RLS or $\widehat{\gamma}_c^{(-1)}$ in IIS. Choose K different significance levels $\mathcal{S}_K = \{\gamma_{c_k}\}_{k=1}^K$ with corresponding cut-offs $\mathcal{C}_K = \{c_k\}_{k=1}^K$ such that $\gamma_{c_1} > \gamma_{c_2} > \cdots > \gamma_{c_K}$ and $c_1 < c_2 < \cdots < c_K$. Then under H_0 and as $n \to \infty$*

$$S_{sup} = \sup_{c \in \mathcal{C}_K} |n^{1/2}(\widetilde{\gamma}_c - \gamma_c)| \overset{\mathsf{D}}{\to} \sup_{c \in \mathcal{C}_K} |\mathsf{GP}(c; 0, \Sigma)|, \tag{2.20}$$

where $\mathsf{GP}(c)$ for $c \in [c_0, \infty)$ is the Gaussian process with mean $\mathsf{E}\{\mathsf{GP}(t)\} = 0$ and variance-covariance structure $\Sigma_{st} = \mathsf{Cov}\{\mathsf{GP}(s), \mathsf{GP}(t)\}$ so

$$\Sigma_{st} = \gamma_t(1-\gamma_s) + s\mathsf{f}(s)(\tau_2^t + \gamma_t - 1) + t\mathsf{f}(t)(\tau_2^s + \gamma_s - 1) + st\mathsf{f}(s)\mathsf{f}(t)(\tau_4 - 1),$$

for any $c_0 \leq s \leq t < \infty$.

Proof of Theorem 2.8. Apply the continuous mapping theorem to the weak convergence results in Theorem 2.3, 2.4 by Assumption 2.1(ia, ii), then the limiting distribution of the test statistic follows. ∎

2.5.3 *Scaling Global Test*

The third proposed scaling test for the presence of outliers is based on repeated testing using the proportion or count outlier tests from sections 2.4 and then using the Simes (1986) (see also Seeger 1968) correction for repeated testing. Similar to the sum and supremum scaling tests, run RLS or IIS at levels of significance $\mathcal{S}_K = \{\gamma_{c_k}\}_{k=1}^K$. For each γ_{c_k} conduct either a proportion or count test of the individual null hypothesis of no outliers. Let $\tilde{p}_k$ denote the kth observed p-value of the K hypothesis tests using either the proportion or count tests. Order the observed p-values of each test in ascending order as $\tilde{p}_{(1)} \leq \tilde{p}_{(2)} \leq \ldots, \leq \tilde{p}_{(K)}$. The rejection rule of the global hypothesis is then: reject the global hypothesis of no outliers if there exists a k in $(j = 1, \ldots, K)$ for which:

$$\tilde{p}_{(k)} \leq \frac{k}{K} p_{glob} \tag{2.21}$$

where p_{glob} is the chosen level of significance when testing the global hypothesis. In other words, reject the overall null hypothesis of no outliers H_0 if there exists an observed p-value from one of the tests that is lower than the threshold p-value adjusted for multiple testing. This approach controls the false-rejection rate at $\leq p_{glob}$ for independent tests (see Simes, 1986) as well as dependent tests (see Sarkar & Chang 1997, and Benjamini & Yekutieli 2001).

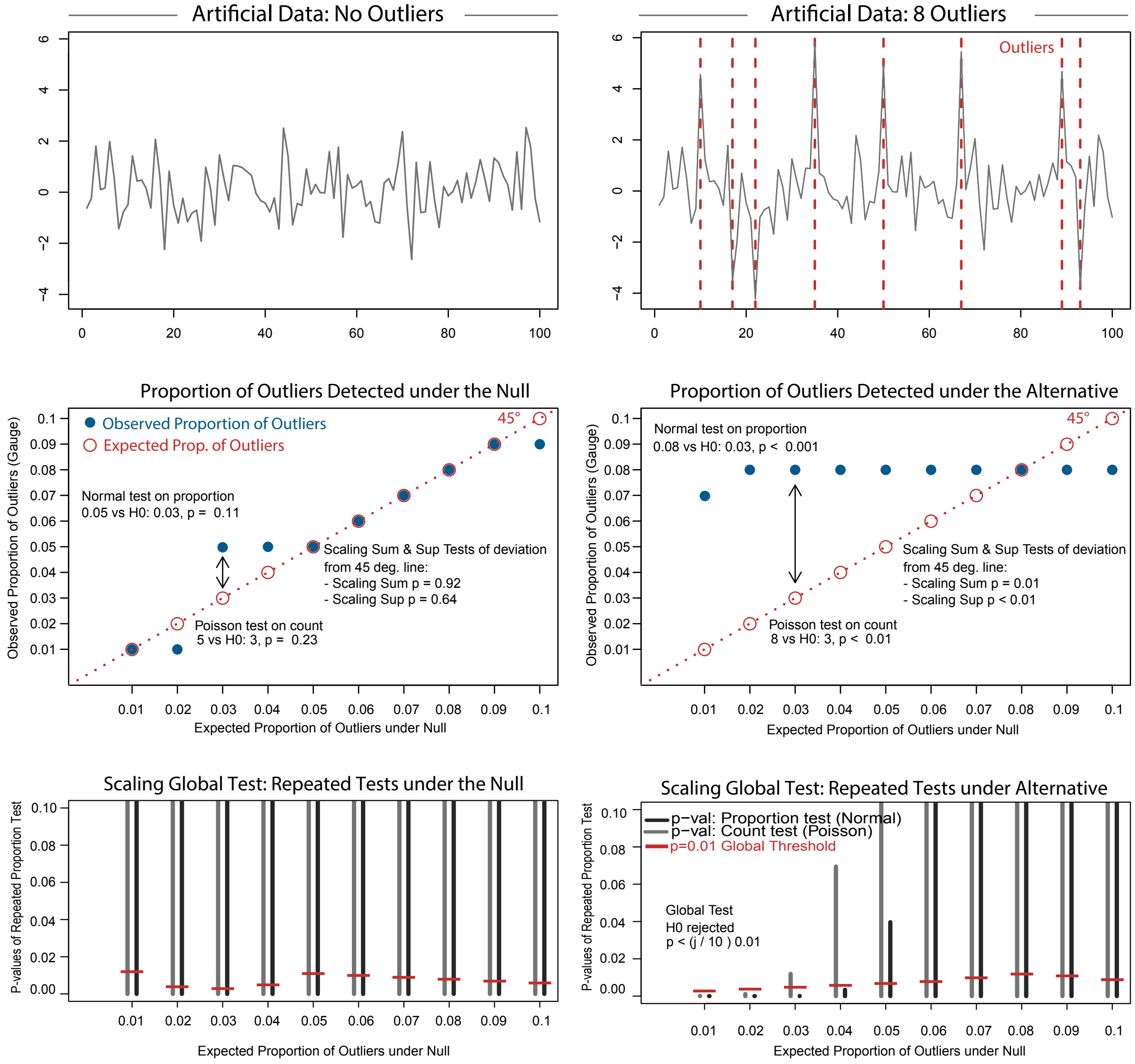
Yet another alternative approach taking outlier magnitude into account would be to assess the probability of observing the impulse t-values under the null-hypothesis of no outliers (as suggested by Johansen & Hendry 2015 for general variable selection), or combining both the proportion and magnitude into a single test statistic capturing the average outlier magnitude. This could be done by bias-correcting the t-statistics of the individual impulses using Hendry & Krolzig (2005) to fill-in the truncated distributions, and subsequently testing the average of bias-corrected t-statistic across all impulses. Such extensions will be the subject of future research.

2.6 Implementation and Use of the Tests

We provide an illustration of the proposed proportion, count, scaling sum, supremum, and global tests under the null of no outliers as well as under an alternative in Figure 1. Subsequently we investigate their performances in simulations in section 3 to provide guidance on which test to apply in practise. The proportion and count tests assess the difference between expected (red empty circles) and observed (blue solid points) proportions of outliers for a single level of selection significance in RLS or IIS – the results shown for a test when 3 outliers are expected as marked by the arrow. The sum and supremum scaling tests assess the entire deviation of the observed (blue solid) from the expected (red hollow, along 45 the degree line) proportion of outliers for varying levels of selection significance γ_c. The global scaling test (bottom panel) compares the p-values of multiple individual hypothesis tests (bottom panel) using either the count (light gray) or proportion (dark gray) tests repeatedly against the threshold given by the correction for repeated testing (red lines). The scaling tests have power where the proportion or count tests do not, as can be seen when for example testing for a deviation from 8 outliers under the alternative (right) where observed $\tilde{\gamma}_c = 0.08$ equals the expected value under the null $\gamma_c = 0.08$ despite the presence of outliers.

The proportion and count outlier tests require a single run of RLS or IIS, while the scaling and global tests require multiple (K) runs of RLS or IIS which can be computationally intensive in large samples. IIS can be applied using the algorithm *Autometrics* (Doornik 2009) in PcGive (Doornik & Hendry 2013) or using the `isat` function in the R-package *gets* (Pretis, Reade, & Sucarrat 2017). Code to implement the gauge test and compute the approximate variance is available from the authors compatible with the R-package *gets*. The most direct use of the test is to assess the presence of outlying observations in regression models occurring for any number of reasons: from measurement errors, non-normality, to model misspecification. Alternatively one can consider y_i as a series of forecast errors (difference between predicted and actual observations) and thus the tests lend themselves to assess forecast misspecification: if forecasts (e.g. from economic dynamic stochastic general equilibrium models, or general circulation models in climate) track observations well, then only a small number of forecast errors will be classed as outliers, while if the forecasts fail, a large number of outliers will be detected (see Ericsson 2017 and Pretis et al. 2015).

Figure 1: Illustrating the Proportion, Count, and Scaling (Sum, Sup., and Global) tests when there are no outliers (left) and in the presence of 8 outliers (right) in artificial data. Top panels show artificial data, middle panels show the detected proportion of outliers using IIS plotted against the expected proportion of outliers for the different levels of significance of selection. The bottom panels show the observed p-values of repeated tests together with the corresponding thresholds (red bars) for the scaling global test.



3 Simulation Study

Here we consider the performance of the proportion test (2.14), the count test (2.16), as well as the scaling sum (2.18), supremum (2.19), and global, (2.21) tests in a set of Monte Carlo simulations under the null hypothesis of no outliers and under different alternatives. All simulations are coded using the indicator saturation implementation `isat` in the R-package *gets* (Pretis, Reade, & Sucarrat 2017) with 4000 replications for the proportion and count tests, and 1000 replications for the scaling tests. Note that relative to the simulation study of Croux & Wilms (2016) and previous simulation assessments of IIS (e.g. Marczak & Proietti 2016) and step-indicator saturation (Castle et al. 2015), here we assess the properties of testing the presence of outliers on the proportion and number of outliers itself, rather than assessing what fraction of outliers are recovered in simulations (often referred to as the potency in previous simulations). Thus, the proposed tests here are more classical tests in the sense of them being evaluated by studying their size and power.

3.1 Under the Null of No Outliers

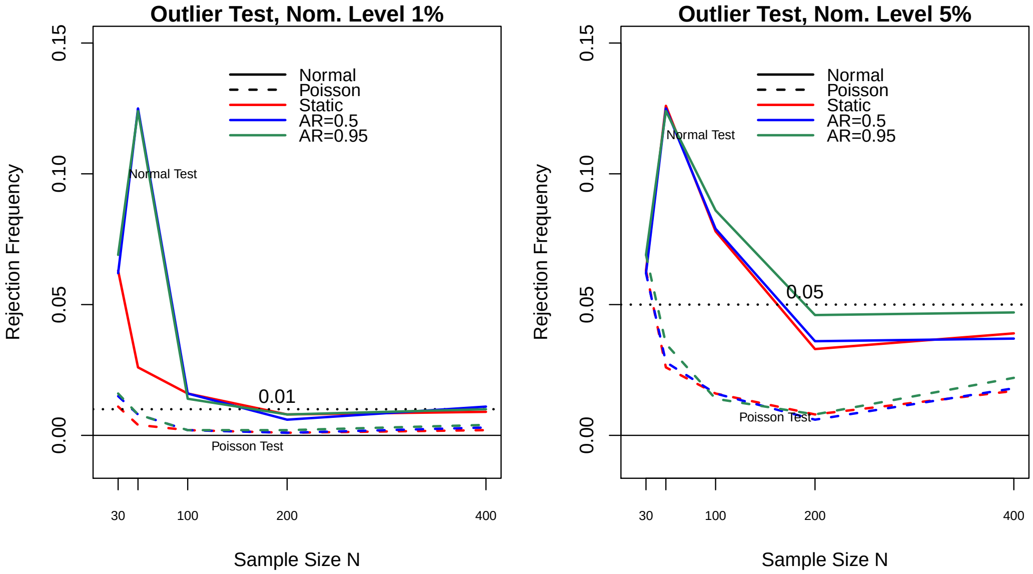
To study the tests under the null hypothesis, the simulated IIS model nests the DGP and there are no outlying observations. For the proportion and count tests different sample lengths n, specifications for x_i, and levels of significance of selection γ_c are explored. For the scaling tests we additionally consider different levels K of selection significance where γ_{c_k} are chosen such that the expected number of outliers equals distinct integers $\lambda_{\gamma_k} = 1, 2, ..., K$. The global scaling test is assessed using both the proportion and count tests repeatedly. The DGPs considered are static white noise (equivalent to a cross-section, $x_i = 1 \forall i$), as well as first-order autoregressive (AR) $x_i = y_{i-1}$ with coefficients $\beta = 0.5$ and $\beta = 0.95$.

