Outliers and Persistence in Threshold Autoregressive Processes: A Puzzle?

By

Yamin Ahmad
(UW-Whitewater)

and

Luiggi Donayre
(University of Minnesota - Duluth)

Working Paper 14 - 02

University of Wisconsin – Whitewater
Department of Economics
4th Floor Hyland Hall
800 W. Main Street
Whitewater, WI 53190

Tel: (262) 472 -1361

Outliers and Persistence in Threshold Autoregressive Processes: A Puzzle?[*]

Yamin Ahmad[†]
Department of Economics
University of Wisconsin-Whitewater

Luiggi Donayre[‡]
Department of Economics
University of Minnesota-Duluth

This version: March 17th, 2014

Abstract

We conduct Monte Carlo simulations to investigate the effects of outlier observations on the properties of linearity tests against threshold autoregressive (TAR) processes. By considering different specifications and levels of persistence of the data generating processes, we find that outliers distort the size of the test and that the distortion increases with the level of persistence. However, contrary to what one might expect, we also find that larger outliers could help improve the power of the test in the case of persistent TAR processes.

JEL Classification Code: C15, C22

Keywords: Outliers, Persistence, Monte Carlo Simulations, Threshold Autoregression, Size, Power.

[*] All errors are our own.

[†] Department of Economics, University of Wisconsin - Whitewater, 800 W Main Street, Whitewater, WI 53190. Email: ahmady@uww.edu

[‡] *Corresponding author.* Department of Economics, University of Minnesota - Duluth, 1318 Kirby Dr., Duluth, MN 55812. Email: adonayre@d.umn.edu

1 Introduction

An important area of research in recent theoretical and empirical time series analysis has focused on potential nonlinear features of macroeconomic relationships using threshold autoregressive (TAR) specifications. Much of the impetus for this interest stems from a large number of studies that have found evidence of asymmetric behavior in such relationships that have led to forecasts that depend on the state of certain economic conditions (Potter, 1995; Galbraith, 1996; Pesaran and Potter, 1997; Koop and Potter, 2004; Juvenal and Taylor, 2008; Gonzalo and Pitarakis, 2013).[1] While many of these studies have found extensive support for nonlinearity in many macroeconomic time series, the evidence is, however, not overwhelming (see for example, Enders, Falk, and Siklos (2007)).

In this environment, it is important to recognize that this evidence for nonlinearity could result from distortions introduced by the presence of outlier observations. For example, most macroeconomic aggregates are subject to substantial variation, given changes in economic conditions, financial or political crises and other isolated disturbances. These turbulent histories may appear as outliers in those time series and, in the presence of some atypical observations, statistical tests may incorrectly reject linear specifications for the data-generating process (DGP). At the same time, the sample size of most macroeconomic aggregates is relatively short because they are usually sampled quarterly or annually. It has been argued by van Dijk, Franses, and Lucas (1999), Ahmad (2008) and López Villavicencio (2008) that the nonlinear properties of the relevant series may only be reflected in a few observations. A researcher may view these nonlinear data points as atypical observations and remove them, thus destroying intrinsic nonlinearity (van Dijk et al., 1999; López Villavicencio, 2008).[2]

In this paper, we investigate the effects of outliers on regular testing procedures for TAR processes (Hansen, 1996, 1997) by means of Monte Carlo simulations. Our objective is three-fold. First, we evaluate the performance of linearity tests for autoregressive (AR) and TAR processes when outliers are absent. Second, we introduce outliers into the DGP and examine how the presence of these outlier observations influences the results for different sample sizes, degrees of contamination and magnitudes of the outlier observations. Finally, we analyze the link between outliers and the degree of persistence in AR and TAR processes, focusing specifically on how they affect the size and power of the tests.

Addressing these issues has important implications for the time series literature. Incorrectly assuming linearity in macroeconomic relationships could lead to misleading inferences, policy implications and out-of-sample forecasts. Therefore, if the presence of outliers distorts the identification of linear vis-à-vis

[1] For a comprehensive review of TAR applications in Economics, see Hansen (2011) and Tong (2011).

[2] Moreover, Ahmad (2008) argues that even if the underlying data generating process is nonlinear, the presence of nonlinearity in a small number of observations leads to small sample bias when estimating parameters.

nonlinear properties of time series, then researchers should exert caution when making inferences about them. Meanwhile, a number of macroeconomic time series exhibit a high degree of persistence. Given that such persistence has important economic implications, understanding how linearity tests perform in the presence of outliers for persistent processes becomes relevant.[3] Moreover, if the nonlinear behavior of the series is only reflected in a small number of observations, then persistent processes may make the identification and estimation of the process easier since the probability that the process remains in the alternative regime is higher.

Our paper is similar in spirit to Koop and Potter (2001), who propose a Bayesian model comparison approach to determine whether departures from linearity in macroeconomic aggregates are endogenously generated, as in a TAR process, or whether they merely reflect structural instabilities that linear fixed parameter models cannot account for. They use their approach to investigate the presence of nonlinearities in several artificial and real data series and find little support for threshold-type nonlinearities alone. It is also related to the work of van Dijk et al. (1999), who study whether apparent evidence of nonlinearity in smooth transition autoregressive (STAR) processes is due to the presence of outlier observations in economic time series. They propose tests for smooth transition nonlinearity that are robust to the presence of outliers and conduct Monte Carlo experiments to evalute the performance of such tests. Their results show that the proposed tests have a better level and power behavior than standard nonrobust tests in series contaminated with outliers.

