Additive Outlier Detection via Extreme-Value Theory*

Peter Burridge
City University

A.M.Robert Taylor
University of Birmingham

July 2005

Abstract

This paper is concerned with detecting additive outliers using extreme value methods. The test recently proposed for use with possibly non-stationary time series by Perron and Rodriguez (2003), is, as they point out, extremely sensitive to departures from Gaussianity. As an alternative, we investigate the robustness to distributional form of a test based on weighted spacings of the sample order statistics. Difficulties arising from uncertainty about the number of potential outliers are discussed, and a simple algorithm requiring minimal distributional assumptions is proposed and its performance evaluated. The new algorithm has dramatically less level inflation in face of departures from Gaussianity than the PR test, yet retains good power in the presence of outliers.

Keywords: additive outliers, extreme order statistics, standardised spacings.

1 Introduction

There are at least three motives for attempting to detect outliers in sample data. One is pure data quality, in which the reliability of the recording mechanism may come under scrutiny, a second is the identification of extraneous causal factors, as in Basmann (2003), while a third concerns the robustness of subsequent statistical analysis to the presence of anomalous observations. In the latter case, the aim is to detect observations that are too extreme to have been produced by the same mechanism as the rest of the sample and to eliminate their effect on the subsequent analysis, for example by the introduction of suitable dummy variables in fitted regressions.

Two kinds of outliers have recently been studied in the context of unit root testing; Harvey, Leybourne and Newbold (2001), for example, analyse breaks in level caused by Innovational Outliers under the unit root null hypothesis, while Perron and Rodriguez

*Correspondence to: Peter Burridge, Department of Economics, City University, Northampton Square, London EC1V 0HB, U.K. Email: P.Burridge@City.ac.uk, Tel. (44) 20 7040 8919

(2003) (PR hereafter) propose a method for the detection and treatment of Additive Outliers (AO) in a similar context. AO, also termed "Type I" outliers as in Fox (1972), affect only the single observation to which they are attached. In the present paper we explore the extent to which the detection of AO can be made independent of distributional assumptions.

The starting point for this study is the test for AO in Gaussian processes suspected to contain a unit autoregressive root, proposed by PR, which we describe in Section 2. Using numerical evidence we demonstrate that the PR test is remarkably robust to the dynamics of the series under test, but is incredibly sensitive to the form of the distribution of the innovations. In particular, rejection rates under the null hypothesis of no outliers can be close to unity in the presence of data from distributions with heavier tails than the Normal. Of course, an outlier can only be defined with respect to a statistical model, and wrongly assuming Gaussianity when the true model has heavier tails leads to a very high false rejection rate; that is, the Gaussian assumption underlying the PR test makes it arguably unusable for the type of sample data that are common in econometric work.

PR mention the possibility of using EV methods to develop tests for outliers which are robust to the underlying distribution of the shocks but do not report any details. In Section 3 we develop this idea and argue that when the distribution of the underlying shocks is unknown, but is in the domain of attraction of a known Extreme Value (EV) distribution (here the Gumbel distribution), outliers can be defined by reference to the distribution of the spacings between sample order statistics. We initially motivate our approach through the IID case but later generalise to the time series setting, presenting a suggested practical testing algorithm. A numerical comparison of our recommended procedure with that of PR for various sample sizes and innovation distributions demonstrates the comparative robustness of our proposal, although both level and power are somewhat affected by the underlying dynamics of the series involved. Problems of aliasing arise, and a simple modification that reduces these is proposed. Section 4 concludes.

2 Additive Outliers in Time Series and the Perron-Rodriguez Test.

Following PR and Fox (1972), consider the following time-series process which is contaminated with additive outliers from some unspecified exogenous source:

$$z_t = y_t + \sum_j \mathcal{I}_{\tau_j} x_{\tau_j},\ t = 1, ..., T \tag{2.1}$$

$$y_t = \phi y_{t-1} + u_t \tag{2.2}$$

where $\mathcal{I}_{\tau_j} = 1(t = \tau_j), = 0(t \neq \tau_j)$ and $\{u_t\}$ is some ARMA process. Here, we observe z_t but might want to make inferences about ϕ. The problem is that we generally do not know either the relevant dates, τ_j, or the magnitudes of the x_{τ_j}.

When $\phi = 1$, as is a common hypothesis in econometric work, there are good reasons for basing outlier detection on the first differenced data, as shown by PR. They adopt a sequential approach based on estimating x_{τ_j} by

$$\begin{aligned} \hat{x}_{\tau_j} &= \frac{1}{2}(\Delta z_{\tau_j} - \Delta z_{\tau_j+1}) \\ &= \frac{1}{2}(\Delta y_{\tau_j} - \Delta y_{\tau_j+1} + 2x_{\tau_j}) \end{aligned} \tag{2.3}$$

at every possible date, τ_j, in the sample. If $\phi = 1$ and $u_t \sim IID(0, \sigma_u^2)$ this delivers an efficient estimate, while more generally, if Δy_t is a (covariance) stationary process with autocovariances, γ_0, γ_1, etc., then the variance of $\hat{x}_{\tau_j}$ is $(\gamma_0 - \gamma_1)/2$, which can easily be estimated to yield a studentised statistic, $\hat{\delta}_{\tau_j}$, say. We note that this is only the case if τ_j is isolated from neighbouring outlier dates.

Assuming that any outliers present *are* isolated, however, it becomes possible to calculate $|\hat{\delta}_{\tau_j}|$ for *every* date in the sample except the first and last, taking the supremum as test statistic. PR give critical values for $\sup |\hat{\delta}_{\tau_j}|$ computed by Monte Carlo with $\phi = 1$ and $u_t \tilde{} NIID(0, \sigma_u^2)$ and report that the test level is robust to departures from these dynamic assumptions, though not to a large departure from Gaussianity. They demonstrate that with Gaussian shocks, sequential use of the test, with detected outliers removed at each step, leads to a procedure with very little level inflation, and good power against at least a few specific alternatives.

We demonstrate numerically the extent to which the PR outlier detection test is over-sized in the presence of non-Gaussian shocks. In Table 1 we report empirical significance levels at nominal 5% when the data generating process is as in (2.1)-(2.2) with $\phi = 1$ and $u_t \sim IID$ from various parent distributions. In each case PR's test statistic is referred to critical values calculated assuming Normal shocks.