Proportion and Count Tests under the Null Simulation results of the proportion and count tests under the null are shown in Figure 2 (as well as Tables 3, 4, and 5). As the Tables show – consistent with theory – the estimated proportion of outliers in IIS approaches the theory gauge γ_c as the sample size increases and the theory standard deviation (2.15) matches the simulation standard deviation closely for large samples. While the standard normal test on the proportion is oversized for small samples due to both a higher than expected gauge and variance, the size is close to the nominal level for larger samples (approx. $n \geq 100$) for both static and dynamic DGPs. The small-sample effect is particularly pronounced for low levels of significance of selection γ_c in IIS. This is intuitive, as in small samples the gauge can only take a small range of possible values. The Poisson approximation yields a size close to the nominal level of the test for small sample sizes, however, appears under-sized for large samples. The suggestion is thus to rely on the count test in small samples, and the proportion test in large samples to yield a stable size close to the nominal level.

Scaling Tests under the Null Simulation results of the scaling sum, supremum, and global tests under the null are shown in Figures 3 and 4 (as well as Tables 6, 7, 8, and 9). The size of the scaling tests closely matches the nominal level for larger sample sizes for both testing at $K = 10$, $K = 15$, and $K = 20$, for static and dynamic models. As expected, the scaling tests over-reject in small samples (as for example $K = 10$

implies very loose significance levels of selection in small samples)[3]. The global test exhibits a size close to the nominal level when based on repeated testing using the count test in small samples, and the proportion test in large samples.

Figure 2: Size: Null rejection frequency of the proportion (standard normal, solid lines) and count (Poisson, dashed lines) IIS outlier test under the null hypothesis of well-specified models for varying sample sizes n at nominal test level of 0.01 (left) and 0.05 (right) for static (red) and dynamic DGPs (blue, green). Selection in IIS was conducted at $\gamma = 0.01$. Full results are shown in Tables 3, 4, and 5.



3.2 Power under Alternatives

To assess the power of rejecting the null hypothesis of no outliers when there are actually outliers in the DGP,[4] the simulation DGPs are specified with varying proportions and magnitudes of (randomly placed with opposite signs) additive[5] outliers. The percentage of outlying observations is increased from 1% to 25% of the sample, with outlier magnitudes of $\delta/\sigma_\epsilon = 2, 3, 4, 6$.

[3] , E.g. for $T = 50$, $K = 10$ implies $\gamma = 0.1$ when λ_{γ_k} are chosen to be distinct integers.

[4] This is notably different to studying the proportion of outliers correctly identified – referred to as potency in previous simulation studies.

[5] IIS can detect both innovational and additive outliers in time series – the first by a single indicator, the latter by multiple subsequent indicators with opposing signs. In a setting with observations independent over i (e.g. cross-section), additive and innovational outliers are identical and can be matched by a single indicator. For a discussion on additive relative to innovational outliers see Nielsen (2004).

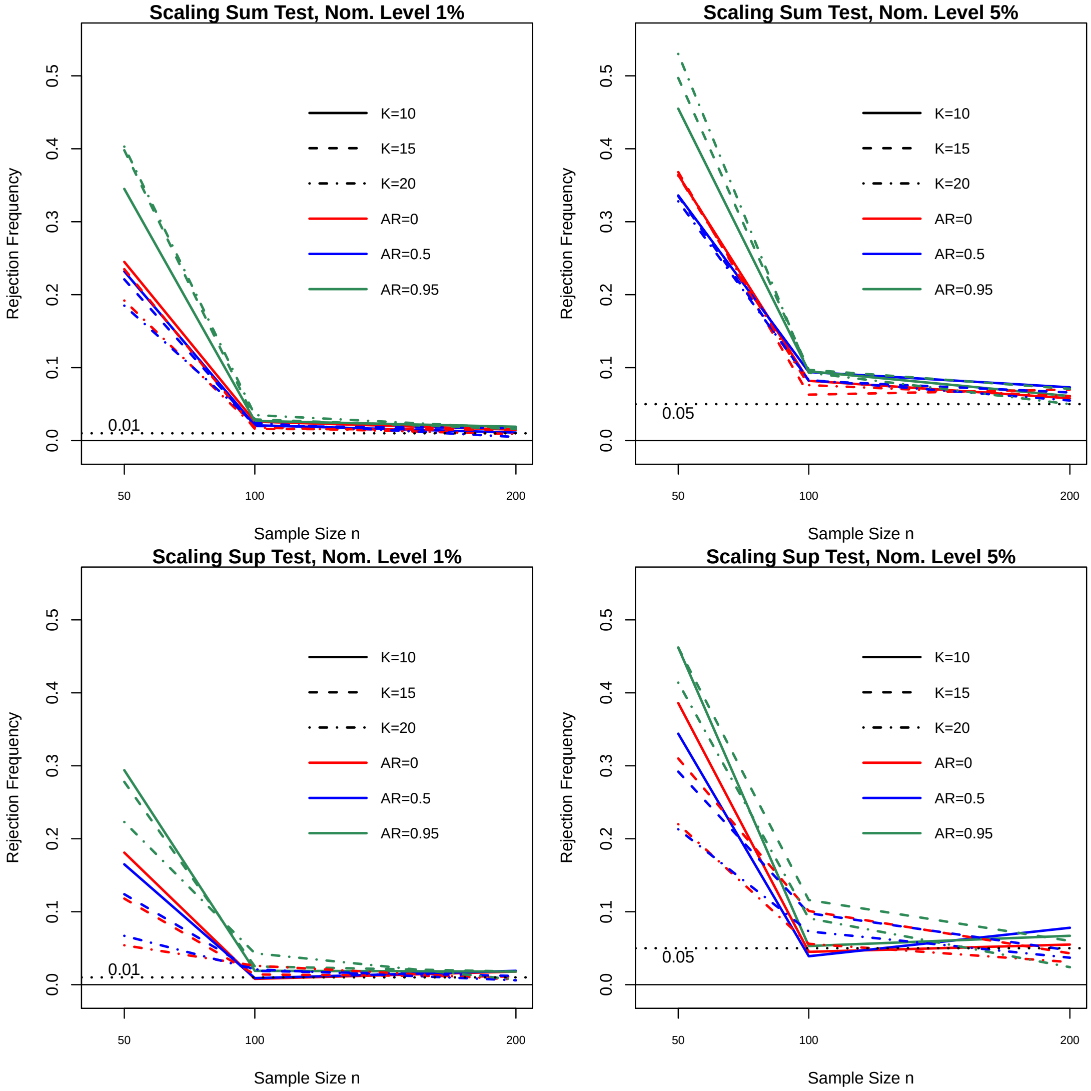
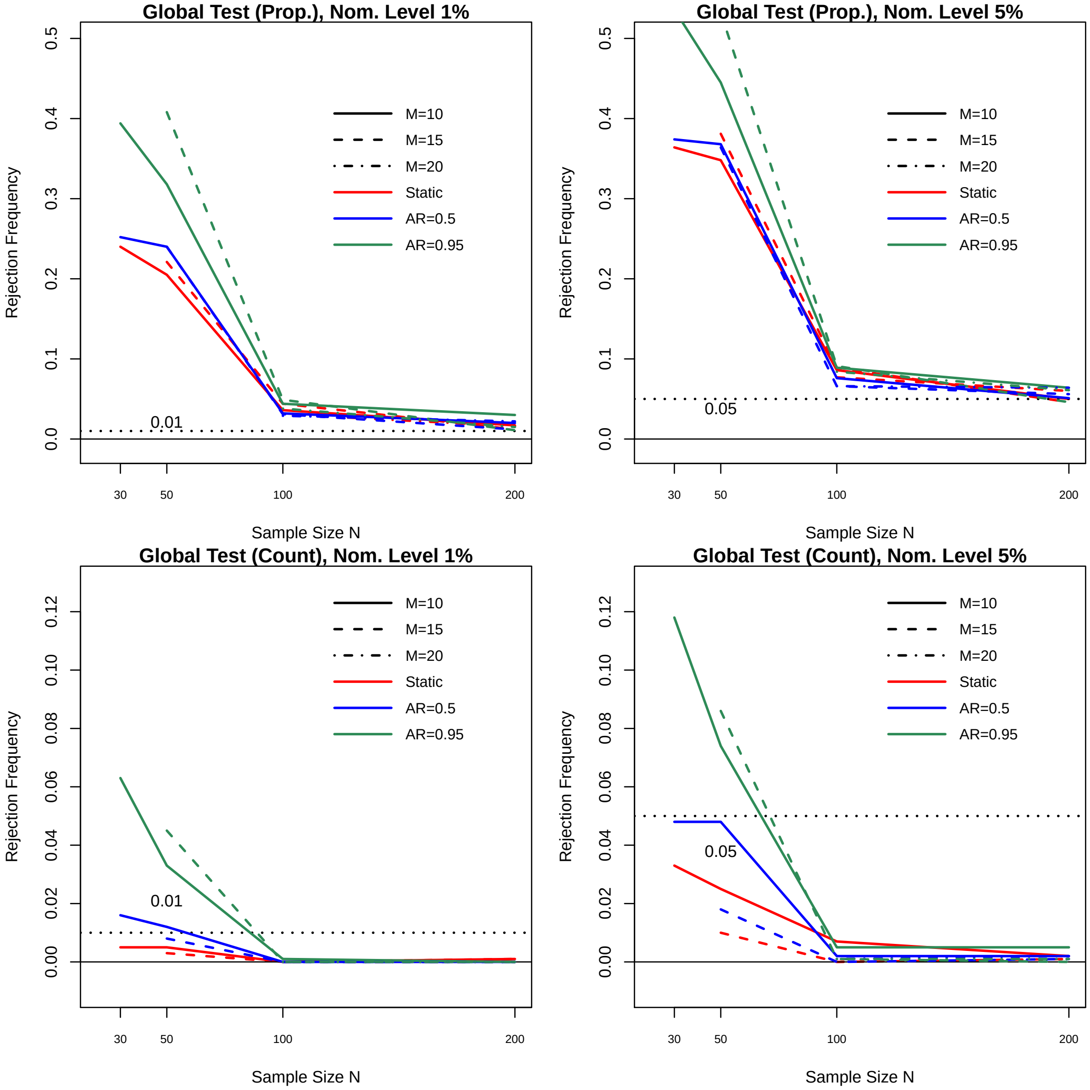


Figure 3: Size: Null rejection frequency of the scaling sum and sup tests under the null hypothesis of no outliers for scale levels $K = (10, 15, 20)$ for static and dynamic models AR $\beta = (0, 0.5, 0.95)$ with error variance 1 for sample size $n = (50, 100, 200)$ for nominal test levels $p = (0.01, 0.05)$ with 1000 replications.

Figure 4: Size: Null rejection frequency of the scaling global test using repeated proportion tests (top) and count tests (bottom) with $K = 10$ (solid) $K = 15$ (dashed), and $K = 20$ (dot-dashed), different significance levels in IIS (corresponding to expected outliers of $1, 2..., K$) under the null hypothesis of well-specified models for varying sample sizes n at nominal test level of 0.01 (left) and 0.05 (right) for static (red) and dynamic DGPs (blue, green).



Proportion and Count Tests under Alternatives Simulation results of the proportion and count tests under the alternative are shown in Figure 5 (as well as Tables 10, 11, 12). The power of the proportion and count tests increases with the sample size n. As expected, the standard normal proportion test exhibits higher power than the count test (though note that the power results are not size-corrected). The tests also exhibit increasing power as the proportion of outliers increases, however, once the proportion of outliers becomes very large (here simulated as 25% of the sample), the null-rejection frequency decreases as the high number of outliers can be mis-interpreted as a higher variance, thus leading to low retention of impulses.

Scaling and Global Tests under Alternatives Simulation results of the scaling tests under the alternative are shown in Figures 6 for the sum and supremum scaling tests, and in Figure 7 for the scaling global test. The results are also given in Tables 13, 14 and 15 for the scaling sum test, in Tables 16, 17, and 18 for the scaling supremum test, in Tables 19, 20, and 21 for the scaling global test using repeated proportion testing, and using repeated count tests in Tables 22, 23, and 24. The scaling tests exhibits good power even for a small number of outliers (5% of the sample), and their power increases in the magnitude of outliers. The power of the scaling sum and supremum tests increases in both the outlier magnitude as well as proportion. The global test using repeated proportion testing performs better than the global test using repeated count testing, consistent with the latter being under-sized in larger samples under the null.