By contrast, our study differs from the existing literature in that, to the best of our knowledge, it is the only one that specifically explores the effects of outliers on TAR processes. More importantly, we also investigate the extent to which the degree of persistence affects the size and power of Hansen's bootstrap linearity test. Our results show that the size of the test is affected by the presence of outliers only in the case of persistent processes. The size distortion increases monotonically with the degree of persistence and the sample size. Meanwhile, the presence of outliers marginally reduces the power of the test, particularly in the kinds of sample sizes common for macroeconomic analysis. Notably, we find that the power of the test increases with both the degree of persistence, and the magnitude of the outlier, even in small samples and for small threshold effects. This implies a puzzling result: larger outlier observations in persistent TAR processes could actually help to improve the performance of statistical tests.

The rest of the document is organized in the following way. In the next section, we discuss how to

[3] For example, the persistence of inflation is a major determinant of the economic costs of disinflation. Similarly, a change in the persistence of cyclical employment in recent recessions could explain the so-called 'jobless recoveries' phenomenon.

define outlier observations and provide a motivation for how the degree of persistence could affect the linear and TAR processes. Section 3 describes the Monte Carlo experiments and presents the results. We provide some concluding remarks in section 4.

2 Outliers in Economic Time Series

Several empirical studies have found evidence of nonlinearities in different macroeconomic series or relationships using TAR models and their different extensions. For example, different authors suggest threshold-type nonlinearities in business cycle dynamics (Beaudry and Koop, 1993; Potter, 1995; Pesaran and Potter, 1997), unemployment (Koop and Potter, 2004), interest rates (Tsay, 1998; Hansen and Seo, 2002; Gospodinov, 2005), financial conditions (Galbraith, 1996; Balke, 2000; Atanasova, 2003), the effects of monetary policy on output (Sander and Kleimeier, 2004; Donayre, 2014), the effects of fiscal policy on output (Auerbach and Gorodnichenko, 2012, 2013; Fazzari, Morley, and Panovska, 2013), and exchange rates (Taylor, 2001; Bec, Ben-Salem, and Carrasco, 2004; Sarno, Taylor, and Chowdhury, 2004; Juvenal and Taylor, 2008).

Whilst the presence of structural nonlinearities are supported by many economic theories, the majority of studies neglect the possibility that such variables could be contaminated by occasional outliers, as a consequence of a change in regimes, political or economic disturbances. One exception is given by Ahmad and Glosser (2011), who find evidence that nonlinear behavior is reflected in a small number of outlier observations in the context of the real exchange rates that they examine. They find that outlier observations can distort linear relationships and cause the linearity test to reject the correct null hypothesis of linearity too often. That is, apparent nonlinearity found in the empirical literature could be the consequence of outliers observations in linear processes. Since linear and nonlinear models have different implications for the characterization and forecasting of real macroeconomic variables, as well as for policymaking, it is important to examine whether certain features of the series are caused by genuine nonlinearity or by some outlier observations.

Outliers are defined in relation to a specific model. That is, some observations can be considered as outliers in one model and, at the same time, they can be regular observations in a different model (van Dijk et al., 1999). Typically, two types of outliers have been especially important in time series analysis. The additive outlier (AO) generates a one time effect on the level of the time series and, thus, only the current observation is affected. The innovative outlier (IO), in turn, implies that a shock at time t also influences future observations through the same dynamics as the linear part of the process.

The approach we adopt to define outliers in this paper builds on the replacement model of Martin and Yohai (1986):

$$y_t = x_t(1 - \delta_t) + \zeta_t \delta_t \tag{1}$$

for $t = 1, \ldots, T$, where T denotes the sample size and δ_t is a binary random variable such that

$$\delta_t = \begin{cases} 1 & \text{with probability } \pi \\ 0 & \text{otherwise} \end{cases}$$

The observed time series y_t consists of a *core* process, x_t, and a *contaminating* process, ζ_t. In the case of a linear process, we assume that x_t follows an autoregression (AR) of order p; that is, $\phi(L)x_t = \epsilon_t$, where $\phi(L) = 1 - \phi_1 L - \ldots - \phi_p L^p$ is a polynomial in the lag operator L defined as $L^j x_j = x_{t-j}$ for all j, and where $\epsilon_t \sim iid\ (0, \sigma_\epsilon^2)$. Different specifications of the ζ_t process can generate a wide variety of outlier patterns.