Tables 1 – 2 about here

The results in Table 1 are striking. The PR test, although well behaved under Gaussian shocks, is massively over-sized when the data are drawn from asymmetric or heavier-tailed distributions. The effect is largest for the centred exponential, $\chi^2_{(1)}$, and t distributions with small degrees of freedom. Moreover, for a given distribution the distortions also increase, often substantially, with the sample size. For example, for the $\chi^2_{(1)}$ case, level, already as high as 66% for $T = 100$, is effectively unity by $T = 1000$.

As mentioned above, PR are well aware that their statistic has a sampling distribution that is sensitive to shock type; of course, they stress its great merit, which is robustness to variation in the dynamics, as illustrated in Table 2. Here we specify that u_t of (2.2) be generated as the $ARMA(1,1)$ process:

$$(1 - \psi L)u_t = (1 - \theta L)\varepsilon_t \tag{2.4}$$

with $\varepsilon_t \sim NIID(0, \sigma_\varepsilon^2)$ for various relevant values of the design parameters ϕ, ψ and θ. It is evident from the results in Table 2 that the only really problematic case is the model with two unit autoregressive roots, for which the test is very conservative, even for a sample of 1000 observations. Practically speaking, the only obvious defect of the PR test is the extreme level inflation for non-Gaussian shocks revealed in Table 1.

3 Outlier Detection using EV Methods

As noted in the introduction, an outlier can only be defined with respect to a statistical model. The basic reason is that if we have a sample of size, n, the empirical distribution function has a step of height n^{-1} at each sample point; only after some prior information about the population distribution function has been introduced will we have reason to depart from the empirical distribution function. In applications in which the underlying distribution can plausibly be treated as Gaussian, it is commonplace for outliers to be defined via sample standard deviations, for example as points further than 4σ from the mean. In general, however, it is better to define them in terms of tail probabilities of the reference distribution, if this is known. Basmann (2003) explains the latter point very well in an entertaining setting. However, when the reference distribution is unknown, so of course are the tail probabilities, and the key question is exactly how much prior information to introduce. We now approach the implications of this using some results from the theory of extremes.

Initially, and to help motivate our approach to testing for outliers based on EV methods, we will consider in Section 3.1 the case of an independent sample. The independence assumption is subsequently relaxed in Section 3.2.

3.1 IID samples

To fix notation, suppose in the following that we have an independently drawn sample, say $\{X_1, ..., X_n\}$ from a population with cumulative density function (cdf), $F(x)$. Our interest centres on testing, at some pre-specified significance level α, the null hypothesis, H_0, that every sample member is a draw from the same continuous distribution with support the whole real line, against the alternative hypothesis, H_1, that one or more sample members are not drawn from this distribution, and are "too large". We will ignore altogether auxiliary information that could be relevant, such as the incidence of ties, or rounding, that could also signal special observations, and focus exclusively on the upper tail of the distribution involved.

In the case where the parent distribution is known and there is a single possible outlier at a known point in the sample (known index, say j) then H_0 can be rejected if the $p-$value, $p(X_j) := 1 - F(X_j) < \alpha$. That is, we reject H_0 if $X_j > x_\alpha$, where $p(x_\alpha) = \alpha$. In the unknown index case the relevant information is contained in the extreme order statistic of $\{X_j,\ j \in (1, ..., n)\}$. Let the descending order statistics be $X_{1:n} \geq X_{2:n} \geq ... \geq X_{n:n}$. The cdf of the maximum, $X_{1:n}$, is $\{F(x)\}^n$, and so we can choose a critical value, x_α, such that $1 - \{F(x_\alpha)\}^n = \alpha$, and proceed as for the known index case, comparing $X_{1:n}$ with the new critical value, x_α.

The procedure outlined in the previous paragraph can be readily generalised to the case where we have at most k outliers. In the known indices case, taking $k = 2$ to illustrate the point, there are three outcomes: 0 outliers, 1 outlier, or 2 outliers. Determining X_j to be an outlier if $p(X_j) = 1 - F(X_j) < \beta$, then, under H_0, the probability of zero, one and two outliers being detected is $(1-\beta)^2$, $2\beta(1-\beta)$ and β^2, respectively. Consequently, a critical value, x_α, can be chosen so that $p(x_\alpha) = \beta$, with $1 - (1-\beta)^2 = \alpha$. If up to k outliers are possible under the alternative, to keep the overall test level fixed at α we set β to solve $1 - (1-\beta)^k = \alpha$. For the unknown indices case, and since the alternative specifies at most k outliers, we have to set x_α such that $\Pr\{\textit{between } 1 \text{ and } k \text{ sample members exceed } x_\alpha | H_0\} = \alpha$, via the Binomial. A conservative approach would be to solve $1 - \alpha = F(x_\alpha)^n$, thus ignoring the upper bound, k. The latter approach treats the number of outliers as a secondary problem in order to provide a simple means of bounding the level at the target, α. The number of outliers is just given by $\min\{k, \text{number of } X_{j:n} \text{ exceeding } x_\alpha\}$. In both the known and unknown indices cases it can clearly be seen that in order to maintain a fixed significance level α we must sacrifice power to detect outliers.

In reality the assumption made above that the cdf of the parent distribution is known is untenable. We therefore outline the following two cases of primary importance and discuss how EV theory holds the key to analysing this problem.

Case 1: Unknown distribution, at most one outlier, unknown index. Under the null hypothesis, the large sample behaviour of order statistics is governed by the tails of $F(\cdot)$, so we shall have to make some assumption about the tails. If we suppose that F is in the maximum domain of attraction of the Gumbel distribution, with distribution function, $G(x) = \exp\{-\exp\{-x\}\}, -\infty \leq x \leq \infty$, then we can use the following fact. Let $D_{i,n} = X_{i:n} - X_{i+1:n}$ be the *spacings* between successive order statistics, then, for a suitable sequence of constants, a_n, the objects, $\{a_n^{-1} D_{i,n}, i = 1, ..., k\}$ converge in distribution to a vector of k independent exponential variates with means proportional to i^{-1}. See, for example, Weissman (1978, Theorem 3). Consequently, the $iD_{i,n}$, $i = 1, ..., k$, are approximately IID. Assuming the sample is large enough to permit calculation of 20 spacings, for example, we can then use the approximate IID property to construct a test with asymptotic level 5%: $X_{1:n}$ is an outlier if $D_{1,n} > iD_{i,n}$ $\forall i \in (2, ..., 20)$. Similarly, if the sample is large enough, we can get a test with level 2.5% by evaluating 40 spacings, and so on. We shall present some evidence on the robustness and power of this test, below.