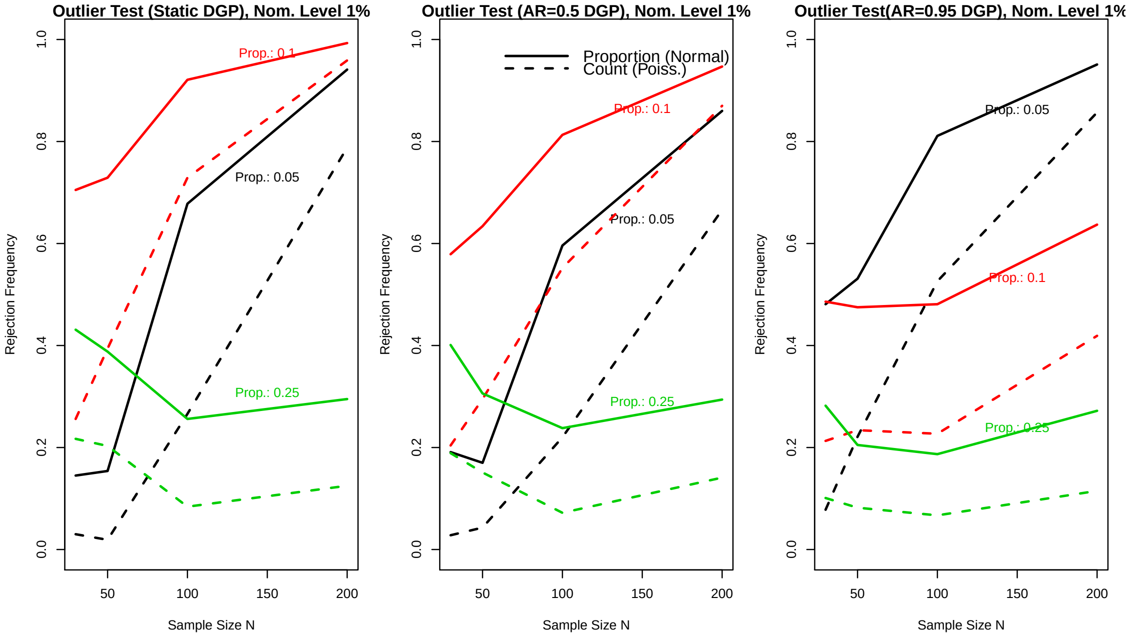
In summary, simulations show that the standard normal test of the proportion of outliers performs well in large ($n \geq 100$) samples with size close to the nominal level and desirable power properties. The count test against a Poisson distribution yields a size close to the nominal level in small samples, however is under-sized (and subsequently exhibits low power) in large samples. The scaling tests are over-sized in small samples ($n \approx 50$), but exhibit size close to the nominal level for larger samples ($n \geq 100$), and show good power against both the magnitude and number of outliers. The scaling sum test shows higher power in practise than the scaling test based on the supremum. There is little difference between the scaling global and sum tests in terms of power. The global test exhibits a size close to nominal levels when using the count test in small samples, and the proportion test in large samples. When taking outlier magnitude into account and handling larger samples, the simulations therefore suggest to rely on the sum or global scaling tests (based on repeated testing using the proportion test).

4 Applications

4.1 Economic Growth in a Cross-Sectional Model

In a cross-sectional analysis we re-visit the study by Mankiw, Romer, & Weil (1992) (henceforth MRW) by estimating their augmented cross-country Solow growth model using IIS to assess model specification through the detection of outliers when modelling GDP per capita as a function of the investment share, growth of effective labour, and human capital. We first test their augmented Solow model for mis-specification by assessing the proportion and number of detected outliers, second we apply the scaling and global tests to also take outlier magnitude into account.

Figure 5: Power: Null rejection frequency of the proportion and count test for varying sample sizes n and proportion of outliers in the sample when testing at nominal level of 0.01 (left) and 0.05 (right) for static (top) and dynamic (bottom) DGPs and models with outlier magnitude $\delta/\sigma_\epsilon = 4$. Selection in IIS was conducted at $\gamma_c = 0.01$. Full results are shown in Tables 10, 11, and 12.



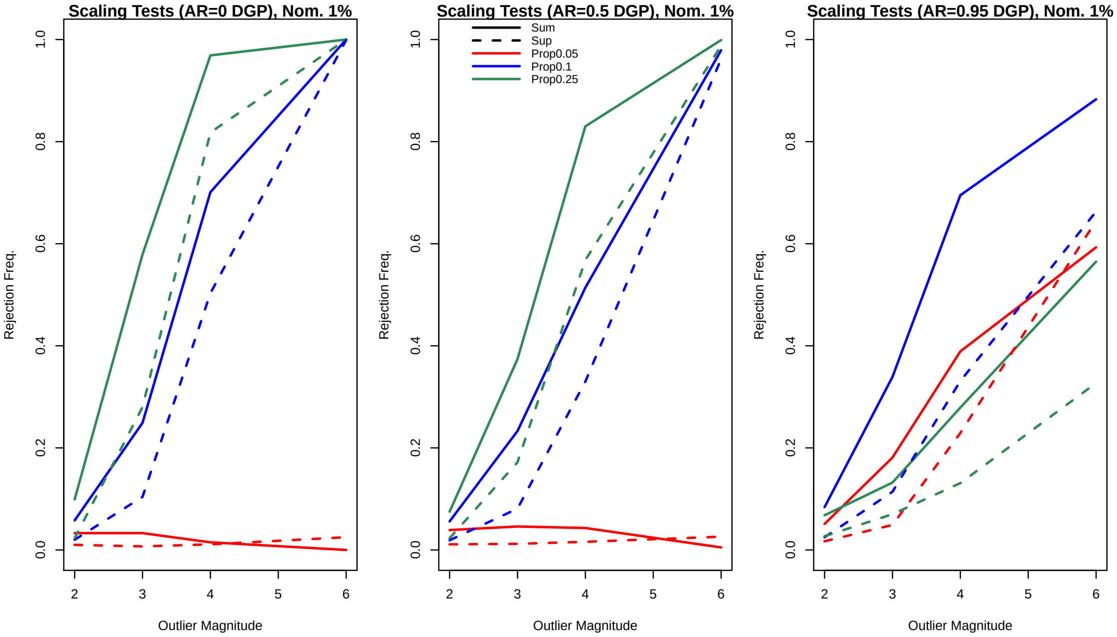
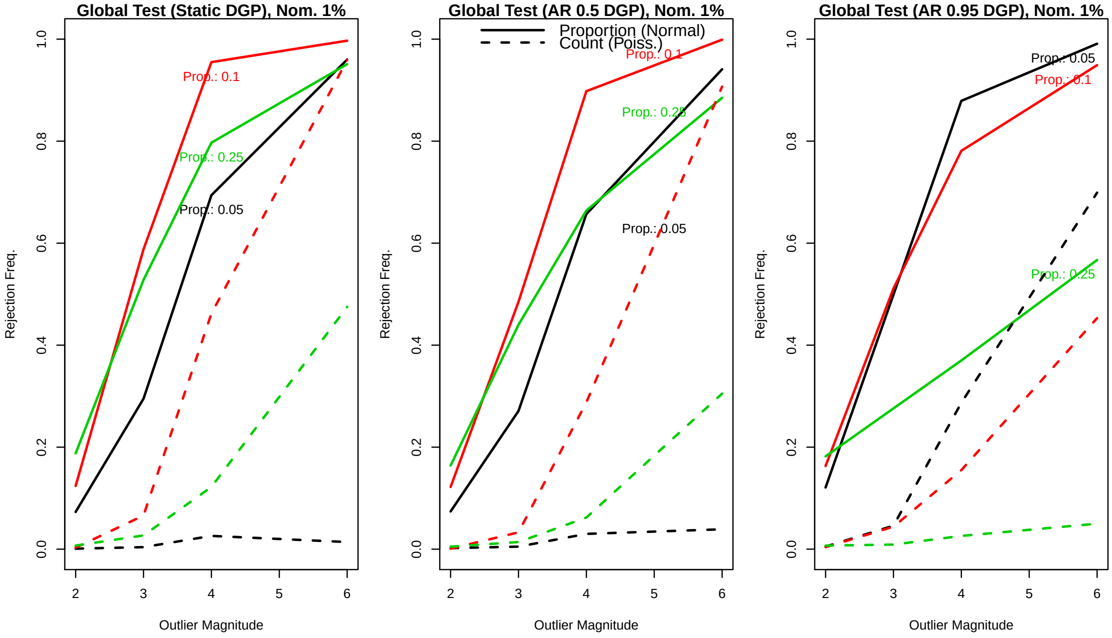


Figure 6: Power: Null rejection frequency of scale sum and sup tests under the alternative hypothesis for scale levels $K = 10$ for static and dynamic models AR $\beta = (0, 0.5, 0.95)$ with error variance 1 for sample size $n = 100$ for proportion of outliers $(0.05, 0.10, 0.25)$ and magnitude of outliers $(2, 3, 4, 6)$ for nominal test levels $p = 0.01$ with replications 1000.

Figure 7: Power: Null rejection frequency of the global test using repeated proportion tests (solid) and count tests (dashed)for increasing outlier magnitude δ/σ_ϵ and proportion of outliers in the sample when testing at nominal level of 0.01 (left) for static (left) and dynamic (middle, right) DGPs. Selection in IIS was conducted at $K = 10$ levels of significance of selection in IIS in a sample size of $n = 100$.



Testing on the number of outliers (*Proportion* and *Count* Tests) in MRW Applying IIS at a significance level of selection of $\gamma_c = 0.05$ results in seven outlying countries[6] denoted as I_{Country} in equation (4.1) identified in the sample of countries excluding oil producers ($n = 98$):[7]

$$\begin{aligned} \text{log(GDPpc)}_i \quad = \quad & -\underset{(0.91)}{3.82} + \underset{(0.12)}{0.60}\text{Inv}_i - \underset{(0.36)}{1.79}ngd_i + \underset{(0.06)}{0.67}\text{School}_i - \underset{(0.44)}{1.26}I_{\text{Ghana}} - \underset{(0.43)}{0.91}I_{\text{Togo}} \\ & -\underset{(0.44)}{1.42}I_{\text{Zaire}} - \underset{(0.44)}{0.95}I_{\text{Zambia}} - \underset{(0.44)}{1.01}I_{\text{Burma}} + \underset{(0.44)}{1.06}I_{\text{HK}} - \underset{(0.43)}{1.05}I_{\text{India}} \end{aligned} \tag{4.1}$$

This corresponds to an estimated outlier proportion of $\tilde{\gamma}_c = 7/98 = 0.071$. Testing whether the observed proportion (0.071) significantly differs from the expected proportion ($\gamma_c = 0.05$), the corresponding test statistic is given by $z_\gamma = (0.071 - 0.05)/\sqrt{\omega_\gamma} = 0.019/0.0144 = 1.46$. A two-tailed comparison against the standard normal distribution yields a p-value of 0.24, suggesting no rejecting of the null-hypothesis of the proportion of outlying observations differing from the expected proportion. Testing whether the number of observed outliers differs from the expected number using the Poisson count test equally does not lead to a rejection of the null hypothesis: a two-sided Poisson test of $\tilde{\gamma}_c n = 7$ against the expected number $\gamma_c n = 4.9$ yields a p-value of $p = 0.36$.

The coefficients on investment (Inv.), effective labour growth (ngd), and human capital (School) are not statistically different to the original estimates of MRW (with t-values for tests of difference of -0.25, -0.17, and -0.17 respectively) further confirming that applying IIS on a well specified model has little effect on other included covariates.[8] The detected outliers in the non-oil producing sample overlap in parts with the outyling countries in the robust estimation of the MRW model in Temple (1998), however, the tests presented here suggest that the number of outliers detected do not differ significantly from their expected value under the null hypothesis of no outliers.[9]

Testing both number and magnitude of outliers through scaling sum, supremum, and global tests in MRW The above tests based on the estimates of $\tilde{\gamma}_c$ and $\tilde{\gamma}_c n$ use a single level of significance of selection γ_c in IIS and have power only against the number of outliers. Here we apply the set of scaling tests (sum, supremum, global) to take outlier magnitude into account by running IIS repeatedly using multiple levels of selection significance γ_{c_k} and recording $\tilde{\gamma}_{c_k}$ at each run. Table 1 and Figure 8 show the number of outliers detected together with the corresponding $K = 10$ different values of selection significance γ_{c_k}, chosen such that the expected number of outliers are distinct integers ($\gamma_{c_k} n = 1, 2, ..., 10$). The observed proportion of outliers $\tilde{\gamma}_{c_k}$ is compared to the expected number γ_{c_k} using the sum, supremum, and global tests (see Figure 8).