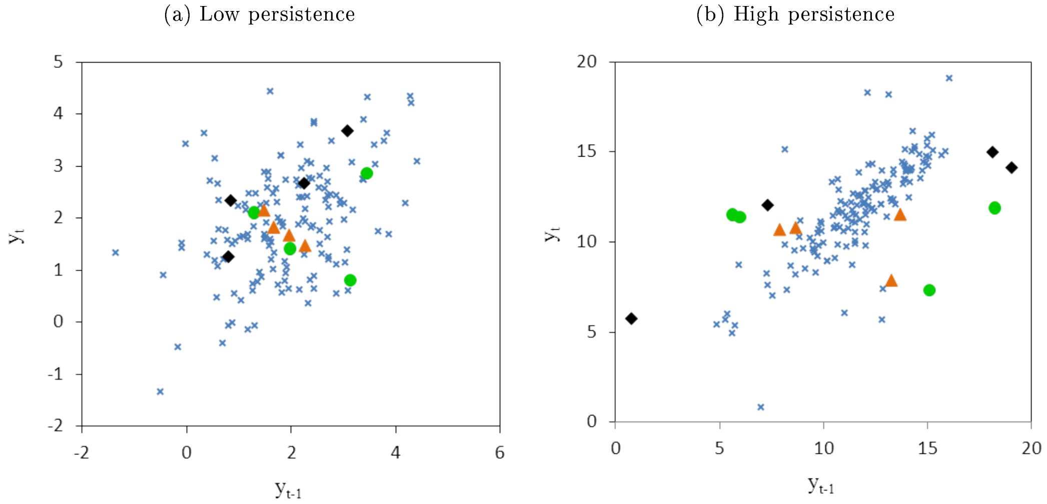
While it has been shown that the presence of both AO and IO can distort the results of testing and estimation, additive outliers have a much more significant effect on ordinary least squares (OLS) estimates (van Dijk et al., 1999; van Dijk, Teräsvirta, and Franses, 2002). Since the estimation and testing of TAR processes relies on an iterative procedure of OLS estimates for different values of the threshold parameter, we focus on outliers of the additive type in this paper. An additive outlier is obtained if $\zeta_t = x_t + \zeta$ for some constant ζ, such that (1) reduces to

$$y_t = x_t + \zeta \delta_t \tag{2}$$

A vast literature exists that highlights different approaches to detecting outliers, and procedures for correcting for the presence of outliers. For the purposes of this paper, equations (1) and (2) are all we need to understand the effect of outlier observations on TAR processes. For a review of the literature on outliers, we refer the reader to van Dijk et al. (1999) and Lucas, Franses, and van Dijk (2002).

To gain some insight regarding the possible effects of outlier observations on AR processes, and how they might depend on the degree of persistence, we simulate both linear AR and TAR processes to gain some insight into how outliers may influence the cloud of observations. Figure 1 displays two AR(1) processes. The left panel of figure 1 exhibits an AR(1) process with a low degree of persistence ($\phi = 0.4$) while the right panel of figure 1 exhibits an AR(1) process with a high degree of persistence ($\phi = 0.9$). In both cases, the DGP is the same, except for the autoregressive coefficient, and outliers have been introduced according to equation (2) for $\zeta = \sigma$, 2σ, 3σ, where σ is the standard deviation of the given process. For visual purposes, outlier observations are depicted in orange triangles for $\zeta = \sigma$, black rhombi for $\zeta = 2\sigma$ and green circles for $\zeta = 3\sigma$.

Figure 1: Outliers in AR processes


(a) Low persistence

(b) High persistence

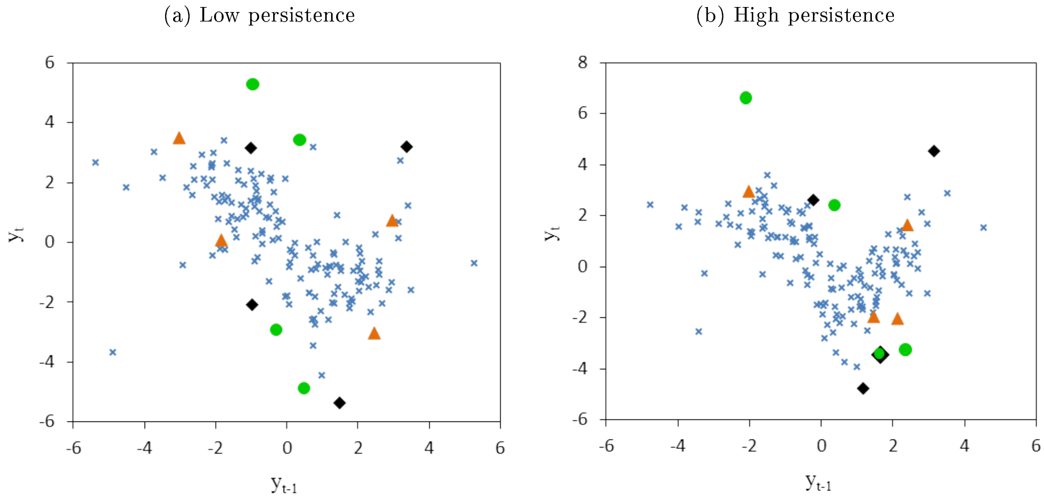
Artificial AR processes generated according to $y_t = 1.2 + \phi y_{t-1} + e_t$, where $e_t \sim N(0,1)$ for values of $\phi = 0.4$ and $\phi = 0.9$. Outliers are depicted in orange triangles for $\zeta = \sigma$, black rhombi for $\zeta = 2\sigma$ and green circles for $\zeta = 3\sigma$ according to equation (2).