Case 2: Unknown distribution, unknown number of outliers, unknown indices. Clearly we must continue to make some assumption about tail behaviour, so we will maintain that F is in the maximum domain of attraction of the Gumbel distribution. We will surely wish to say that there are k outliers if $X_{1:n}, X_{2:n}, ..., X_{k:n}$ are "too large" relative to $X_{k+1:n}, X_{k+2:n}, ...X_{n:n}$, but how shall we decide the value of k? Suppose n is moderately large, so that we may evaluate a conveniently large number of spacings, S, say. Let K be an upper bound on the number of outliers to be entertained. Now it seems obvious that kD_k must be large, and $(k+1)D_{k+1}$ not, if we are to declare that there are k outliers but not $k+1$. This suggests (at least) two potentially interesting criteria on which to base tests.

C1: There are $0 < k_{c1} \leq K$ outliers if and only if $S_{k_{c1}} = k_{c1} D_{k_{c1}}$ is the largest of the S standardised spacings.

C2: There are $0 < k_{c2} \leq K$ outliers if k_{c2} is the largest k such that $S_k \geq S_j$ for all $j > k$ and $k \leq K$.

Notice that the two criteria lead to tests of the same overall size, since $\Pr\{k_{c1} = 0\} = \Pr\{k_{c2} = 0\}$. This is so because $k_{c1} = 0 \Longleftrightarrow$ {largest spacing is S_j for some $j > K$}, while {largest spacing is S_j for some $j > K$} $\Longleftrightarrow$ {no S_k exceeds S_j for all $j > k$ and $k \leq K$} $\Longleftrightarrow k_{c2} = 0$. Assume that the S spacings are IID when no outliers are present. Then $E\{k_{c1}\} = K(K+1)/2S$ because $\Pr\{k_{c1} = k\} = 1/S$ for each $k \in (1, 2, ..., K)$. On the other hand, k_{c2} has a higher mean than k_{c1}. We experimented with C2, but found that the results were not interestingly different from those for C1, so they are not reported. We do find, however, that in the time-series setting C1 can be misleading unless used in a sequential fashion; see the numerical results in Section 3.2.

3.1.1 Numerical Evidence: EV methods in IID samples

We now present some evidence on the above detection rules using extreme order statistics from IID samples. We consider first the use of the asymptotic spacings result mentioned in the previous section, and then discuss modifications that provide better finite sample performance. While we demonstrate that use of distributional information can improve the level, we have found that such information is extremely difficult to extract from a sample. A workable compromise is to adopt a weighting scheme for the spacings that gives good empirical size for Gaussian samples, while being slightly liberal in other leading cases. We show subsequently that this approach is much more robust to variations in distributional form than the outlier test of PR.

An approach based on the Asymptotic Distribution of Standardised Spacings

Let $D_{i,n} := X_{i:n} - X_{i+1:n}$, $i = 1, ..., S$, as above, then we define the *standardised spacings*, using the asymptotic weights discussed in Case 1 above, as $iD_{i,n}$, $i = 1, ..., S$. Then, criterion C1 of Case 2 states that "if the jth standardised spacing is the largest, and $j \leq K$, then there are j outliers." We illustrate test level using IID pseudo-random variables, either distributed as $N(0,1)$ or as $\chi^2_{(1)} - 1$, or as $Exponential(\lambda = 1) - 1$.

Tables 3 – 7 about here

If the spacings were independent and identically distributed, with $E\{iD_i\} = \lambda_1(n)$, $i = 1, ..., 40$, then criterion C1 should have significance level 1/40 for each possible number of outliers (i.e. the proportions in the first, second and third columns of Tables 3-5 should be .025, .025 and .025, respectively). It is apparent from Table 3 that this rule is highly conservative in the case of Gaussian shocks, but that its level does gradually approach the nominal level as the sample size increases. From Table 4 we see that with shocks drawn from a centred Chi-square distribution with one degree of freedom, the test is somewhat liberal in small samples, but again its level approaches the nominal level as sample size increases. In Table 5 we see that when the shocks are exponential the rule is correctly sized at all sample sizes, as indeed it should be.

On the face of it these results seem a little disappointing: the test is either too liberal or too conservative in small samples according to the distribution from which the sample is drawn, and still conservative in very large Gaussian samples. However, when we compare these results with those from Table 1 for the PR test we see that our procedure is, in fact, relatively robust in comparison. Moreover, the unadjusted power of our testing procedure is encouraging, as illustrated in Tables 6 and 7.

Two things are clear from the results in Tables 6 and 7. Firstly, the test detects a very large outlier (with associated tail probability of less than 6×10^{-7}) only slightly less than half the time with a Gaussian sample (reflecting the very conservative level in this case), but does much better in the Chi-square case. Secondly, for given level, power declines with increasing sample size, as it should, since any fixed extreme observation necessarily becomes less implausible as sample size increases, given that the underlying distribution has unbounded support.

To see how far it is possible to adjust the test level using knowledge of the parent distribution, we examine the weights that should be applied to the spacings to give them a common mean in finite samples. To do this we consider an expansion of the quantile function, as described by Arnold, Balakrishnan and Nagaraja (1992, p.128) [ABN], whom we follow below (except that we continue to denote the *largest* order statistic by $X_{1:n}$).