To apply the scaling sum test, we compute $S_{sum} = 1.91$ using equation (2.17) and compare it agains the

[6] These are: Ghana, Togo, Zaire, Zambia, Burma, Hong Kong, and India. Selection at $\gamma = 0.01$ results in one outlier, Zaire, being detected.

[7] The variable Inv_i refers to investment relative to GDP_i, ngd is the growth in effective units in labour plus depreciation, and School_i refers to the percentage of working age population in secondary schooling – see MRW 1992 for more details.

[8] When both an efficiency correction on the standard errors and a consistency correction for the estimate of σ_ϵ are applied – see Johansen & Nielsen (2009) and Johansen and Nielsen (2016b).

[9] Temple (1998) considers robust estimation of multiple sub-samples beyond the example considered here, and finds additional outliers in these samples.

critical values of the normal distribution with the variance given by $\mathsf{Var}_{\mathsf{c_k}}$ in equation (2.18). This yields a p-value of $p = 0.046$, suggesting no rejection of the null hypothesis of no outliers when testing at the 1% level. For the scaling supremum test we take the supremum using (2.19) to construct $S_{sup} = 0.40$ which is assessed against the critical values obtained by simulating the Gaussian process in equation (2.20), which yields a p-value of 0.06. Finally, applying the global scaling test, for each γ_{c_k} we conduct a proportion outlier test, the recorded p-values of these tests are shown in Table 1 and Figure 8 (bottom panel). These are ordered in ascending order and compared to the corresponding threshold given by (2.21) (red bars in Figure 8 and bottom row in Table 1). All observed p-values exceed their corresponding threshold value, thus the global null hypothesis of no outliers is not rejected.

All scaling tests (sum, supremum, and global) tests show that the number of detected outliers scales proportionally to the level of significance of selection consistent with the null hypothesis of no outliers and the model not being misspecified. Overall, tests on both the number and magnitude of detected outliers suggest that the model is well specified and outlying countries are found consistent with expected values observed by chance under the null.

Figure 8: Scaling tests on the Mankiw-Romer-Weil model of economic growth. Top panel shows observed (blue solid) and expected (red empty) number of outliers in the MRW augmented Solow Model estimated using IIS at $K = 10$ selection significance levels (see Table 1 for full results). Under the null of no outliers the observed outliers are expected to fall on the 45 degree line (red). Using the IIS scaling sum and sup tests the null hypothesis of no outliers is not rejected (p=0.045 for sum, and p=0.06 for the supremum tests). Bottom panel shows the p-values from repeated proportion tests (gray) together with the global test thresholds at 1% (red bars). All observed p-values exceed their respective thresholds and the global null hypothesis of no outliers is not rejected.

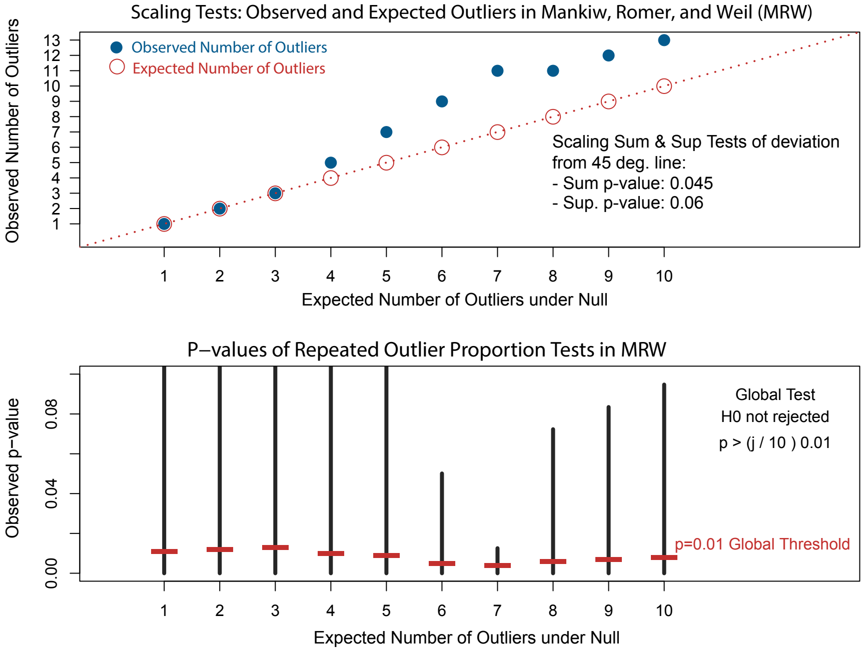


Table 1: Detected ($\tilde{\gamma}_{c_k}$) and expected (γ_{c_k}) proportion of outliers in the MRW augmented Solow growth model for $K = 10$ levels of significance γ_{c_k} used to construct the scaling tests. Lower rows show p-values of repeated proportion test and the corresponding threshold of the global test. All proportion test values exceed their global thresholds and the null hypothesis of no outliers is not rejected.

Lev. of Sign.: γ_{c_k}	0.0102	0.0204	0.0306	0.0408	0.051	0.0612	0.0714	0.0816	0.0918	0.102
Detected Prop.: $\tilde{\gamma}_{c_k}$	0.0102	0.0204	0.0306	0.051	0.071	0.091	0.112	0.112	0.122	0.133
Expected Num.: $\gamma_{c_k} n$	1	2	3	4	5	6	7	8	9	10
Detected Num: $\tilde{\gamma}_{c_k} n$	1	2	3	5	7	9	11	11	12	13
Scal. Sum: 1.91 ($p \approx 0.046$)										
Scal. Sup.: 0.40 ($p \approx 0.06$)										
Prop. Test p-value $\tilde{p}^g$:	1.000	1.000	1.000	0.462	0.168	0.050	0.013	0.072	0.083	0.095
Global threshold (1%)	0.008	0.009	0.01	0.007	0.006	0.002	0.001	0.003	0.004	0.005
$\tilde{p}^g \leq$ threshold:	x	x	x	x	x	x	x	x	x	x

4.2 Re-visiting existing studies using Impulse Indicator Saturation

As the proportion and count tests only require the detected number of outliers, sample size, and the nominal significance level of selection γ_c, both tests can readily be applied to existing studies already using indicator saturation without requiring re-estimation.[10] For example, building on Martinez (2015), Ericsson (2017) uses IIS to assess systematic forecast bias in forecasts of US government debt. In a sample of $n = 29$ forecast error observations applying IIS at $\gamma_c = 0.01$, Ericsson (2017) detects 4 ($\tilde{\gamma}_c = 4/29 = 0.14$), 5 ($\tilde{\gamma}_c = 5/29 = 0.17$), and 7 ($\tilde{\gamma}_c = 7/29 = 0.24$) outlying observations in the government forecasts of three different agencies respectively. Applying the proportion test results in test statistics of 8.16 ($p \approx 0$), 10.36 ($p \approx 0$), and 14.76 ($p \approx 0$) respectively. Simulations in Section 3 suggest, however, that the standard normal test over-rejects in small samples. Thus, to address concerns about the small number of forecast observations ($n = 29$) the count test is also considered. Under the null hypothesis the expected rate is equal to $\gamma_c n = 0.01 \times 29 = 0.29$. A two-sided test against a $\mathsf{Poisson}(\gamma_c n = 0.29)$ distribution given the observed rates of $\tilde{\gamma}_c n = (4, 5, 7)$ yields the p-values of $p = 0.0002, p \approx 0$, and $p \approx 0$ respectively. Thus, both the standard normal proportion as well as the Poisson count test reject that the observed proportion (number) of outliers equals the expected proportion (number).

Expanding on the example of re-visiting Ericsson's (2017) results, we apply the outlier count and proportion tests to a set of 12 published studies which used IIS in a range of applications: from modelling UK wages to forecasts of food prices and the economic impacts of climate change. Table 2 provides an overview of the studies together with the results of applying the proportion and count tests. Overall, interpreted model misspecification by the authors closely coincides with the null rejection of the outlier test.

[10] The scaling tests require estimation of the models at multiple levels of significance, therefore requires re-estimation and is thus less straightforward to apply to published studies.

Table 2: The normal and Poisson outlier tests applied to 12 published studies using IIS: $\tilde{\gamma}$ shows the estimated proportion of outliers as the number $\tilde{\gamma}_c n$ of impulse indicators retained over the number of observations, γ_c the nominal significance level of selection, and "Prop." and "Count" show the p-values of applying the IIS *proportion* and *count* tests to the author's results.

Authors	**IIS Application**	$\tilde{\gamma}_c$	γ_c	***Prop.***	***Count***
Castle & Hendry (2009)	UK Wage Equation	6/141=0.043	0.001	p≈0	p≈0
Hendry & Mizon (2011b)	US Food expenditure (results from dynamic model)	7/72=0.097	0.01	p≈0	p≈0
Hendry (2011)	UK Consumer Expenditure	3/65=0.046	0.025	p=0.15	p=0.22
Castle & Hendry (2012)	US Returns to Education	301/5173=0.06	0.001	p≈0	p≈0
Hendry & Pretis (2013)	Atmospheric CO_2 Concentrations	1/246=0.004	0.001	p=0.12	p=0.21
Dreger & Wolters (2014)	Euro Area Money demand equation	6/112=0.054	0.05	p=0.79	p=0.83
Anundsen (2015)	US Rent Prices	4/140=0.028	0.01	p=0.01	p=0.054
Nymoen & Sparrman (2015)	Unemployment Rate in Panel of OECD Countries	157/994=0.158	0.025	p≈0	p≈0
Stillwagon (2016)	US/GBP Exchange Rate	2/373=0.005	0.001	p=0.006	p=0.05
Ahumada & Cornejo (2016)	Food price forecasts (reported for soybean prices, 1-step fixed sample)	4/28 = 0.143	0.01	p≈0	p≈0
Martinez (2017)	Cross-sectional model of hurricane damages.	7/98 = 0.07	0.01	p≈0	p≈0
Pretis, Schwarz, et al. (2017)	Panel estimation of climate impacts on economic growth.	81/7007 = 0.012	0.001	p≈0	p≈0

5 Conclusion

The proportion and numbers of outliers detected in regression models can be used to test model misspecification. We propose two sets of testing procedures for the null hypothesis of no outliers using robustified least squares and impulse indicator saturation algorithms. First, a set of tests on the proportion and count of outliers, and second, a set of scaling tests. The proposed tests on the proportion and count have power against the number of outliers, while the scaling tests have power against both the number and magnitude of outliers. The tests are valid for both stationary or stochastically trending regressors. It is worth highlighting that rejection of the null hypothesis of the observed outlier proportion equalling the expected proportion is indicative of model misspecification, but non-rejection does not rule out that individual outliers (even in a small proportion) denote actual outlying observations, as the outlier magnitude is not considered when testing on the proportion alone. The outlier magnitude can be taken into account through the proposed scaling tests by assessing the number of outliers in indicator saturation or robustified least squares at varying levels of the nominal level of significance used to detect outliers. The tests perform well with size close to nominal levels and high power against both magnitude and number of outliers. The testing procedure are readily available through *Autometrics* in PcGive and *gets* in R, and can be applied in a wide range of settings: from evaluating model misspecification, to the accuracy of forecasts in simulated models, to assessing the presence of outliers in time series, panel, or cross sectional models.

References

Adler, R. J., & Taylor, J. E. (2009). *Random fields and geometry*. Springer Science & Business Media.

Ahumada, H., & Cornejo, M. (2016). Forecasting food prices: The case of corn, soybeans and wheat. *International Journal of Forecasting*, *32*(3), 838–848.

Anundsen, A. K. (2015). Econometric regime shifts and the us subprime bubble. *Journal of Applied Econometrics*, *30*(1), 145–169.

Benjamini, Y., & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, 1165–1188.

Bickel, P. J. (1975). One-step huber estimates in the linear model. *Journal of the American Statistical Association*, *70*(350), 428–434.

Billingsley, P. (2013). *Convergence of probability measures*. John Wiley & Sons.

Castle, J. L., Doornik, J. A., Hendry, D. F., & Pretis, F. (2015). Detecting locations shifts by step-indicator saturation during model selection. *Econometrics*, *3*, 240-264.

Castle, J. L., & Hendry, D. F. (2009). The long-run determinants of UK wages, 1860–2004. *Journal of Macroeconomics*, *31*(1), 5–28.

Castle, J. L., & Hendry, D. F. (2012). Automatic selection for non-linear models. In *System identification, environmental modelling, and control system design* (pp. 229–250). Springer.

Castle, J. L., & Shephard, N. (Eds.). (2009). *The methodology and practice of econometrics*. Oxford: Oxford University Press.

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM computing surveys (CSUR)*, *41*(3), 15.

Croux, C., & Wilms, I. (2016). Discussion of asymptotic theory of outlier detection algorithms for linear time series regression models. *Scandinavian Journal of Statistics*, *43*(2), 353–356.

Doornik, J. A. (2009). Autometrics. In J. L. Castle & N. Shephard (Eds.), (pp. 88–121). Oxford: Oxford University Press.

Doornik, J. A., & Hendry, D. F. (2013). PcGive 14. *London: Timberlake Consultants*.