As it can be observed from figure 1, the cloud of points is more scattered in the less persistent AR process. As a consequence, the outlier observations seem to blend it with those from the core process x_t. In this sense, they are less evident than in the case of the more persistent AR process, even for large outlier observations (*e.g.*, green circles associated with $\zeta = 3\sigma$). In contrast, outlier observations are more likely to be further away from the cloud of points given a higher degree of persistence (see right panel of figure 1). Therefore, they are more likely to distort the results of linearity tests. As the right panel suggests, the distortion seems to increase with the size of the outlier observations, which is to be expected.

Figure 2, on the other hand, displays two TAR(1) processes generated using the same DGP, except that they differ in their degree of persistence. Outliers have been introduced according to equation (2) for $\zeta = \sigma$, 2σ, 3σ and they are shown in the same colors and patterns as figure 1. The left panel of figure 2 exhibits a TAR process with a low degree of persistence while the right panel of figure 2 exhibits a TAR(1) process with a high degree of persistence.

From just observing the figures above, we note that the outlier observations in the less persistent TAR process seem to blend in within the cloud of points, even in the case of large outliers associated

Figure 2: Outliers in TAR processes



(a) Low persistence

(b) High persistence

Artificial TAR processes generated according to $y_t = (0.7 - 0.5y_{t-1})I[y_{t-1} \leq 0] + (-1.8 + \phi y_{t-1})I[y_{t-1} > 0] + e_t$, where $e_t \sim N(0, 1)$ for values of $\phi = 0.4$ and $\phi = 0.9$. Outliers are depicted in orange triangles for $\zeta = \sigma$, black rhombi for $\zeta = 2\sigma$ and green circles for $\zeta = 3\sigma$ according to equation (2).

with $\zeta = 3\sigma$ (in green circles). Meanwhile, those in the more persistent TAR process, in the right panel of figure 2, lie further away from the cloud of points, making it easier to visually identify the two regimes. Whilst the evidence here is anecdotal, the data in figure 2 would suggest that larger outlier observations may allow the identification of a change in regimes, thus helping to increase the power of linearity tests.

3 Monte Carlo Analysis

In this section, we evaluate the effects of additive outliers on linearity tests against TAR processes for different specifications of several, artificial time series. Section 3.1 describes the Monte Carlo experiment and discusses the linear and threshold autoregressive processes considered, as well as the bootstrap test for linearity against TAR processes (Hansen, 1996). In section 3.2, we investigate the performance of the test in a setting in which the data generating process (DGP) is linear. In the third subsection, we evaluate the relative performance of the test when the DGP follows a TAR process.

We consider the cases where the AR and TAR processes exhibit different degrees of contamination with outlier observations, as well as the case where they are not contaminated. Furthermore, we also examine the behavior of the linearity test for different degrees of persistence for both linear and TAR processes. In all case we examine, we set the sample size to $T = 40, 80, 160, 320, 640$. Outliers occur with probability π, and we set their magnitude equal to $\zeta = 0, 1, 2, 3$ standard deviations. We consider symmetric contamination in the sense that outliers are equally likely to be positive or negative. Hence, δ_t takes the values 1, 0, -1 with probabilities $\pi/2$, $1 - \pi$ and $\pi/2$, respectively. Finally, we consider 1,000 simulations for each case we examine. Although we examine different values of π, we present the results in the paper for the case of $\pi = 0.05$.[4]

3.1 Core processes and linearity test

In the first set Monte Carlo experiments, we consider a linear DGP as in equation (2), for which the core process x_t is generated from a pth-order autoregression according to:

$$x_t = \alpha + \sum_{j=1}^{p} \phi_j x_{t-j} + \epsilon_t \tag{3}$$

where $\epsilon_t \sim NID\ (0, \sigma_\epsilon^2)$, $\sigma_\epsilon = 1$, α is the intercept term and all roots of $\phi(L)$ lie outside the unit circle. To address the power of the linearity test, the AR(p) process for x_t is replaced by a TAR(p) one according to:

$$x_t = \left(\alpha_1 + \sum_{j=1}^{p} \phi_{j,1} x_{t-j}\right) I(s_t \leq \gamma) + \left(\alpha_2 + \sum_{j=1}^{p} \phi_{j,2} x_{t-j}\right) I(s_t > \gamma) + \epsilon_t \tag{4}$$

where $I(.)$ denotes the indicator function; α_1 and α_2 correspond to the intercepts in regimes 1 and 2, respectively; the roots of $\phi_1(L)$ and $\phi_2(L)$ lie outside the unit circle; s_t is the threshold variable; γ is the unknown threshold parameter; and $\epsilon_t \sim NID\ (0, \sigma_\epsilon^2)$, $\sigma_\epsilon = 1$. In this way, when $s_t \leq \gamma$, the dynamics of the series is captured by the $1 \times (p+1)$ vector of coefficients, $(\alpha_1 \quad \phi_1{}')'$, and when $s_t > \gamma$, it is captured by the alternative $1 \times (p+1)$ vector of coefficients, $(\alpha_2 \quad \phi_2{}')'$. For further details on TAR processes, refer to Hansen (1996, 1997).

In all cases, the starting value x_0 is set equal to 0. To eliminate possible dependencies of the results on this intial condition, the first 200 observations of each series are discarded. Furthermore, we obtain the contaminated series y_t by adding AOs according to (2). In our experiments, we consider the effects of varying the persistence parameters in the AR and TAR models, as well as the sample size T, and the magnitude of the outliers, ζ. In practice, the econometrician has to decide on the order p of the linear

[4] Additional results are available upon request from the authors.