A Procedure based on an Expansion for Mean Spacings

Suppose F is a continuous distribution function on $(-\infty, \infty)$, and let $\{U_{i:n}\}$ be the order statistics from an IID Uniform $[0, 1]$ sample, then $\{X_{i:n}\} = \{F^{-1}(U_{i:n})\}$ are the order statistics from an IID sample with common continuous distribution, F. Assuming F^{-1} has as many derivatives as we require, we may construct a Taylor series expansion of $F^{-1}(U_{i:n})$ around the point, $E\{U_{i:n}\} = (n - i + 1)/(n + 1) = p_i$ as follows:

$$X_{i:n} = F^{-1}(p_i) + F_1^{-1}(p_i)(U_{i:n} - p_i) + \frac{1}{2}F_2^{-1}(p_i)(U_{i:n} - p_i)^2 + ... \qquad (3.5)$$

where $F_j^{-1}(p_i)$ is the j^{th} derivative of F^{-1} evaluated at p_i. The expectation of this expression may be rearranged into an expansion in inverse powers of n as in David and Johnson (1954) using the central moments of uniform order statistics, as

$$E\{X_{i:n}\} = \mu_{i:n} = F^{-1}(p_i) + \frac{p_i(1 - p_i)}{2(n + 2)}F_2^{-1}(p_i) + O(n^{-2}) \qquad (3.6)$$

which in turn yields for the spacings,

$$\begin{aligned} E\{D_i\} &= \mu_{i:n} - \mu_{i+1:n} \\ &= F^{-1}(p_i) - F^{-1}(p_{i+1}) + \\ &\quad [p_i(1-p_i)F_2^{-1}(p_i) - p_{i+1}(1-p_{i+1})F_2^{-1}(p_{i+1})]/(2(n+2)) + O(n^{-2}). \end{aligned} \tag{3.7}$$

Generally, letting $x = F^{-1}(p)$, one obtains for the first two derivatives,

$$\begin{aligned} F_1^{-1}(p) &= \frac{\partial F^{-1}(p)}{\partial p} = \frac{\partial x}{\partial p} = \frac{1}{\partial p/\partial x} \\ &= \frac{1}{f(x)} = \frac{1}{f\left(F^{-1}(p)\right)} \end{aligned} \tag{3.8}$$

and,

$$\begin{aligned} F_2^{-1}(p) &= \frac{\partial}{\partial p}\left(\frac{1}{f\left(F^{-1}(p)\right)}\right) \\ &= \left(\frac{-1}{f^2\left(F^{-1}(p)\right)}\right) \times \frac{\partial}{\partial p}\left(f(F^{-1}(p))\right) \\ &= \left(\frac{-1}{f^3\left(F^{-1}(p)\right)}\right) f'(F^{-1}(p)) \end{aligned} \tag{3.9}$$

Following ABN, we note that the derivatives appearing in (3.6) are readily evaluated for the Gaussian distribution, since $\frac{\partial}{\partial x}(\phi(x)) = -x\phi(x)$, giving

$$\mu_{i:n} = \Phi^{-1}(p_i)\{1 + \frac{p_i(1-p_i)}{2(n+2)} \times \frac{1}{[\phi(\Phi^{-1}(p_i))]^2}\} + O(n^{-2}). \tag{3.10}$$

It is easily seen that in general, to $O(n^{-1})$, for a fixed quantile (i.e. holding p_i fixed)

$$E\{D_i\} \simeq F^{-1}(p_i) - F^{-1}(p_{i+1}) \tag{3.11}$$

the change in the quantile function at p_i. However, it is also clear that the second term in (3.6) and (3.10) could be large. The reciprocal of the mean spacings, that is of the "weights" that would be applied to spacings to give them a common mean, is, to $O(n^{-1})$, linear for the exponential distribution, but convex for the Chi-square and concave for the Gaussian. These shapes explain why the spacings test using the asymptotic (linear) weights is conservative for Gaussian samples and liberal for Chi-square samples.

Further confirmation of the importance of the $O(n^{-1})$ term is provided by the three experiments using weights correct to $O(n^{-1})$ which follow. For these weights we need the second derivatives of the quantile function for the three distributions. We already have this for the Gaussian, while for the Exponential, the object required is

$$\frac{\partial}{\partial p}\left(\frac{1}{f\left(F^{-1}(p)\right)}\right) = \left(\frac{-1}{f^3\left(F^{-1}(p)\right)}\right) f'(F^{-1}(p)) \tag{3.12}$$

$$= (1-p)^{-2} \tag{3.13}$$

and is particularly simple because $F^{-1}(p) = \ln(1/(1-p))$ for this distribution. We thus obtain for the expectation,

$$E\{X_{i:n}\} = \ln(1/(1-p_i)) + \frac{p_i}{2(n+2)(1-p_i)} + O(n^{-2}). \qquad (3.14)$$

In passing, we note that for $i = 1, 2, 3$, a little calculation using (3.14) yields

$$\begin{array}{lcll} E(D_1) & \approx & \ln(2) + \frac{n+1}{4(n+2)} & \approx .94 \\ E(D_2) & \approx & \ln(3/2) + \frac{n+1}{12(n+2)} & \approx .49 \\ E(D_3) & \approx & \ln(4/3) + \frac{n+1}{48(n+2)} & \approx .31 \end{array}$$

so the linear decline of the expected spacings is apparent, but these numbers are not satisfyingly close to $1, 1/2, 1/3$ suggesting that the $O(n^{-2})$ term may be necessary for extreme order statistic spacings. This is unfortunately the case; adding the next term in the expansion yields the sequence, $1, .5, .33$, etc. In the experiments we use numerical derivatives for the Chi-square case.

Tables 8 to 10 are indicative, being based on only 1000 replications, and rounded to 2 significant digits. Figures in the body of the tables are empirical sizes; using 40 spacings, nominal size is thus .025.

Tables 8 – 10 about here

Evidently, the $O(n^{-1})$ term is important, since empirical size is approximately correct in the *(b)* columns, while the test is liberal in the *(a)* columns of Tables 6 to 8, where only the first order term in the expansion is used to generate the weights. The aberrant behaviour evident in Table 8 for sample size, 100, may arise from the truncation of the expansion at the $O(n^{-1})$ term.

We next consider whether it is practicable to choose the weights with sufficient accuracy using sample information. To do this we first consider estimation of the quantile function, then smoothing weighted spacings, neither of which proves satisfactory, then finally we adopt fixed weights appropriate to the Gaussian distribution, which work rather well.

Non-Parametric Quantile Function Estimates

We saw above that if the first two terms of the Taylor expansion are used in determining the weights, our 40-spacings test has essentially correct level for Gaussian, Chi-square, and Exponential shocks for samples of at least 200 observations. Though encouraging, this result is of little practical value unless the weights can be estimated from the sample.

This suggests that it might be useful to consider an approach based around an estimator of the quantile function. The simplest such estimator is the empirical quantile function,

$$\begin{array}{rcll} Q_n(p) & = & X_{n:n} & (0 \le p < 1/n) \\ & = & X_{j:n} & (1 - j/n \le p < 1 - (j-1)/n) \end{array} \qquad (3.15)$$

although this has the disadvantage that it is not differentiable. Conversely, it has the advantage, away from the tails, of a natural robustness to outliers: only the extreme upper quantile estimates will be affected by any outliers that may be present. Of course, the use we have in mind for the quantile estimates places a premium on accuracy in the tails: exactly the region in which most difficulty is likely to arise.