Doornik, J. A., & Hendry, D. F. (2016). Outliers and model selection: Discussion of the paper by Søren Johansen and Bent Nielsen. *Scandinavian Journal of Statistics*, *43*(2), 360–365.

Dreger, C., & Wolters, J. (2014). Money demand and the role of monetary indicators in forecasting euro area inflation. *International Journal of Forecasting*, *30*(2), 303–312.

Ericsson, N. R. (2017). How biased are us government forecasts of the Federal debt? *International Journal of Forecasting*.

Grubbs, F. E. (1969). Procedures for detecting outlying observations in samples. *Technometrics*, *11*(1), 1–21.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., & Stahel, W. A. (1986). Linear models: robust estimation. *Robust Statistics: The Approach Based on Influence Functions*, 307–341.

Hendry, D. F. (2011). Revisiting uk consumers expenditure: cointegration, breaks and robust forecasts. *Applied Financial Economics*, *21*(1-2), 19–32.

Hendry, D. F., Johansen, S., & Santos, C. (2008). Automatic selection of indicators in a fully saturated regression. *Computational Statistics*, *23*, 337-339.

Hendry, D. F., & Krolzig, H.-M. (2005). The properties of automatic Gets modelling. *Economic Journal*, *115*, C32–C61.

Hendry, D. F., & Mizon, G. E. (2011a). Econometric modelling of time series with outlying observations. *Journal of Time Series Econometrics*, *3 (1)*, DOI: 10.2202/1941-1928.1100.

Hendry, D. F., & Mizon, G. E. (2011b). Econometric modelling of time series with outlying observations. *Journal of Time Series Econometrics*, *3*(1).

Hendry, D. F., & Pretis, F. (2013). Anthropogenic influences on atmospheric CO2. *Handbook on Energy and Climate Change*, 287.

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

Hodge, V. J., & Austin, J. (2004). A survey of outlier detection methodologies. *Artificial intelligence review*, *22*(2), 85–126.

Jiao, X., & Nielsen, B. (2017). Asymptotic analysis of iterated 1-step huber-skip m-estimators with varying cut-offs. In *Workshop on analytical methods in statistics* (pp. 23–52).

Johansen, S., & Hendry, D. (2015). Model discovery and trygve haavelmo's legacy. *Econometric Theory*.

Johansen, S., & Nielsen, B. (2009). An analysis of the indicator saturation estimator as a robust regression estimator. In J. L. Castle & N. Shephard (Eds.), (pp. 1–36). Oxford: Oxford University Press.

Johansen, S., & Nielsen, B. (2013). Asymptotic theory for iterated one-step huber-skip estimators. *Econometrics*.

Johansen, S., & Nielsen, B. (2016a). Analysis of the forward search using some new results for martingales and empirical processes. *Bernoulli*, *22*(2), 1131–1183.

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

Jurečková, J., Sen, P. K., & Picek, J. (2013). *Methodology in robust and nonparametric statistics*. CRC Press.

Mankiw, N. G., Romer, D., & Weil, D. N. (1992). A contribution to the empirics of economic growth. *The Quarterly Journal of Economics*, *107*(2), 407-437. doi: 10.2307/2118477

Marczak, M., & Proietti, T. (2016). Outlier detection in structural time series models: The indicator saturation approach. *International Journal of Forecasting*, *32*(1), 180–202.

Martinez, A. B. (2015). How good are us government forecasts of the federal debt? *International Journal of Forecasting*, *31*(2), 312–324.

Nielsen, H. B. (2004). Cointegration analysis in the presence of outliers. *Econometrics Journal*, *7*, 249–271.

Nymoen, R., & Sparrman, V. (2015). Equilibrium unemployment dynamics in a panel of oecd countries. *Oxford Bulletin of Economics and Statistics*, *77*(2), 164–190.

Prescott, P. (1975). An approximate test for outliers in linear models. *Technometrics*, *17*(1), 129–132.

Pretis, F., Mann, M. L., & Kaufmann, R. K. (2015). Testing competing models of the temperature hiatus: assessing the effects of conditioning variables and temporal uncertainties through sample-wide break detection. *Climatic Change*, *131*(4), 705–718.

Pretis, F., Reade, J., & Sucarrat, G. (2017). General-to-specific (gets) modelling and indicator saturation with the R package gets. *Journal of Statistical Software*, *in press*.

Pretis, F., Schwarz, M., Tang, K., Haustein, K., & Allen, M. (2017). Uncertain impacts on economic growth when stabilising temperatures at 1.5 or 2C warming. *Philosophical Transactions of the Royal Society, A*, *in press*.

Ruppert, D., & Carroll, R. J. (1980). Trimmed least squares estimation in the linear model. *Journal of the American Statistical Association*, *75*(372), 828–838.

Sarkar, S. K., & Chang, C.-K. (1997). The simes method for multiple hypothesis testing with positively dependent test statistics. *Journal of the American Statistical Association*, *92*(440), 1601–1608.

Seeger, P. (1968). A note on a method for the analysis of significances en masse. *Technometrics*, *10*(3), 586–593.

Simes, R. J. (1986). An improved bonferroni procedure for multiple tests of significance. *Biometrika*, 751–754.

Srivastava, M. S., & von Rosen, D. (1998). Outliers in multivariate regression models. *Journal of Multivariate Analysis*, *65*(2), 195–208.

Stillwagon, J. R. (2016). Non-linear exchange rate relationships: An automated model selection approach with indicator saturation. *The North American Journal of Economics and Finance*, *37*.

Temple, J. R. (1998). Robustness tests of the augmented solow model. *Journal of Applied Econometrics*, 361–375.

Tietjen, G. L., Moore, R., & Beckman, R. (1973). Testing for a single outlier in simple linear regression. *Technometrics*, *15*(4), 717–721.

Van Der Vaart, A. W., & Wellner, J. A. (1996). Weak convergence. In *Weak convergence and empirical processes* (pp. 16–28). Springer.

Welsh, A. H., & Ronchetti, E. (2002). A journey in single steps: robust one-step m-estimation in linear regression. *Journal of Statistical Planning and Inference*, *103*(1), 287–310.

6 Appendix

6.1 Appendix: Proofs

Here we provide the proofs of the Lemmas and theorems in the main text.

Proof of Lemma 2.1. The term of interest is $\mathcal{M}_n = n^{-1/2}\sum_{i=1}^n v_i^{a,b,c}$.

1. *Decompose* $\mathcal{M}_n$. Write $\mathcal{M}_n = \mathcal{M}_{n,1} + \mathcal{M}_{n,2} + \mathcal{M}_{n,3}$, where

$$\mathcal{M}_{n,1} = n^{-1/2}\sum_{i=1}^n 1_{(|\varepsilon_i|\le\sigma c)}, \quad \mathcal{M}_{n,2} = n^{-1/2}\sum_{i=1}^n \mathsf{E}_{i-1}\{v_i^{a,b,c} - 1_{(|\varepsilon_i|\le\sigma c)}\},$$
$$\mathcal{M}_{n,3} = n^{-1/2}\sum_{i=1}^n\{v_i^{a,b,c} - 1_{(|\varepsilon_i|\le\sigma c)}\} - n^{-1/2}\sum_{i=1}^n \mathsf{E}_{i-1}\{v_i^{a,b,c} - 1_{(|\varepsilon_i|\le\sigma c)}\}.$$

Therefore, the first term in stochastic expansion is $\mathcal{M}_{n,1}$. We will linearize $\mathcal{M}_{n,2}$ to obtain the second term, and argue that $\mathcal{M}_{n,3}$ is small in probability.

2. *Linearize* $\mathcal{M}_{n,2}$. Theorem 10 in Jiao and Nielsen (2017) by Assumption 2.1(ia, iic) shows $\mathcal{M}_{n,2} = \overline{\mathcal{M}}_{n,2} + \mathrm{O}_{\mathsf{P}}(n^{-2\eta})$, where $\overline{\mathcal{M}}_{n,2} = 2c\mathsf{f}(c)a/\sigma$. Note $0 < \eta \le 1/4$. Thus, we have $\mathcal{M}_{n,2} = \overline{\mathcal{M}}_{n,2} + \mathrm{o}_{\mathsf{P}}(1)$ uniformly in $0 < c < \infty$ and $|a|, |b| \le n^{1/4-\eta}B$.

3. *Bounding* $\mathcal{M}_{n,3}$. Due to Assumption 2.1(ia, iib, iic), Theorem 9 in Jiao and Nielsen (2017) shows $\mathcal{M}_{n,3} = \mathrm{o}_{\mathsf{P}}(1)$ uniformly in a, b, c. ∎

Proof of Theorem 2.1. The definition (2.9) of the sample gauge gives

$$n^{1/2}(\widehat{\gamma}_c^{(m)} - \gamma_c) = n^{-1/2}\sum_{i=1}^n\{(1 - v_{i,c}^{(m)}) - \gamma_c\}.$$

Then express the weight $v_{i,c}^{(m)}$ in (2.4) as

$$v_{i,c}^{(m)} = 1_{(|y_i - x_i'\widehat{\beta}_c^{(m)}|\le\widehat{\sigma}_c^{(m)}c)} = 1_{(|\varepsilon_i - x_{in}'\widehat{b}_c^{(m)}|\le\sigma c + n^{-1/2}\widehat{a}_c^{(m)}c)} = v_i^{\widehat{a}_c^{(m)},\widehat{b}_c^{(m)},c},$$

where $\widehat{b}_c^{(m)} = N^{-1}(\widehat{\beta}_c^{(m)} - \beta)$ and $\widehat{a}_c^{(m)} = n^{1/2}(\widehat{\sigma}_c^{(m)} - \sigma)$ are the m step estimation errors for β and σ. Since $|\widehat{b}_c^{(m)}| + |\widehat{a}_c^{(m)}| = \mathrm{O}_{\mathsf{P}}(1)$ and by Assumption 2.1(ia, ii), apply the asymptotic expansion in Lemma 2.1 to obtain for any $0 \le m < \infty$

$$n^{1/2}(\widehat{\gamma}_c^{(m)} - \gamma_c) = n^{-1/2}\sum_{i=1}^n\{1_{(|\varepsilon_i|>\sigma c)} - \gamma_c\} - 2c\mathsf{f}(c)n^{1/2}(\frac{\widehat{\sigma}_c^{(m)}}{\sigma} - 1) + R(\widehat{a}_c^{(m)}, \widehat{b}_c^{(m)}, c),$$

where the remainder term $R(a, b, c)$ vanishes in probability uniformly in $c \in \mathbb{R}_+$ and $|a|, |b| \le B$. ∎

Proof of Theorem 2.2. By Assumption 2.1(ia, ii, iii) with $\eta = 1/4$, Theorem 2 in Jiao and Nielsen (2017) shows

$$\sup_{0\le m<\infty}\ \sup_{c_0\le c<\infty} |N^{-1}(\widehat{\beta}_c^{(m)} - \beta)| + |n^{1/2}(\widehat{\sigma}_c^{(m)} - \sigma)| = \mathrm{O}_{\mathsf{P}}(1).$$

Thus apply Theorem 2.1 to get

$$n^{1/2}(\widehat{\gamma}_c^{(m)} - \gamma_c) = n^{-1/2}\sum_{i=1}^{n}\{1_{(|\varepsilon_i|>\sigma c)} - \gamma_c\} - 2c\mathsf{f}(c)n^{1/2}(\frac{\widehat{\sigma}_c^{(m)}}{\sigma} - 1) + \mathsf{o}_\mathsf{P}(1),$$

where, uniformly in $m \in [0,\infty)$ and $c \in [c_0,\infty)$, the first term is $\mathrm{O}_\mathsf{P}(1)$ due to the empirical process CLT, see Billingsley (1968, Theorem 16.4) or Van Der Vaart & Wellner 1996 (Theorem 2.12.1), while the second term is $\mathrm{O}_\mathsf{P}(1)$ since $\sup_{c_0\le c<\infty} c\mathsf{f}(c) < \infty$ by Assumption 2.1(ia) and tightness of $\widehat{\sigma}_c^{(m)}$. Therefore, the above term of interest is bounded in probability uniformly in m, c. ∎

Proof of Theorem 2.3. Since $\widehat{\beta}_c^{(0)} = \widetilde{\beta}$, $(\widehat{\sigma}_c^{(0)})^2 = \widetilde{\sigma}^2$ are full sample least squares and by moment conditions for regressors in Assumption 2.1(ii), the initial estimators satisfy $N^{-1}(\widehat{\beta}_c^{(0)} - \beta) = \mathrm{O}_\mathsf{P}(1)$, $n^{1/2}(\widehat{\sigma}_c^{(0)} - \sigma) = \mathrm{O}_\mathsf{P}(1)$ and

$$n^{1/2}(\widehat{\sigma}_c^{(0)} - \sigma) = n^{-1/2}(2\sigma)^{-1}\sum_{i=1}^{n}(\varepsilon_i^2 - \sigma^2) + \mathrm{O}_\mathsf{P}(n^{-1/2}).$$