AR(p) and the TAR(p) models. To reduce the computational burden, since our choice here is purely for expositional purposes, we focus on the simplest case and set $p = 1$. Hence, equation (3) reduces to:

$$x_t = \alpha + \phi_1 x_{t-1} + \epsilon_t \tag{5}$$

with $|\phi_1| < 1$, while equation (4) simplifies to:

$$x_t = (\alpha_1 + \phi_{1,1} x_{t-1}) I(s_t \leq \gamma) + (\alpha_2 + \phi_{1,2} x_{t-1}) I(s_t > \gamma) + \epsilon_t \tag{6}$$

with $|\phi_{1,1}| < 1$ and $|\phi_{1,2}| < 1$. When testing linearity against a TAR process, the relevant null hypothesis is given by $H_0 : \phi_{1,1} = \phi_{1,2}$. Thus, the linearity test is based on a bootstrap procedure to approximate the asymptotic distribution of the likelihood ratio (LR) statistic:

$$LR = \underset{\gamma \in \Gamma}{sup} \{LR_n(\gamma)\} \tag{7}$$

where γ is assumed to be restricted to a bounded set[5] $\Gamma = \left[\underline{\gamma}, \overline{\gamma}\right]$ and

$$LR_n(\gamma) = 2\left[lnf(Y|\hat{\gamma}) - lnf(Y|\gamma)\right] \tag{8}$$

is the LR statistic against the alternative $H_1 : \phi_{1,1} \neq \phi_{1,2}$ when γ is known, $f(Y|\gamma)$ corresponds to the values of the likelihood function for each γ, and $\hat{\gamma}$ is estimate of the threshold parameter, considering a general process such as the one given by equation (2).

Since γ is not identified under the null hypothesis, the asymptotic distribution of (8) is non-standard. Hansen (1996) shows that the asymptotic distribution may be approximated by the following bootstrap procedure for a TAR process like that in (6). Let u_t^* be *iid* $\mathbb{N}(0, 1)$ random draws for $t = 1, \ldots, T$ and set $x_t^* = u_t^*$. Regressing x_t^* on past values of x_t, we can obtain the likelihood functions $lnf^*(X^*|\gamma)$ and $lnf^*(X^*|\hat{\gamma})$ to form the likelihood profile $LR_n^*(\gamma) = 2\left[lnf(X^*|\hat{\gamma}) - lnf(x^*|\gamma)\right]$ and $LR_n^* = sup_{\gamma \in \Gamma} LR_n^*(\gamma)$. Hansen (1996) shows that the distribution of LR_n^* converges weakly in probability to the null distribution of LR_n so that repeated (bootstrap) draws from LR_n^* may be used to approximate the asymptotic null distribution of LR_n. For further details, refer to Hansen (1996).

[5] This is standard in the literature to avoid end-of-sample distortions.

3.2 Size properties

We generate data according to equation (5) and set $\alpha = 1.2$ and allow the AR coefficient ϕ to take the following values: $\{-0.9, -0.5, -0.1, 0.1, 0.5, 0.9\}$. Table 1 shows the rejection frequencies of the null hypothesis of linearity when the DGP is linear for different values of ζ, T and ϕ. That is, it shows the frequency with which the linearity test falsely rejects the null hypothesis when it is true, considering 5 percent critical values for the linearity test.

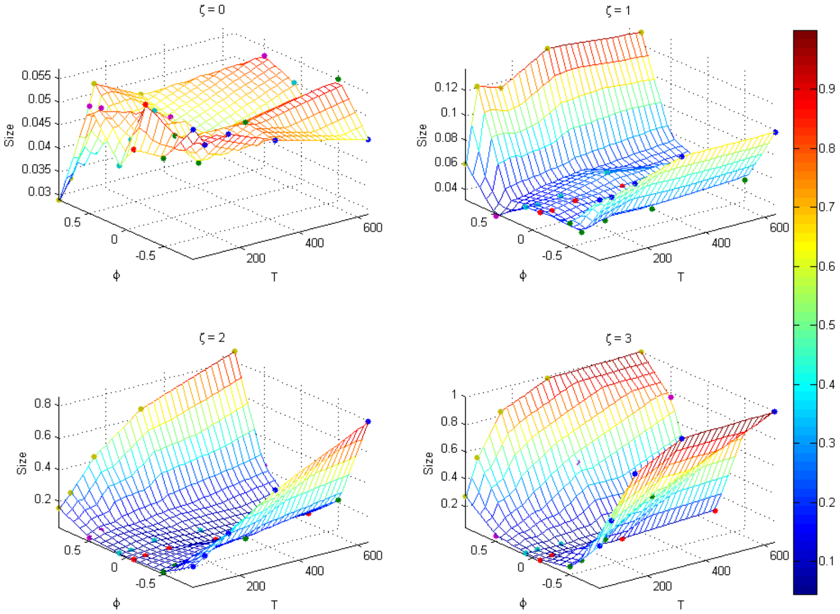
The results reported in table 1 demonstrate that size of the test is close to the nominal level of 5 percent across different sample sizes and levels of persistence in the absence of outliers (i.e., for $\zeta = 0$). However, the rejection frequency for the contaminated series y_t increases with the degree of contamination, ζ, for a given sample size, T. For example, for $T = 80$ and $\phi = 0.5$, the percentage of false rejections is 4.0, 8.3 and 14.4 for $\zeta = 1, 2, 3$, respectively. The pattern is similar for all values of ϕ. In general, the size of the test increases with the sample size T, with the exception of processes with very low persistence ($\phi = -0.1, 0.1$). In all other cases, when the processes are relatively persistent, or very persistent, the rejection frequencies increase to very high levels for high values of ζ and/or high values of T. This is consistent with the results found in van Dijk et al. (1999).