It appears to be standard practice to employ kernel estimators when empirical quantile functions are to be smoothed, and we implemented a recent proposal by Cheng and Peng (2002) combining kernel smoothing with local quadratic fitting. However, and contrary to the evidence presented by Cheng and Peng, the results were incredibly sensitive to choice of bandwidth. Moreover, we were unable to improve upon the asymptotic weights for any bandwidth. For that reason we do not report these results here.

An Expected Weighted Spacing Estimator

We also explored a simple one-sided estimator of the expected weighted spacing,

$$i\hat{D}_i = h^{-1} \sum_{j=i+1}^{i+h} jD_j \tag{3.16}$$

where the D_j are the observed spacings, and $h \to \infty$ at a suitable rate. The de-meaned weighted spacings were then defined as

$$\tilde{S}_i = iD_i - i\hat{D}_i \tag{3.17}$$

and a test constructed as follows.

For a test with asymptotic 10% significance level, declare that there is an outlier present if $\tilde{S}_1 > \tilde{S}_i \ \forall i \in (2, ..., 10)$, and at 5% significance level if $\tilde{S}_1 > \tilde{S}_i \ \forall i \in (2, ..., 20)$, and so on. Unfortunately, although this test worked reasonably well for non-Gaussian data, (3.16) did not adapt sufficiently to the lighter tailed Gaussian case, giving excessively conservative inference in that case no matter how we set the bandwidth, h. We decided, therefore, to construct a test using fixed weights that would be approximately correctly sized for Gaussian shocks, in the hope that it would have some robustness against centred $\chi^2_{(1)}$ shocks. It turns out that the size inflation, though significant, is within tolerable limits.

A Test Using Fixed Weighted Spacings

Because the expansion, (3.7), can be inaccurate for small n, (see again Table 8) we estimated the mean spacings of the largest 61 order statistics for the absolute values of draws from the Gaussian distribution by Monte Carlo simulation, using 10,000 replications of samples of lengths, $n = 100, 200, 500, 1000$ and 3000. Denoting these estimated mean spacings by $\bar{D}_i(n)$, $i = 1, ..., S = 60$, we constructed weights, $W_i(n) = \bar{D}_1(n)/\bar{D}_i(n)$, $i = 1, ..., 60$. For $n = 3000$ these weights are given by

$W(3000) = \{1.000, 0.531, 0.362, 0.280, 0.230, 0.193, 0.169, 0.147, 0.132, 0.123, 0.113, 0.104, 0.096, 0.088, 0.083, 0.080, 0.075, 0.071, 0.067, 0.065, 0.062, 0.060, 0.058, 0.056, 0.053, 0.052, 0.050, 0.048, 0.047, 0.046, 0.044, 0.042, 0.042, 0.040, 0.040, 0.039, 0.037, 0.037, 0.036, 0.035, 0.035, 0.034, 0.033, 0.032, 0.032, 0.031, 0.031, 0.031, 0.030, 0.029, 0.028, 0.028, 0.027, 0.027, 0.027, 0.026, 0.026, 0.025, 0.025, 0.025\}$.

There was in fact almost no difference between the $W_i(1000)$ weights and the corresponding $W_i(3000)$ weights given above. Moreover, although $W_i(3000) \neq W_i(\infty) = i^{-1}$, the convergence of the weights as n increases appears to be so slow that $W_i(3000)$ will remain appropriate, in practice, even for very large Gaussian samples.

Using these weights in Table 11, based on 2000 replications, we now report approximate empirical sizes of the following test procedure:

(i) compute the order statistics of $|X_i|$ and denote these in descending order of magnitude by $X_{i:n}$, $i = 1, ..., n$;
(ii) compute

$$S_i = (X_{i:n} - X_{i+1:n})/W_i(3000); \tag{3.18}$$

(iii) For a test with nominal level approximately 5%, reject the null hypothesis of no outliers in favour of the alternative of a single outlier if $S_1 = \max\{S_1, ..., S_{20}\}$.

Table 11 about here

Although slightly liberal for non-Gaussian shocks, a comparison of the results in Table 11 with those for the PR test in Table 1 show that our proposed outlier test performs dramatically better than that of PR.

3.2 The Serially Dependent Case

In the time series setting our attention now turns from testing for outliers directly in the observed sample $\{X_1, ..., X_n\}$ considered in Section 3.1 to the set-up considered in PR where we need to test for outliers in the $|\hat{x}_{\tau_j}|$ of (2.3).

As a first step in robustifying the PR procedure against departures from Gaussianity, suppose it is reasonable to assume that the data contain at most one additive outlier, and that the sample is large enough to permit calculation of 10, 20, 40, or 60 spacings. Let the upper order statistics of the $|\hat{x}_{\tau_j}|$ be denoted by $\{X_{1:n} \geq X_{2:n} \geq \cdots \geq X_{n:n}\}$, and again assume that the $|\hat{x}_{\tau_j}|$ are in the domain of attraction of the Gumbel distribution.

Next define the spacings,

$$D_{i,n} = X_{i:n} - X_{i+1,n},\ i = 1, ..., S,$$

with corresponding weighted standardised spacings as in (3.18). As we saw in Section 3.1, a test with level approximately 5% in the Normal case, and with level no more than 10% when shocks are centred $\chi^2_{(1)}$ is obtained when the $|\hat{x}_{\tau_j}|$ are independent with the decision rule: "reject the null hypothesis when $S_1 = \max\{S_1, ..., S_{20}\}$."

For the $|\hat{x}_{\tau_j}|$ to be independent requires that the process admits *two* unit roots but no other dynamics; that is, $\Delta^2 y_t = \epsilon_t$ with ϵ_t IID. In practice the $|\hat{x}_{\tau_j}|$ might be expected to be far from independent. To that end consider the correlation between successive $\hat{x}_{\tau_j}$. Under the null hypothesis,

$$\hat{x}_{\tau_j} = 0.5(\Delta y_{\tau_j} - \Delta y_{\tau_j+1})$$

so that if Δy has autocovariance generating function, $\Gamma_{\Delta y}(z)$, then the acgf of $\hat{x}$ is

$$\Gamma_{\hat{x}}(z) = 0.5(1-z)(1-z^{-1})\Gamma_{\Delta y}(z).$$

A few leading examples should help reveal the implications of the above. Writing $\rho_1(\hat{x})$ for the correlation between successive "outlier" estimates under the null, and $\Delta_z = (1-z)$ we obtain the results in Table 12.