Further by Assumption 2.1(ia, ii), the asymptotic expansion in Theorem 2.1 shows

$$n^{1/2}(\widehat{\gamma}_c^{(0)} - \gamma_c) = n^{-1/2}\sum_{i=1}^{n}\{1_{(|\varepsilon_i|>\sigma c)} - \gamma_c\} - c\mathsf{f}(c)n^{-1/2}\sum_{i=1}^{n}(\frac{\varepsilon_i^2}{\sigma^2} - 1) + \mathsf{o}_\mathsf{P}(1),$$

uniformly in $c \in [c_0,\infty)$. The finite dimensional convergence follows by CLT so for any $c \in [c_0,\infty)$ then

$$\mathbb{G}_n(c) = n^{1/2}(\widehat{\gamma}_c^{(0)} - \gamma_c) \overset{\mathsf{D}}{\to} \mathsf{N}\{0, \gamma_c(1-\gamma_c) + 2c\mathsf{f}(c)(\tau_2^c + \gamma_c - 1) + c^2\mathsf{f}^2(c)(\tau_4 - 1)\}.$$

Moreover, with the tightness result of $\mathbb{G}_n$ in Theorem 2.2 by Assumption 2.1(ia, ii), weak convergence arises so

$$\mathbb{G}_n \rightsquigarrow \mathsf{GP}(0, \Sigma).$$

The rest is to calculate the covariance for the above process between the time s and t where $c_0 \le s \le t < \infty$. Replace c by s and t respectively in the above stochastic expansion of $n^{1/2}(\widehat{\gamma}_c^{(0)} - \gamma_c)$, and then compute asymptotic covariance as

$$\Sigma_{st} = \gamma_t(1-\gamma_s) + s\mathsf{f}(s)(\tau_2^t + \gamma_t - 1) + t\mathsf{f}(t)(\tau_2^s + \gamma_s - 1) + st\mathsf{f}(s)\mathsf{f}(t)(\tau_4 - 1). \quad ∎$$

Proof of Theorem 2.4. Since Assumption 2.1(ii) holds for $\mathcal{I}_1, \mathcal{I}_2$, the initial estimators satisfy $N_j^{-1}(\widehat{\beta}_j - \beta) = \mathrm{O}_\mathsf{P}(1)$, $n_j^{1/2}(\widehat{\sigma}_j - \sigma) = \mathrm{O}_\mathsf{P}(1)$ and

$$n_j^{1/2}(\widehat{\sigma}_j - \sigma) = n_j^{-1/2}(2\sigma)^{-1}\sum_{i\in\mathcal{I}_j}(\varepsilon_i^2 - \sigma^2) + \mathrm{O}_\mathsf{P}(n_j^{-1/2}),$$

for $j = 1, 2$. Insert this into the asymptotic expansion in Theorem 2.1 by Assumption 2.1(ia, ii) and use

$n_1 = n_2 = n/2$ to get

$$\begin{aligned} n_j^{-1/2}\sum_{i\in\mathcal{I}_j}\{1_{(|y_i-x_i'\widehat{\beta}_{3-j}|>\widehat{\sigma}_{3-j}c)}-\gamma_c\} &= n_j^{-1/2}\sum_{i\in\mathcal{I}_j}\{1_{(|\varepsilon_i|>\sigma c)}-\gamma_c\} \\ &\quad - c\mathsf{f}(c)n_j^{-1/2}\sum_{i\in\mathcal{I}_{3-j}}(\frac{\varepsilon_i^2}{\sigma^2}-1)+\mathrm{o}_{\mathsf{P}}(1), \end{aligned}$$

for $j = 1, 2$ and uniformly in $c \in [c_0, \infty)$. Combine counts of outliers for two subsamples $\mathcal{I}_1, \mathcal{I}_2$ to obtain

$$\begin{aligned} n^{1/2}(\widehat{\gamma}_c^{(-1)}-\gamma_c) &= n^{-1/2}\sum_{j=1}^{2}\sum_{i\in\mathcal{I}_j}\{1_{(|y_i-x_i'\widehat{\beta}_{3-j}|>\widehat{\sigma}_{3-j}c)}-\gamma_c\} \\ &= n^{-1/2}\sum_{j=1}^{2}[\sum_{i\in\mathcal{I}_j}\{1_{(|\varepsilon_i|>\sigma c)}-\gamma_c\} - c\mathsf{f}(c)\sum_{i\in\mathcal{I}_{3-j}}(\frac{\varepsilon_i^2}{\sigma^2}-1)]+\mathrm{o}_{\mathsf{P}}(1), \end{aligned}$$

uniformly in c. This reduces to the expansion for the Robustified Least Squares. Thus argue along the lines of Theorem 2.3 to attain the same limiting Gaussian process. ▮

Proof of Theorem 2.5. By Assumption 2.1(ia, ii, iii) with $\eta = 1/4$, Theorem 3 in Jiao and Nielsen (2017) shows, when m, n are large, that the iterated 1-step Huber-skip M-estimator converges in probability to the fixed point $\widehat{\beta}_c^*$, $\widehat{\sigma}_c^*$ where $N^{-1}(\widehat{\beta}_c^*-\beta) = \mathrm{O}_{\mathsf{P}}(1)$, $n^{1/2}(\widehat{\sigma}_c^*-\sigma) = \mathrm{O}_{\mathsf{P}}(1)$ and

$$n^{1/2}(\widehat{\sigma}_c^*-\sigma) = \frac{1}{2\sigma\{\tau_2^c - c(c^2-\varsigma_c^2)\mathsf{f}(c)\}}n^{-1/2}\sum_{i=1}^{n}(\varepsilon_i^2-\varsigma_c^2\sigma^2)1_{(|\varepsilon_i|\le\sigma c)}+\mathrm{o}_{\mathsf{P}}(1),$$

uniformly in c. Thus the gauge also has the fixed point $\widehat{\gamma}_c^* = n^{-1}\sum_{i=1}^{n}1_{(|y_i-x_i'\widehat{\beta}_c^*|>\widehat{\sigma}_c^*c)}$, and then substitute $n^{1/2}(\widehat{\sigma}_c^*-\sigma)$ into the stochastic expansion in Theorem 2.1 to obtain

$$\begin{aligned} n^{1/2}(\widehat{\gamma}_c^*-\gamma_c) &= n^{-1/2}\sum_{i=1}^{n}\{1_{(|\varepsilon_i|>\sigma c)}-\gamma_c\} \\ &\quad - \frac{c\mathsf{f}(c)}{\tau_2^c - c(c^2-\varsigma_c^2)\mathsf{f}(c)}n^{-1/2}\sum_{i=1}^{n}(\frac{\varepsilon_i^2}{\sigma^2}-\varsigma_c^2)1_{(|\varepsilon_i|\le\sigma c)}+\mathrm{o}_{\mathsf{P}}(1), \end{aligned}$$

uniformly in $c \in [c_0, \infty)$. CLT shows finite dimensional convergence so for any c

$$\mathbb{G}_n^*(c) = n^{1/2}(\widehat{\gamma}_c^*-\gamma_c) \overset{\mathsf{D}}{\to} \mathsf{N}[0, \gamma_c(1-\gamma_c) + \{\frac{c\mathsf{f}(c)}{\tau_2^c - c(c^2-\varsigma_c^2)\mathsf{f}(c)}\}^2\{\tau_4^c - \frac{(\tau_2^c)^2}{1-\gamma_c}\}].$$

Moreover, with tightness of $\mathbb{G}_n^*$ by Theorem 2.2, weak convergence is established so

$$\mathbb{G}_n^* \rightsquigarrow \mathsf{GP}(0, \Sigma).$$

For $c_0 \le s \le t < \infty$, replace c by s and t respectively in the expansion of $n^{1/2}(\widehat{\gamma}_c^*-\gamma_c)$, and then compute their asymptotic covariance as

$$\Sigma_{st} = \gamma_t(1-\gamma_s) - \frac{t\mathsf{f}(t)}{\tau_2^t - t(t^2-\varsigma_t^2)\mathsf{f}(t)}\{\varsigma_t^2(1-\gamma_s) - \tau_2^s\}$$
$$+ \frac{s\mathsf{f}(s)}{\tau_2^s - s(s^2-\varsigma_s^2)\mathsf{f}(s)}\frac{t\mathsf{f}(t)}{\tau_2^t - t(t^2-\varsigma_t^2)\mathsf{f}(t)}\{\tau_4^s - \frac{(\tau_2^s)^2}{1-\gamma_s}\}. \blacksquare$$

Proof of Theorem 2.7. Notice first $\sum_{c\in\mathcal{C}_K} n^{1/2}(\widetilde{\gamma}_c - \gamma_c) = \sum_{k=1}^K n^{1/2}(\widetilde{\gamma}_{c_k} - \gamma_{c_k})$. Due to Assumption 2.1$(ia, ii)$, proofs of Theorem 2.3, 2.4 show the asymptotic expansion

$$n^{1/2}(\widetilde{\gamma}_c - \gamma_c) = n^{-1/2}\sum_{i=1}^n\{1_{(|\varepsilon_i|>\sigma c)} - \gamma_c\} - c\mathsf{f}(c)n^{-1/2}\sum_{i=1}^n(\frac{\varepsilon_i^2}{\sigma^2} - 1) + \mathsf{o}_\mathsf{P}(1),$$

uniformly in $c \in [c_0, \infty)$. Note c_1 is chosen in a way so that $c_1 \geq c_0$. Then replace c by $c_k, k = 1, 2, \ldots, K$ respectively in the above expansion and insert these into the test statistic to obtain

$$\sum_{k=1}^K n^{1/2}(\widetilde{\gamma}_{c_k} - \gamma_{c_k}) = \sum_{k=1}^K n^{-1/2}\sum_{i=1}^n\{1_{(|\varepsilon_i|>\sigma c_k)} - \gamma_{c_k}\}$$
$$- \{\sum_{k=1}^K c_k\mathsf{f}(c_k)\}\{n^{-1/2}\sum_{i=1}^n(\frac{\varepsilon_i^2}{\sigma^2} - 1)\} + \mathsf{o}_\mathsf{P}(1).$$

Therefore CLT shows

$$\sum_{c\in\mathcal{C}_K} n^{1/2}(\widetilde{\gamma}_c - \gamma_c) \overset{\mathsf{D}}{\to} \mathsf{N}(0, \mathsf{Var}_{\mathcal{C}_K}),$$

where

$$\mathsf{Var}_{\mathcal{C}_K} = \sum_{k=1}^K \gamma_{c_k}(1-\gamma_{c_k}) + 2\sum_{1\leq k<l\leq K}(\gamma_{c_l} - \gamma_{c_k}\gamma_{c_l}) + \{\sum_{k=1}^K c_k\mathsf{f}(c_k)\}^2(\tau_4 - 1)$$
$$+ 2\{\sum_{k=1}^K c_k\mathsf{f}(c_k)\}\{\sum_{k=1}^K(\tau_2^{c_k} + \gamma_{c_k} - 1)\}. \blacksquare$$

6.2 Appendix: Simulation Tables

Here we provide the full simulation results of the proposed tests under the null as well as under a range of alternatives.

6.2.1 Simulation Results Tables under the Null

Table 3: Properties of the IIS Outlier Tests Under the Null in a Static Model: gauge, simulation- & theory-standard deviation of the gauge, and rejection frequencies for different nominal test levels of $p = (0.01, 0.05)$ where "Prop" refers to the standard normal proportion test, "Cnt" refers to the count Poisson test. M=4000 replications.

$\gamma = 0.01$	$\hat{\gamma}$	$sd(\hat{\gamma})$	$sd(\hat{\gamma})$ (theor.)	p=0.01 (Prop)	p=0.05 (Prop)	p=0.01 (Cnt)	p=0.05 (Cnt)
T=30	0.013	0.022	0.015	0.063	0.063	0.011	0.063
T=50	0.012	0.016	0.012	0.026	0.126	0.004	0.026
T=100	0.011	0.01	0.008	0.016	0.078	0.002	0.016
T=200	0.01	0.006	0.006	0.008	0.033	0.001	0.008
T=400	0.01	0.004	0.004	0.009	0.039	0.002	0.017

$\gamma = 0.05$	$\hat{\gamma}$	$sd(\hat{\gamma})$	$sd(\hat{\gamma})$ (theor.)	p=0.01 (Prop)	p=0.05 (Prop)	p=0.01 (Cnt)	p=0.05 (Cnt)
T=30	0.073	0.045	0.027	0.154	0.154	0.014	0.051
T=50	0.072	0.033	0.021	0.125	0.299	0.016	0.125
T=100	0.05	0.016	0.015	0.022	0.101	0	0.004
T=200	0.049	0.011	0.01	0.009	0.036	0	0.002
T=400	0.05	0.007	0.007	0.012	0.062	0	0.002

Table 4: Properties of the IIS Outlier Tests Under the Null in a Dynamic Model (AR $\beta = 0.5$): gauge, simulation- & theory-standard deviation of the gauge, and rejection frequencies for different nominal test levels of $p = (0.01, 0.05)$ where "Prop" refers to the standard normal proportion test, "Cnt" refers to the Poisson count test. M=4000 replications.