The reason behind the higher rejection frequencies associated with higher sample sizes is that, in larger samples, the fraction of outlier observations in the contaminated series, y_t, is also larger. With more outlier observations, linearity tests are more likely to reject the null hypothesis. These rejection frequencies are substantially higher in the case of more persistent series. Intuitively, when processes are less persistent, the distribution of the observations generated is wider and, therefore, outlier observations tend to blend in within a more scattered cloud of points, as shown in the left panel of figure 1. In contrast, for more persistent processes, the distribution of observations generated from the core process x_t is much tighter. Hence, outliers are further away from the cloud of points, distorting inference.

The distortion of the size of the test for large T and high ϕ becomes more evident from a graphical perspective. The results from table 1 are summarized in figure 3. Each panel, which corresponds to different magnitudes of the outlier observations, ζ, shows rejection frequencies for varying values of T and ϕ. In the upper left panel, where $\zeta = 0$, the size of the test is close to the nominal level of 5 percent in most cases.

In the presence of outliers, the size is clearly distorted. For example, in the upper right panel, where $\zeta = 1$, rejection frequencies remain close to the 5 percent nominal level for values of ϕ close to zero, even for large samples. However, as the persistence of the AR process (ϕ) departs from zero, the size increases forming a U-shaped surface. The same pattern arises for the cases of $\zeta = 2$ and $\zeta = 3$. Note that the size is substantially higher for the bottom panels, reaching rejection frequencies as high 80 percent for $\zeta = 2$ and as high as 100 percent for $\zeta = 3$.

Figure 3: Size of Hansen (1996)'s bootstrap test in the presence of outlier observations



Rejection frequencies of the null hypothesis of linearity, using Hansen (1996)'s bootstrap linearity test, based on 1,000 Monte Carlo simulations. The DGP is generated according to equation (5) with $\alpha = 1.2$ and different values of the magnitude of the outliers, ζ, sample size, T, and level of persistence ϕ. Outliers occur with probability $\pi = 0.05$.

3.3 Power properties

Table 2 shows rejection frequencies of the null hypothesis of linearity when the core process x_t is described by a threshold autoregression of order one, TAR(1). Specifically, the model in (4) reduces to:

$$x_t = \begin{cases} \alpha_1 + \phi_1 x_{t-1} + \epsilon_t, & \text{if } x_{t-d} \leq \gamma \\ \alpha_2 + \phi_2 x_{t-1} + \epsilon_t, & \text{if } x_{t-d} > \gamma \end{cases} \tag{9}$$

where $d = 1$, and we set $\gamma = 0$, $\alpha_1 = 0$, $\phi_1 = 0.6$, $\phi_2 = 0$ and allow α_2 to vary from 0.1 to 0.6 (to assess sensitivity with respect to the threshold effect). This DGP corresponds to one used in Hansen (1997). The contaminated series y_t are obtained by adding AO's to (9) according to equation (2). It is important to notice that this DGP is not very persistent in regime 1 (when $x_{t-1} \leq \gamma$) and has zero persistence in regime 2.

From the results reported in table 2, it can be seen that the power of the test increases with the sample size, T, as expected. For small sample sizes, the power of the test is very low, reflecting the little information available to identify both regimes accurately. However, for large sample sizes, the power of the test increases dramatically, approaching 100 percent even in the presence of large outliers. We also note that, for a given sample size T, the power of the test also increases with the threshold effect (*i.e.* as α_2 increases, the rejection frequency also increases for given T and ζ), regardless of the existence or size of outliers.