Table 12 about here

More generally, we see that provided $\Delta^2 y_t$ is stationary, and ε_t is IID, the behaviour of the $\hat{x}_{\tau_j}$ under the null is fully described by the behaviour of the extreme values of the stationary ARMA process with autocovariance generating function, $\Gamma_{\hat{x}}(z)$. As mentioned above, the $\hat{x}_{\tau_j}$ are independent for the model with *two* unit roots, and so this is the case for which the simple order statistic-based test should have correct asymptotic level. Notice, however, that the base case used by PR for calibrating the critical values of their test is the pure random walk, for which $\rho_1(\hat{x}) = -0.5$; the test's robustness to different dynamics can be seen to arise from the fact that $\rho_1(\hat{x})$ is similar in the other rows of the table, *except* that for the double unit root model, for which the test is very conservative; cf. Table 2. We now illustrate the behaviour of the weighted spacings test, using decision rule C1, in this setting.

3.2.1 Numerical Evidence: One-step Test using 20 Spacings

Tables 13 and 14 report the results from a small experiment with the one-step spacings test using criterion C1 with weighted spacings calculated as in (3.18). The entries in the *Level* columns are the proportions of samples in which $S_1 = \max\{S_1, ..., S_{20}\}$ under the null, while the entries in the *Power* columns are the proportion of samples in which $S_j = \max\{S_1, ..., S_{20}\}$, $j = 1, 2, 3$, when in fact there is a single AO present equal to 5 for the $N(0, 1)$ distribution and 25 for the $\chi^2_{(1)}$ distribution. The entries in the columns for 2 and 3 outliers are thus false positives induced by the taking of double differences.

Tables13 − 14 about here

Observe that for Gaussian shocks the test levels are not much different from those for the IID case reported in Table 11. In particular, there is little difference between the results for the non-stationary and stationary cases. Conversely, for $\chi^2_{(1)}$ shocks there is less aliasing of one outlier by three for the stationary case, giving less conservative levels, but still very misleading "power".

To understand the remarkable differences between the results in Tables 13 and 14, consider the quantity on which the test is based, (2.3), for dates surrounding the true outlier date, taken to be $t = \tau$:

$$\begin{array}{lcl} \vdots & \vdots & \vdots \\ \Delta z_{\tau-1} & = & \Delta y_{\tau-1} \\ \Delta z_{\tau} & = & \Delta y_{\tau} + x_{\tau} \\ \Delta z_{\tau+1} & = & \Delta y_{\tau+1} - x_{\tau} \\ \Delta z_{\tau+2} & = & \Delta y_{\tau+2} \\ \vdots & \vdots & \vdots \end{array} \tag{3.19}$$

giving, under the pure random walk,

$$\begin{aligned} \hat{x}_{\tau-1} &= (\Delta z_{\tau-1} - \Delta z_{\tau})/2 = (\varepsilon_{\tau-1} - \varepsilon_{\tau} - x_{\tau})/2 \\ \hat{x}_{\tau} &= (\Delta z_{\tau} - \Delta z_{\tau+1})/2 = x_{\tau} - \Delta\varepsilon_{\tau+1}/2 \\ \hat{x}_{\tau+1} &= (\Delta z_{\tau+1} - \Delta z_{\tau+2})/2 = (\varepsilon_{\tau+1} - x_{\tau} - \varepsilon_{\tau+2})/2 \end{aligned} \tag{3.20}$$

Evidently, if there is an outlier, and x_{τ} is very large, then it is likely that there will be not one, but *three* apparent outliers in the order statistics of $|\hat{x}|$. We term this phenomenon *aliasing*. Also, if the largest of these three were removed (this being the central one, most likely), then the other two would vanish. Worse, the spacings test might not detect the largest outlier, if only one outlier is tested for, because the first spacing may not be twice as big as the second, let alone three times as big as the third. This is what happens most obviously with the centred $\chi^2_{(1)}$ shocks. We now present a simple practical solution to this problem.

3.2.2 Numerical Evidence: Sequential Test using 60 Spacings

The algorithm we propose for detecting a single outlier, avoiding the aliasing problem outlined above, and with nominal significance level 5% is as follows:

Algorithm 1 *(1) Using the $W(3000)$ set of weights given in Section 3.1, check if any of the first three standardised spacings is the largest of the first 60 spacings; if not, stop.(2) Dummy out the largest "outlier" by subtracting $|\hat{x}_{\tau}|$ from the observation dated, τ.(3) Repeat until none of the first three spacings is the largest, or three outliers have been found.(4) The number of outliers is the number dummied out.*

Tables15 – 16 about here

Tables 15 and 16 reveal that this algorithm performs rather well under the null, and in the presence of a single large outlier, in both stationary and non-stationary data. Of particular significance are the facts that the aliasing problem has virtually disappeared, and that the test significance level is around 10-12% for the centred $\chi^2_{(1)}$ shocks, compared to levels ranging from 66% to 99% for the Perron-Rodriguez test reported in Table 1.

Although we have restricted attention in this section to the case where we wish to test against a single outlier allowing for the problem of aliasing in the time series setting, the approach can be easily extended to the case where we wish to test against k outliers by choosing a correspondingly larger number of standardised spacings.

4 Conclusions

When the series under study is driven by Gaussian shocks, an advantage of the PR test is that any putative outliers are referred to critical values that are unaffected by other sample characteristics. If more than one outlier is present, then there is an impact on the denominator of the PR statistic, but this will not generally be very great provided outliers are not consecutive, so the risk of one outlier masking another is not too large. Further, the sequential method of searching for multiple outliers ensures that no aliasing occurs: we test for a second outlier only after removal of the first, and so on. The test has impressive power, equal to unity for the outliers used in our experiments. However, when the shocks are non-Gaussian, their test has uncontrolled significance level. In Algorithm 1, we propose a method based on examining the weighted spacings between the order statistics of the outlier estimate implicit in the PR test, and show that size inflation can be greatly reduced, though not entirely eliminated. Our test is somewhat less powerful than the PR test, but some small power reduction is inevitable when the unrealistic requirement that the shocks are Gaussian is abandoned. We explored methods of data-based correction of significance level, but found that these were ineffective in practice, and so opted for a compromise in which the spacings are weighted *as if the sample is large, and Gaussian* but as we show, the resulting test is much more robust than the PR test from which it is developed. In a large sample, our test can be conducted at any desired significance level simply by varying the number of spacings compared. In a large sample, if more than 60 spacings were to be compared, suitable weights could be computed using (3.7) for the normal distribution. We therefore recommend the adoption of our outlier detection procedure in preference to that suggested in PR.