$\gamma = 0.01$	$\hat{\gamma}$	$sd(\hat{\gamma})$	$sd(\hat{\gamma})$ (theor.)	p=0.01 (Prop)	p=0.05 (Prop)	p=0.01 (Cnt)	p=0.05 (Cnt)
T=30	0.013	0.023	0.015	0.062	0.062	0.015	0.062
T=50	0.012	0.017	0.012	0.125	0.125	0.008	0.028
T=100	0.011	0.01	0.008	0.016	0.079	0.002	0.016
T=200	0.01	0.006	0.006	0.006	0.036	0.001	0.006
T=400	0.01	0.004	0.004	0.011	0.037	0.003	0.018

$\gamma = 0.05$	$\hat{\gamma}$	$sd(\hat{\gamma})$	$sd(\hat{\gamma})$ (theor.)	p=0.01 (Prop)	p=0.05 (Prop)	p=0.01 (Cnt)	p=0.05 (Cnt)
T=30	0.077	0.05	0.027	0.165	0.373	0.028	0.064
T=50	0.073	0.034	0.021	0.126	0.284	0.021	0.126
T=100	0.049	0.016	0.015	0.028	0.116	0	0.005
T=200	0.05	0.011	0.01	0.008	0.064	0	0.001
T=400	0.05	0.007	0.007	0.01	0.061	0	0.003

Table 5: Properties of the IIS Outlier Tests Under the Null in a Dynamic Model (AR $\beta = 0.95$): gauge, simulation- & theory-standard deviation of the gauge, and rejection frequencies for different nominal test levels of $p = (0.01, 0.05)$ where "Prop" refers to the standard normal proportion test, "Cnt" refers to the Poisson count test. M=4000 replications.

$\gamma = 0.01$	$\hat{\gamma}$	$sd(\hat{\gamma})$	$sd(\hat{\gamma})$ (theor.)	p=0.01 (Prop)	p=0.05 (Prop)	p=0.01 (Cnt)	p=0.05 (Cnt)
T=30	0.014	0.025	0.015	0.069	0.069	0.016	0.069
T=50	0.012	0.018	0.012	0.124	0.124	0.008	0.035
T=100	0.011	0.01	0.008	0.014	0.086	0.002	0.014
T=200	0.011	0.007	0.006	0.008	0.046	0.002	0.008
T=400	0.01	0.004	0.004	0.01	0.047	0.004	0.022

$\gamma = 0.05$	$\hat{\gamma}$	$sd(\hat{\gamma})$	$sd(\hat{\gamma})$ (theor.)	p=0.01 (Prop)	p=0.05 (Prop)	p=0.01 (Cnt)	p=0.05 (Cnt)
T=30	0.083	0.057	0.027	0.195	0.411	0.043	0.088
T=50	0.075	0.036	0.021	0.138	0.301	0.024	0.138
T=100	0.051	0.017	0.015	0.03	0.121	0.001	0.007
T=200	0.049	0.011	0.01	0.008	0.058	0	0.001
T=400	0.05	0.007	0.007	0.011	0.06	0	0.003

Table 6: Size of the scale sum test under the null hypothesis of no outliers for scale levels $K = (10, 15, 20)$ for static and dynamic models AR $\beta = (0, 0.5, 0.95)$ with error variance 1 for sample size $n = (50, 100, 200)$ for nominal test levels $p = (0.01, 0.05)$ with replications 1000.

	AR $\beta = 0$		AR $\beta = 0.5$		AR $\beta = 0.95$	
$K = 10$	$p = 0.01$	$p = 0.05$	$p = 0.01$	$p = 0.05$	$p = 0.01$	$p = 0.05$
$n = 50$	0.245	0.364	0.232	0.336	0.345	0.455
$n = 100$	0.026	0.082	0.021	0.094	0.027	0.095
$n = 200$	0.015	0.058	0.011	0.073	0.019	0.061
$K = 15$	$p = 0.01$	$p = 0.05$	$p = 0.01$	$p = 0.05$	$p = 0.01$	$p = 0.05$
$n = 50$	0.235	0.368	0.221	0.335	0.398	0.497
$n = 100$	0.016	0.063	0.020	0.082	0.029	0.097
$n = 200$	0.015	0.070	0.017	0.066	0.016	0.071
$K = 20$	$p = 0.01$	$p = 0.05$	$p = 0.01$	$p = 0.05$	$p = 0.01$	$p = 0.05$
$n = 50$	0.192	0.363	0.185	0.328	0.403	0.530
$n = 100$	0.018	0.076	0.024	0.083	0.035	0.093
$n = 200$	0.009	0.061	0.005	0.055	0.017	0.050

Table 7: Size of the scale sup test under the null hypothesis of no outliers for scale levels $K = (10, 15, 20)$ for static and dynamic models AR $\beta = (0, 0.5, 0.95)$ with error variance 1 for sample size $n = (50, 100, 200)$ for nominal test levels $p = (0.01, 0.05)$ with replications 1000.

	AR $\beta = 0$		AR $\beta = 0.5$		AR $\beta = 0.95$	
$K = 10$	$p = 0.01$	$p = 0.05$	$p = 0.01$	$p = 0.05$	$p = 0.01$	$p = 0.05$
$n = 50$	0.181	0.386	0.165	0.344	0.294	0.462
$n = 100$	0.008	0.045	0.009	0.039	0.019	0.053
$n = 200$	0.018	0.055	0.019	0.078	0.018	0.067
$K = 15$	$p = 0.01$	$p = 0.05$	$p = 0.01$	$p = 0.05$	$p = 0.01$	$p = 0.05$
$n = 50$	0.118	0.310	0.124	0.292	0.278	0.462
$n = 100$	0.014	0.101	0.019	0.098	0.025	0.116
$n = 200$	0.011	0.043	0.012	0.047	0.018	0.060
$K = 20$	$p = 0.01$	$p = 0.05$	$p = 0.01$	$p = 0.05$	$p = 0.01$	$p = 0.05$
$n = 50$	0.054	0.220	0.067	0.213	0.223	0.414
$n = 100$	0.026	0.056	0.021	0.073	0.043	0.091
$n = 200$	0.007	0.031	0.006	0.037	0.007	0.024

Table 8: Properties of the scaling global test using repeated proportion tests under the null of no outliers: Null rejection frequency for different nominal test levels of $p = (0.01, 0.05)$ for two different sets of significance levels of outlier detection $m = 10, 15, 20$ for static, and dynamic models (AR $\beta = 0.5, 0.95$).

	Static		AR $\beta = 0.5$		AR $\beta = 0.95$	
$m = 10$	p=0.01 (prop)	p=0.05 (prop)	p=0.01 (prop)	p=0.05 (prop)	p=0.01 (prop)	p=0.05 (prop)
T=30	0.240	0.364	0.252	0.374	0.394	0.536
T=50	0.205	0.348	0.240	0.368	0.318	0.445
T=100	0.036	0.086	0.032	0.076	0.044	0.089
T=200	0.017	0.050	0.020	0.051	0.030	0.064
$m = 15$	p=0.01 (prop)	p=0.05 (prop)	p=0.01 (prop)	p=0.05 (prop)	p=0.01 (prop)	p=0.05 (prop)
T=50	0.221	0.381	0.24	0.364	0.408	0.532
T=100	0.044	0.089	0.031	0.066	0.049	0.091
T=200	0.012	0.046	0.012	0.056	0.011	0.046
$m = 20$	p=0.01 (prop)	p=0.05 (prop)	p=0.01 (prop)	p=0.05 (prop)	p=0.01 (prop)	p=0.05 (prop)
100	0.031	0.077	0.029	0.066	0.038	0.084
200	0.016	0.060	0.022	0.064	0.015	0.061

Table 9: Properties of the scaling global test using repeated count tests under the null of no outliers: Null rejection frequency for different nominal test levels of $p = (0.01, 0.05)$ for two different sets of significance levels of outlier detection $m = 10, 15, 20$ for static, and dynamic models (AR $\beta = 0.5, 0.95$).

	Static		AR $\beta = 0.5$		AR $\beta = 0.95$	
$m = 10$	p=0.01 (count)	p=0.05 (count)	p=0.01 (count)	p=0.05 (count)	p=0.01 (count)	p=0.05 (count)
T=30	0.005	0.033	0.016	0.048	0.063	0.118
T=50	0.005	0.025	0.012	0.048	0.033	0.074
T=100	0.000	0.007	0.000	0.002	0.001	0.005
T=200	0.001	0.002	0.000	0.002	0.000	0.005
$m = 15$	p=0.01 (count)	p=0.05 (count)	p=0.01 (count)	p=0.05 (count)	p=0.01 (count)	p=0.05 (count)
T=50	0.003	0.010	0.008	0.018	0.045	0.086
T=100	0.000	0.000	0.000	0.000	0.000	0.001
T=200	0.001	0.001	0.000	0.001	0.000	0.000
$m = 20$	p=0.01 (count)	p=0.05 (count)	p=0.01 (count)	p=0.05 (count)	p=0.01 (count)	p=0.05 (count)
100	0.000	0.000	0.000	0.001	0.000	0.001
200	0.000	0.001	0.000	0.002	0.000	0.001

6.2.2 Simulation Results Tables under Alternatives

Table 10: Power of the Proportion and Count Tests: Rejection frequency of the null hypothesis of the IIS Outlier tests when selecting at $\gamma = 0.01$ in a static model, for varying outlier magnitude δ, sample size T and proportion of the sample that are outliers in the DGP. "P" refers to the normal proportion test, "C" refers to the Poisson count test.

$\gamma = 0.01$	$\delta = 2\sigma$				$\delta = 4\sigma$			
T=30 / Prop Outl.	P 0.01	P 0.05	C 0.01	C 0.05	P 0.01	P 0.05	C 0.01	C 0.05
0.05	0.087	0.087	0.013	0.087	0.145	0.145	0.03	0.145
0.1	0.15	0.15	0.028	0.15	0.705	0.705	0.256	0.705
0.25	0.151	0.151	0.024	0.151	0.431	0.431	0.217	0.431
T=50 / Prop Outl.	P 0.01	P 0.05	C 0.01	C 0.05	P 0.01	P 0.05	C 0.01	C 0.05
0.05	0.051	0.215	0.01	0.051	0.154	0.768	0.019	0.154
0.1	0.071	0.321	0.016	0.071	0.729	0.943	0.394	0.729
0.25	0.104	0.354	0.019	0.104	0.388	0.624	0.203	0.388
T=100 / Prop Outl.	P 0.01	P 0.05	C 0.01	C 0.05	P 0.01	P 0.05	C 0.01	C 0.05
0.05	0.053	0.205	0.004	0.053	0.678	0.93	0.266	0.678
0.1	0.089	0.286	0.017	0.089	0.921	0.983	0.729	0.921
0.25	0.122	0.337	0.035	0.122	0.256	0.545	0.084	0.256
T=200 / Prop Outl.	P 0.01	P 0.05	C 0.01	C 0.05	P 0.01	P 0.05	C 0.01	C 0.05
0.05	0.05	0.159	0.017	0.05	0.941	0.99	0.788	0.941
0.1	0.093	0.25	0.027	0.093	0.993	0.998	0.959	0.993
0.25	0.166	0.35	0.05	0.166	0.295	0.522	0.125	0.295

Table 11: Power of the Proportion and Count Tests: Rejection frequency of the null hypothesis of the IIS Outlier test when selecting at $\gamma = 0.01$ in a dynamic model (AR $\beta = 0.5$), for varying outlier magnitude δ, sample size T and proportion of the sample that are outliers in the DGP. “P” refers to the normal proportion test, “C” refers to the Poisson count test.

$\gamma = 0.01$	$\delta = 2\sigma$				$\delta = 4\sigma$			
T=30 / Prop Outl.	P 0.01	P 0.05	C 0.01	C 0.05	P 0.01	P 0.05	C 0.01	C 0.05
0.05	0.093	0.093	0.019	0.093	0.191	0.191	0.028	0.191
0.1	0.13	0.13	0.03	0.13	0.579	0.579	0.204	0.579
0.25	0.149	0.149	0.041	0.149	0.401	0.401	0.189	0.401
T=50 / Prop Outl.	P 0.01	P 0.05	C 0.01	C 0.05	P 0.01	P 0.05	C 0.01	C 0.05
0.05	0.051	0.221	0.01	0.051	0.17	0.738	0.043	0.17
0.1	0.072	0.283	0.012	0.072	0.634	0.885	0.295	0.634
0.25	0.094	0.333	0.036	0.094	0.306	0.582	0.151	0.306
T=100 / Prop Outl.	P 0.01	P 0.05	C 0.01	C 0.05	P 0.01	P 0.05	C 0.01	C 0.05
0.05	0.054	0.22	0.007	0.054	0.596	0.893	0.22	0.596
0.1	0.081	0.294	0.02	0.081	0.813	0.955	0.552	0.813
0.25	0.093	0.291	0.015	0.093	0.238	0.531	0.072	0.238
T=200 / Prop Outl.	P 0.01	P 0.05	C 0.01	C 0.05	P 0.01	P 0.05	C 0.01	C 0.05
0.05	0.053	0.169	0.02	0.053	0.86	0.957	0.666	0.86
0.1	0.107	0.26	0.036	0.107	0.947	0.99	0.87	0.947
0.25	0.12	0.285	0.032	0.12	0.294	0.529	0.141	0.294

Table 12: Power of the Proportion and Count Tests: Rejection frequency of the null hypothesis of the IIS Outlier tests when selecting at $\gamma = 0.01$ in a dynamic model (AR $\beta = 0.95$), for varying outlier magnitude δ, sample size T and proportion of the sample that are outliers in the DGP. "P" refers to the normal test, "C" refers to the Poisson test.