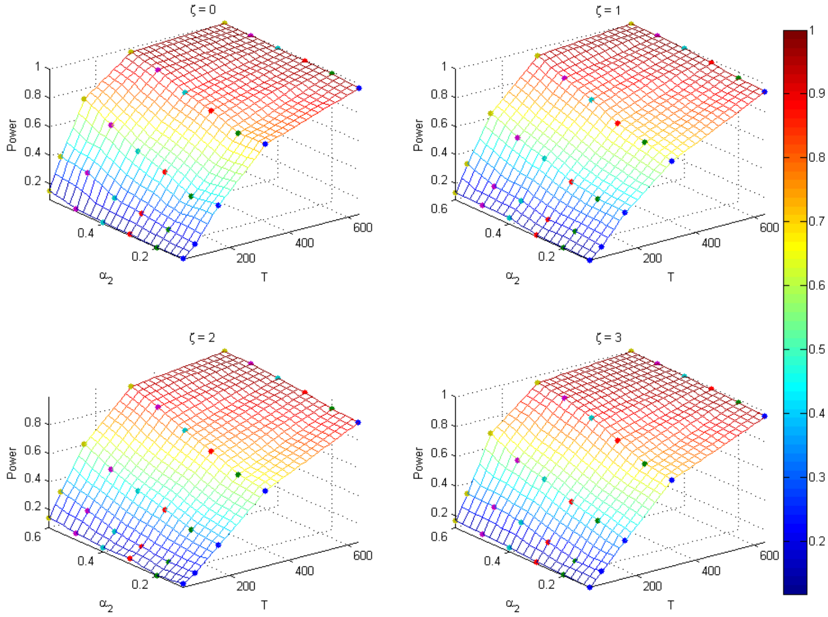
The results here show that the presence (or absence) of outliers seems to only have a small effect on the rejection frequencies. For example, when $\zeta = 0$, the power of the test is not very different from the cases where $\zeta = 1, 2, 3$, for any given T or α_2. Additionally, it is worth noting that the power of the test generally increases with ζ, for $\zeta \neq 0$. That is, in the presence of outliers, the power of the test improves with the magnitude of the outlier approaching levels that are sometimes close to, or even higher than, the power of the test in the absence of outliers. For example, for $T = 80$ and $\alpha_2 = 0.1$, the power of the test is 17.9 percent in the absence of outliers. Once relatively small outliers contaminate the series ($\zeta = 1$), the power drops to 15.7 percent. As ζ increases to 2 and 3, the power of the test improves to 15.8 and 20.7, respectively.

Intuitively, we can think of these results as follows. For small values of ζ, the observations essentially add noise to the series, which decreases the power of the test. However, for larger values of ζ, it helps to highlight observations in different regimes. Consequently, the power of the test increases.

The increase in the power of the test with the sample size T and the threshold effect, α_2, is more easily perceived visually. Figure 4 summarizes the results from table 2. As in the previous figure, each panel corresponds to different magnitudes of the outlier observations, ζ, and shows rejection frequencies for T and α_2 for the TAR process described in (9), with AO's introduced according to equation (2). In the upper left panel, which corresponds to the case of the DGP without outliers ($\zeta = 0$), the hyperplane of rejection frequencies is concave and each contour, on the threshold effect axis, increases with α_2. That is, the power of the test increases with T and α_2. In the presence of outliers, the hyperplanes of rejection frequencies exhibit a similar behavior, regardless of the value of ζ. The power of the test increases with the sample size and the rise in power occurs faster in the presence of a larger the threshold effect (α_2).

To the extent that the degree of persistence drastically distorts the size of the test, as shown in the previous subsection, we next evaluate how the power of the test is affected by the presence of outliers when the degree of persistence changes for a given TAR process. The DGP, in this case, is given by

Figure 4: Power of Hansen (1996)'s bootstrap test in the presence of outlier observations



Rejection frequencies of the null hypothesis of linearity, using Hansen (1996)'s bootstrap linearity test, based on 1,000 Monte Carlo simulations. The DGP is generated according the threshold process described in equation (9) with $d = 1$, $\gamma = \alpha_1 = 0$, $\phi_1 = 0.6$ and different values of the magnitude of the outliers, ζ, sample size, T, and threshold effect α_2. Outliers occur with probability $\pi = 0.05$.

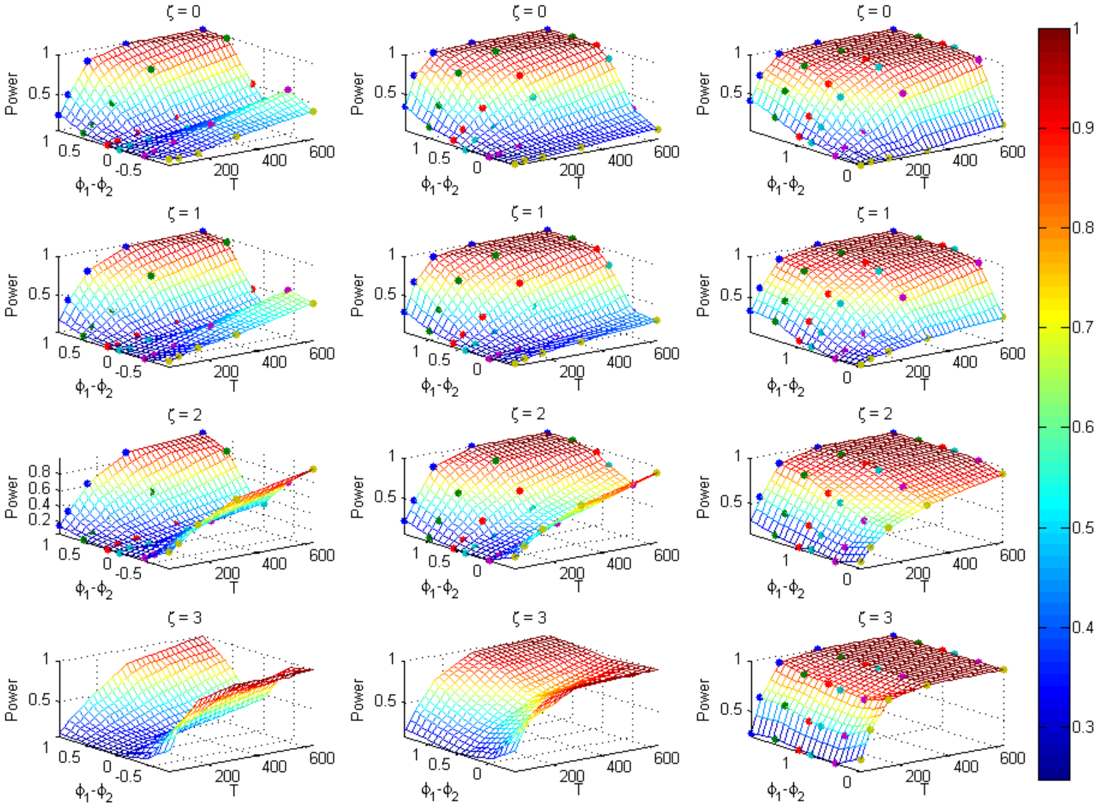
a core process, x_t, that follows the TAR(1) setting described in (9) with $d = 1$, $\gamma = 0$, $\alpha_1 = 0$, and $\phi_1 = \{0.1,\ 0.5,\ 0.9\}$, $\alpha_2 = \{0.1,\ 0.3,\ 0.6\}$ and $\phi_2 = \{-0.9,\ -0.5,\ -0.1,\ 0.1,\ 0.5,\ 0.9\}$.