Acknowledgements: we are grateful for the comments of Jose Olmo, and participants at the SNDE conference, City University, London, April 2005.

References

[1] Arnold, B.C., N. Balakrishnan, and H.N. Nagaraja, 1992, A First Course in Order Statistics, John Wiley & Sons, Inc., New York.

[2] Basmann, R.L., 2003, Statistical outlier analysis in litigation support: the case of Paul F.Engler and Cactus Feeders, Inc., vs. Oprah Winfrey et al., Journal of Econometrics 113, 159-200.

[3] Cheng, M.Y. and L. Peng, 2002, Regression modeling for nonparametric estimation of distribution and quantile functions, Statistica Sinica 12, 1043-1060.

[4] David, F.N. and N.L. Johnson, 1954, Statistical treatment of censored data, Part 1. Fundamental formulae, Biometrika 41 228-240.

[5] Fox, A.J., 1972, Outliers in time series, JRSS(B) 34, 350-363.

[6] Harvey, D.I., S.J.Leybourne and P.Newbold, 2001, Innovational outlier unit root tests with an endogenously determined break in level, Oxford Bulletin of Economics and Statistics 63, 559-575.

[7] Mood, A.M., F.A. Graybill and D.C. Boes, 1974, Introduction to the Theory of Statistics, (3rd edition), McGraw-Hill.

[8] Perron, P. and G. Rodriguez, 2003, Searching for Additive Outliers in Non-stationary Time Series, Journal of Time Series Analysis 24, 193-220.

[9] Weissman, I., 1978, Estimation of parameters and larger quantiles based on the k largest observations, Journal of the American Statistical Association 73, 812-815.

Table 1: Empirical level of PR test in null model with non-Gaussian shocks.

	Sample size			
Distribution	100	200	500	1000
Normal	.05	.05	.05	.05
Exponential	.46	.61	.82	.93
Centred $\chi^2_{(1)}$	.66	.83	.97	.99
Student-t(5)	.33	.45	.65	.82
Student-t(6)	.25	.35	.54	.70
Student-t(10)	.14	.21	.27	.39
Student-t(50)	.075	.068	.075	.088

Table 2: Empirical level of PR test, DGP (2.1)-(2.2). Nominal level 5%.

			Sample size			
ϕ	ψ	θ	100	200	500	1000
1	0	0	.05	.05	.05	.05
1	1	0	.005	.006	.007	.01
0	0	0	.054	.053	.051	.055
.8	0	0	.055	.049	.052	.053
.8	1	0	.026	.034	.043	.051
0	0	1	.057	.047	.048	.049

Table 3: Normal shocks: $S = 40$, $K = 3$. Asymptotic weights.

		number of outliers			
sample size	rule	0	1	2	3
50	C1	.9957	.0000	.0012	.0031
100	C1	.9904	.0016	.0031	.0049
200	C1	.9867	.0021	.0037	.0075
500	C1	.9791	.0048	.0069	.0092
1000	C1	.9730	.0050	.0080	.0140
3000	C1	.9600	.0100	.0130	.0170

Table 4: Chi square shocks: $S = 40$, $K = 3$. Asymptotic weights.

		number of outliers			
sample size	rule	0	1	2	3
50	C1	.7206	.0969	.0956	.0869
100	C1	.7643	.0814	.0795	.0748
200	C1	.8756	.0417	.0403	.0424
500	C1	.9038	.0338	.0315	.0309
1000	C1	.9150	.0290	.0250	.0310
3000	C1	.9070	.0360	.0270	.0300

Table 5: Exponential shocks: $S = 40$, $K = 3$. Asymptotic weights.

		number of outliers			
sample size	rule	0	1	2	3
50	C1	.9312	.0239	.0213	.0236
100	C1	.9263	.0232	.0255	.0250
200	C1	.9277	.0234	.0236	.0253
500	C1	.9242	.0247	.0267	.0244

Table 6: Normal shocks: $S = 40$, $K = 3$.
Asymptotic weighting. One outlier present equal to 5.0

		number of outliers			
sample size	rule	0	1	2	3
50	C1	.72	.27	.008	.005
100	C1	.55	.41	.02	.02
200	C1	.48	.47	.03	.02
500	C1	.49	.46	.03	.02
1000	C1	.45	.49	.04	.01
3000	C1	.53	.41	.04	.02

Table 7: Chi square shocks: $S = 40$, $K = 3$.
Asymptotic weights. One outlier present equal to 25

		number of outliers			
sample size	rule	0	1	2	3
50	C1	.004	.975	.017	.004
100	C1	.005	.964	.026	.005
200	C1	.010	.942	.035	.013
500	C1	.033	.912	.045	.010
1000	C1	.048	.875	.066	.011
3000	C1	.098	.813	.069	.020

Table 8: Empirical significance levels: $S = 40$, $K = 3$.
(a) = Weights using $E\{D_i\} \approx F^{-1}(p_i) - F^{-1}(p_{i+1})$.
(b) = Weights using (3.7)

Gaussian		number of outliers							
sample size	rule	0		1		2		3	
		(a)	*(b)*	*(a)*	*(b)*	*(a)*	*(b)*	*(a)*	*(b)*
50	C1	.78	**.85**	.09	**.05**	.08	**.06**	.05	**.04**
100	C1	.85	**.89**	.05	**.02**	.05	**.04**	.05	**.05**
200	C1	.85	**.90**	.05	**.03**	.05	**.03**	.05	**.04**
500	C1	.86	**.91**	.06	**.03**	.05	**.03**	.03	**.03**
1000	C1	.85	**.91**	.07	**.03**	.05	**.03**	.03	**.03**
3000	C1	.87	**.91**	.06	**.03**	.05	**.03**	.02	**.03**