$\gamma = 0.01$	$\delta = 2\sigma$				$\delta = 4\sigma$			
T=30 / Prop Outl.	P 0.01	P 0.05	C 0.01	C 0.05	P 0.01	P 0.05	C 0.01	C 0.05
0.05	0.116	0.116	0.025	0.116	0.481	0.481	0.078	0.481
0.1	0.169	0.169	0.035	0.169	0.486	0.486	0.213	0.486
0.25	0.143	0.143	0.041	0.143	0.282	0.282	0.101	0.282
T=50 / Prop Outl.	P 0.01	P 0.05	C 0.01	C 0.05	P 0.01	P 0.05	C 0.01	C 0.05
0.05	0.069	0.264	0.012	0.069	0.531	0.869	0.221	0.531
0.1	0.111	0.325	0.031	0.111	0.475	0.766	0.234	0.475
0.25	0.098	0.288	0.034	0.098	0.205	0.452	0.082	0.205
T=100 / Prop Outl.	P 0.01	P 0.05	C 0.01	C 0.05	P 0.01	P 0.05	C 0.01	C 0.05
0.05	0.077	0.279	0.015	0.077	0.811	0.962	0.526	0.811
0.1	0.111	0.306	0.029	0.111	0.481	0.759	0.227	0.481
0.25	0.097	0.283	0.021	0.097	0.187	0.43	0.067	0.187
T=200 / Prop Outl.	P 0.01	P 0.05	C 0.01	C 0.05	P 0.01	P 0.05	C 0.01	C 0.05
0.05	0.094	0.246	0.025	0.094	0.951	0.994	0.857	0.951
0.1	0.13	0.299	0.04	0.13	0.637	0.821	0.419	0.637
0.25	0.117	0.275	0.034	0.117	0.272	0.499	0.115	0.272

Table 13: Power of the scale sum test under the alternative hypothesis for scale levels $K = 10$, static model AR $\beta = 0$ with error variance 1, sample size $n = 100$, and for proportion of outliers $(0.05, 0.10, 0.25)$, magnitude of outliers $(2, 3, 4, 6)$, and for nominal test levels $p = (0.01, 0.05)$ with replications 1000.

$K = 10$, $n = 100$	$p = 0.01$				$p = 0.05$			
Proportion / Magnitude	2	3	4	6	2	3	4	6
0.05	0.033	0.033	0.015	0.000	0.102	0.099	0.059	0.001
0.10	0.058	0.249	0.701	1.000	0.155	0.487	0.877	1.000
0.25	0.099	0.579	0.969	1.000	0.236	0.767	0.989	1.000

Table 14: Power of the scale sum test under the alternative hypothesis for scale levels $K = 10$, stationary model AR $\beta = 0.5$ with error variance 1, sample size $n = 100$, and for proportion of outliers $(0.05, 0.10, 0.25)$, magnitude of outliers $(2, 3, 4, 6)$, and for nominal test levels $p = (0.01, 0.05)$ with replications 1000.

$K = 10$, $n = 100$	$p = 0.01$				$p = 0.05$			
Proportion / Magnitude	2	3	4	6	2	3	4	6
0.05	0.039	0.046	0.043	0.005	0.102	0.153	0.120	0.023
0.10	0.056	0.233	0.514	0.979	0.148	0.393	0.755	0.996
0.25	0.075	0.374	0.83	0.999	0.180	0.573	0.932	1.000

Table 15: Power of the scale sum test under the alternative hypothesis for scale levels $K = 10$, stationary model AR $\beta = 0.95$ with error variance 1, sample size $n = 100$, and for proportion of outliers $(0.05, 0.10, 0.25)$, magnitude of outliers $(2, 3, 4, 6)$, and for nominal test levels $p = (0.01, 0.05)$ with replications 1000.

$K = 10$, $n = 100$		$p = 0.01$				$p = 0.05$		
Proportion / Magnitude	2	3	4	6	2	3	4	6
0.05	0.051	0.181	0.389	0.593	0.143	0.353	0.610	0.762
0.10	0.084	0.339	0.695	0.883	0.191	0.551	0.837	0.963
0.25	0.068	0.132	0.279	0.565	0.163	0.292	0.460	0.735

Table 16: Power of the scale sup test under the alternative hypothesis for scale levels $K = 10$, static model AR $\beta = 0$ with error variance 1, sample size $n = 100$, and for proportion of outliers $(0.05, 0.10, 0.25)$, magnitude of outliers $(2, 3, 4, 6)$, and for nominal test levels $p = (0.01, 0.05)$ with replications 1000.

$K = 10$, $n = 100$		$p = 0.01$				$p = 0.05$		
Proportion / Magnitude	2	3	4	6	2	3	4	6
0.05	0.010	0.007	0.011	0.025	0.041	0.049	0.125	0.483
0.10	0.020	0.104	0.503	0.999	0.058	0.280	0.815	1.000
0.25	0.023	0.280	0.818	1.000	0.103	0.504	0.941	1.000

Table 17: Power of the scale sup test under the alternative hypothesis for scale levels $K = 10$, stationary model AR $\beta = 0.5$ with error variance 1, sample size $n = 100$, and for proportion of outliers $(0.05, 0.10, 0.25)$, magnitude of outliers $(2, 3, 4, 6)$, and for nominal test levels $p = (0.01, 0.05)$ with replications 1000.

$K = 10$, $n = 100$		$p = 0.01$				$p = 0.05$		
Proportion / Magnitude	2	3	4	6	2	3	4	6
0.05	0.011	0.012	0.016	0.026	0.049	0.085	0.135	0.328
0.10	0.019	0.081	0.330	0.962	0.062	0.232	0.647	0.996
0.25	0.023	0.172	0.568	0.988	0.079	0.356	0.791	0.999

Table 18: Power of the scale sup test under the alternative hypothesis for scale levels $K = 10$, stationary model AR $\beta = 0.95$ with error variance 1, sample size $n = 100$, and for proportion of outliers $(0.05, 0.10, 0.25)$, magnitude of outliers $(2, 3, 4, 6)$, and for nominal test levels $p = (0.01, 0.05)$ with replications 1000.

$K = 10$, $n = 100$		$p = 0.01$				$p = 0.05$		
Proportion / Magnitude	2	3	4	6	2	3	4	6
0.05	0.017	0.049	0.229	0.645	0.064	0.195	0.553	0.871
0.10	0.025	0.114	0.331	0.663	0.090	0.308	0.609	0.866
0.25	0.027	0.070	0.131	0.327	0.086	0.165	0.269	0.527

Table 19: Power of the scaling global test using repeated proportion tests: Rejection frequency of the null hypothesis of the scale test over $M = 10$ levels of selection significance in a static model, for varying outlier magnitude δ and nominal testing level $p = 0.01, 0.05$.

$m = 10, T = 100$	p=0.01 (Global Prop. Test)				p=0.05 (Global Prop. Test)			
Prop Outl.	$\delta = 2\sigma$	$\delta = 3\sigma$	$\delta = 4\sigma$	$\delta = 6\sigma$	$\delta = 2\sigma$	$\delta = 3\sigma$	$\delta = 4\sigma$	$\delta = 6\sigma$
0.05	0.073	0.295	0.694	0.960	0.123	0.330	0.709	0.970
0.1	0.124	0.588	0.955	0.997	0.192	0.641	0.967	0.998
0.25	0.188	0.528	0.797	0.951	0.277	0.665	0.896	0.966

Table 20: Power of the scaling global test using repeated proportion tests: Rejection frequency of the null hypothesis of the scale test over $M = 10$ levels of selection significance in a dynamic model (AR $\beta = 0.5$), for varying outlier magnitude δ and nominal testing level $p = 0.01, 0.05$.

$m = 10, T = 100$	p=0.01 (Global Prop. Test)				p=0.05 (Global Prop. Test)			
Prop Outl.	$\delta = 2\sigma$	$\delta = 3\sigma$	$\delta = 4\sigma$	$\delta = 6\sigma$	$\delta = 2\sigma$	$\delta = 3\sigma$	$\delta = 4\sigma$	$\delta = 6\sigma$
0.05	0.074	0.271	0.657	0.941	0.127	0.318	0.680	0.947
0.1	0.122	0.485	0.898	0.999	0.179	0.561	0.921	0.999
0.25	0.164	0.440	0.664	0.885	0.254	0.568	0.786	0.935

Table 21: Power of the scaling global test using repeated proportion tests: Rejection frequency of the null hypothesis of the scale test over $M = 10$ levels of selection significance in a dynamic model (AR $\beta = 0.95$), for varying outlier magnitude δ and nominal testing level $p = 0.01, 0.05$.

$m = 10, T = 100$	p=0.01 (Global Prop. Test)				p=0.05 (Global Prop. Test)			
Prop Outl.	$\delta = 2\sigma$	$\delta = 3\sigma$	$\delta = 4\sigma$	$\delta = 6\sigma$	$\delta = 2\sigma$	$\delta = 3\sigma$	$\delta = 4\sigma$	$\delta = 6\sigma$
0.05	0.121	0.499	0.879	0.991	0.189	0.560	0.907	0.994
0.1	0.163	0.511	0.781	0.949	0.239	0.623	0.876	0.965
0.25	0.182	0.276	0.370	0.567	0.256	0.376	0.484	0.709

Table 22: Power of the scaling global test using repeated count tests: Rejection frequency of the null hypothesis of the scale test over $M = 10$ levels of selection significance in a static model, for varying outlier magnitude δ and nominal testing level $p = 0.01, 0.05$.

$m = 10, T = 100$	p=0.01 (Global Count Test)				p=0.05 (Global Count Test)			
Prop Outl.	$\delta = 2\sigma$	$\delta = 3\sigma$	$\delta = 4\sigma$	$\delta = 6\sigma$	$\delta = 2\sigma$	$\delta = 3\sigma$	$\delta = 4\sigma$	$\delta = 6\sigma$
0.05	0.001	0.004	0.026	0.014	0.008	0.061	0.27	0.821
0.1	0.003	0.066	0.462	0.959	0.022	0.264	0.793	0.991
0.25	0.007	0.027	0.122	0.475	0.042	0.203	0.421	0.743

Table 23: Power of the scaling global test using repeated count tests: Rejection frequency of the null hypothesis of the scale test over $M = 10$ levels of selection significance in a dynamic model (AR $\beta = 0.5$), for varying outlier magnitude δ and nominal testing level $p = 0.01, 0.05$.

$m = 10, T = 100$	p=0.01 (Global Count Test)				p=0.05 (Global Count Test)			
Prop Outl.	$\delta = 2\sigma$	$\delta = 3\sigma$	$\delta = 4\sigma$	$\delta = 6\sigma$	$\delta = 2\sigma$	$\delta = 3\sigma$	$\delta = 4\sigma$	$\delta = 6\sigma$
0.05	0.001	0.004	0.026	0.014	0.008	0.061	0.270	0.821
0.1	0.003	0.066	0.462	0.959	0.022	0.264	0.793	0.991
0.25	0.007	0.027	0.122	0.475	0.042	0.203	0.421	0.743

Table 24: Power of the scaling global test using repeated count tests: Rejection frequency of the null hypothesis of the scale test over $M = 10$ levels of selection significance in a dynamic model (AR $\beta = 0.95$), for varying outlier magnitude δ and nominal testing level $p = 0.01, 0.05$.

$m = 10, T = 100$	p=0.01 (Global Count Test)				p=0.05 (Global Count Test)			
Prop Outl.	$\delta = 2\sigma$	$\delta = 3\sigma$	$\delta = 4\sigma$	$\delta = 6\sigma$	$\delta = 2\sigma$	$\delta = 3\sigma$	$\delta = 4\sigma$	$\delta = 6\sigma$
0.05	0.121	0.499	0.879	0.991	0.189	0.56	0.907	0.994
0.1	0.163	0.511	0.781	0.949	0.239	0.623	0.876	0.965
0.25	0.182	0.276	0.37	0.567	0.256	0.376	0.484	0.709