Table 9: Empirical significance levels: $S = 40$, $K = 3$.
(a) = Weights using $E\{D_i\} \approx F^{-1}(p_i) - F^{-1}(p_{i+1})$.
(b) = Weights using (3.7)

Exponential		number of outliers							
sample size	rule	0		1		2		3	
		(a)	*(b)*	*(a)*	*(b)*	*(a)*	*(b)*	*(a)*	*(b)*
50	C1	.86	**.93**	.07	**.03**	.04	**.02**	.03	**.02**
100	C1	.86	**.92**	.06	**.03**	.04	**.02**	.04	**.03**
200	C1	.84	**.91**	.08	**.04**	.04	**.02**	.04	**.03**
500	C1	.84	**.91**	.08	**.04**	.03	**.02**	.04	**.03**
1000	C1	.86	**.92**	.06	**.03**	.04	**.03**	.04	**.02**
3000	C1	.86	**.91**	.06	**.03**	.04	**.03**	.04	**.03**

Table 10: Empirical significance levels: $S = 40$, $K = 3$.
(a) = Weights using $E\{D_i\} \approx F^{-1}(p_i) - F^{-1}(p_{i+1})$.
(b) = Weights using (3.7)

Centred $\chi^2_{(1)}$		number of outliers							
sample size	rule	0		1		2		3	
		(a)	*(b)*	*(a)*	*(b)*	*(a)*	*(b)*	*(a)*	*(b)*
50	C1	.89	**.93**	.05	**.02**	.04	**.03**	.02	**.02**
100	C1	.72	**.79**	.10	**.06**	.08	**.07**	.09	**.08**
200	C1	.84	**.91**	.07	**.03**	.05	**.03**	.05	**.03**
500	C1	.85	**.92**	.08	**.03**	.05	**.03**	.03	**.02**
1000	C1	.84	**.91**	.08	**.04**	.04	**.03**	.04	**.02**
3000	C1	.86	**.91**	.06	**.03**	.05	**.03**	.03	**.02**

Table 11: Empirical level of a fixed weighted spacings test.
IID sample, nominal level 5%.

Sample size, $n =$	100	200	500	1000	3000
Shock Type					
$N(0,1)$	.018	.027	.032	.047	.042
centred $\chi^2_{(1)}$	.102	.090	.079	.093	.078

Table 12: Correlation of outlier estimates under the null.

y process	$\Gamma_{\Delta y}(z)$	$\Gamma_{\hat{x}}(z)$	$\rho_1(\hat{x})$
$\Delta y_t = \varepsilon_t$	σ^2_ε	$\frac{1}{2}\sigma^2_\varepsilon \Delta_z \Delta_{z^{-1}}$	-0.5
$y_t = \varepsilon_t$	$\sigma^2_\varepsilon \Delta_z \Delta_{z^{-1}}$	$\frac{1}{2}\sigma^2_\varepsilon \{\Delta_z \Delta_{z^{-1}}\}^2$	$\frac{-2}{3}$
$\Delta^2 y_t = \varepsilon_t$		$\frac{1}{2}\sigma^2_\varepsilon$	0
$y_t = \Delta\varepsilon_t$	$\sigma^2_\varepsilon \{\Delta_z \Delta_{z^{-1}}\}^2$	$\frac{1}{2}\sigma^2_\varepsilon \{\Delta_z \Delta_{z^{-1}}\}^3$	$\frac{-3}{4}$
$y_t - .8y_{t-1} = \varepsilon_t$	$\frac{\sigma^2_\varepsilon \Delta_z \Delta_{z-1}}{(1-.8z)(1-.8z^{-1})}$	$\frac{1}{2}\frac{\sigma^2_\varepsilon \{\Delta_z \Delta_{z-1} \Delta_z \Delta_{z-1}\}^2}{(1-.8z)(1-.8z^{-1})}$	$-.51$

Table 13: Empirical level and power of C1 test with nominal level 5%. $S = 20$.

$N(0,1)$	$\Delta y_t = \varepsilon_t$				$y_t - .8y_{t-1} = \varepsilon_t$			
	Level	*Power*[a]			*Level*	*Power*[a]		
n		1	2	3		1	2	3
100	.021	.528	.208	.177	.020	.532	.212	.125
200	.024	.647	.184	.114	.025	.640	.181	.079
500	.035	.719	.166	.076	.033	.714	.158	.051
1000	.049	.782	.141	.043	.045	.767	.126	.037
3000	.052	.835	.118	.023	.044	.803	.102	.022

Table 14: Empirical level and power of C1 test with nominal level 5%. $S = 20$.

$\chi^2_{(1)} - 1$	$\Delta y_t = \varepsilon_t$				$y_t - .8y_{t-1} = \varepsilon_t$			
	Level	*Power*[a]			*Level*	*Power*[a]		
n		1	2	3		1	2	3
100	.009	.007	.001	.979	.012	.019	.004	.959
200	.008	.008	.002	.968	.014	.041	.004	.929
500	.009	.031	.002	.934	.018	.115	.005	.853
1000	.007	.073	.006	.884	.020	.213	.009	.749
3000	.008	.203	.002	.763	.020	.483	.003	.491

Table 15: Empirical level and power C3 test with nominal level 5%. $S = 60$.

$N(0,1)$	$\Delta y_t = \varepsilon_t$				$y_t - .8y_{t-1} = \varepsilon_t$			
	Level	*Power*[a]			*Level*	*Power*[a]		
n		1	2	3		1	2	3
100	.007	.794	.002	.000	.011	.693	.004	.001
200	.012	.897	.007	.000	.015	.807	.006	.001
500	.021	.915	.016	.001	.022	.851	.017	.000
1000	.034	.914	.029	.001	.033	.864	.024	.003
3000	.043	.920	.039	.002	.040	.852	.033	.003

Table 16: Empirical level and power C3 test with nominal level 5%. $S = 60$.

$\chi^2_{(1)} - 1$	$\Delta y_t = \varepsilon_t$				$y_t - .8y_{t-1} = \varepsilon_t$			
	Level	*Power*[a]			*Level*	*Power*[a]		
n		1	2	3		1	2	3
100	.128	.863	.121	.003	.119	.871	.106	.004
200	.132	.861	.116	.001	.113	.879	.092	.002
500	.127	.866	.099	.001	.093	.900	.070	.002
1000	.121	.867	.093	.002	.099	.892	.076	.003
3000	.124	.867	.097	.003	.090	.900	.075	.002