Figure 5 displays the power of the test for different sizes of outliers, ζ, samples sizes, T, and difference in persistence parameters, $(\phi_1 - \phi_2)$, when $\alpha_2 = 0.3$.[6] The columns in figure 5 correspond to the cases where $\phi_1 = 0.1$, $\phi_1 = 0.5$ and $\phi_1 = 0.9$, respectively. The rows correspond to the cases where $\zeta = 0$, $\zeta = 1$, $\zeta = 2$ and $\zeta = 3$, respectively. In general, the results from figure 5 support the findings from figures 4 and 3. The frequency of rejections of the null hypothesis of linearity increases with ϕ_1. That is, a higher degree of persistence in at least one of the regimes increases the power of the test for given ζ. Graphically, the surfaces are higher as ϕ_1 increases from column to column, for each row.

To illustrate the results, suppose that we focus on the first column, where $\phi_1 = 0.1$. When $\phi_1 = 0.1$, the dynamics of the TAR process has very little persistence in regime 1. As we vary the value of ϕ_2 and as the difference $(\phi_1 - \phi_2)$ departs from zero, the dynamics in the second regime become more persistent

[6] The results for the cases of $\alpha_2 = 0.1$ and $\alpha_2 = 0.6$ are not very different and, therefore, not reported here. They are, however, are available upon request

Figure 5: Power of Hansen (1996)'s bootstrap test in the presence of outlier observations (by degree of persistence)



Rejection frequencies of the null hypothesis of linearity, using Hansen (1996)'s bootstrap linearity test, based on 1,000 Monte Carlo simulations. The DGP is generated according the threshold process described in equation (9) with $d = 1$, $\gamma = \alpha_1 = 0$, and different values of the magnitude of the outliers, ζ, sample size, T, threshold effect, α_2, and difference in persistence parameters $(\phi_1 - \phi_2)$. Outliers occur with probability $\pi = 0.05$. The first, second and third columns correspond to the cases where $\phi_1 = 0.1$, $\phi_1 = 0.5$ and $\phi_1 = 0.9$, respectively.

and the power of the test increases. This increase in power is enhanced with the presence of outliers, especially in the case of large outliers ($\zeta = 3$). For example, when the TAR process is not contaminated with outlier observations ($\zeta = 0$), the frequency of rejections of the null of linearity remains small for $(\phi_1 - \phi_2)$ close to zero, even as the sample size T increases, since the threshold effect is, in that case, small. As $(\phi_1 - \phi_2)$ departs from zero, in either direction, the power of the test increases, especially as T increases, since the larger number of observations make the identification of the two regimes easier. Note that the increase in power is not symmetric, however, in the sense that the power is much higher as $(\phi_1 - \phi_2)$ approaches 1 (which implies that $\phi_2 = -0.9$) than for $(\phi_1 - \phi_2)$ approaching -0.8 (which implies that $\phi_2 = 0.9$). In the former case, the threshold effect is much larger than the latter case. With the introduction of outlier observations, the power of the test increases more symmetrically as $(\phi_1 - \phi_2)$ departs from zero. The frequency of rejections as $(\phi_1 - \phi_2)$ approaches -0.8 increases with the magnitude of the outliers and, for $\zeta = 3$, it gets closer to 100 percent, similar to the case when $(\phi_1 - \phi_2)$ approaches 1.

The results are similar for the case where we fix the persistence of regime 1 to $\phi_1 = 0.5$ and we vary the persistence of regime 2. As the difference in persistence $(\phi_1 - \phi_2)$ approaches the maximum difference between the two regimes of 1.4, the degree of persistence in the second regime and the associated threshold effect increase. Consequently, we find that the power of the test increases to 100 percent, even for relatively small samples since the test is better able to distinguish the two regimes. Meanwhile, when the persistence of the second regime is high ($\phi_2 = 0.9$), but the difference in persistence between the regimes is small (and negative), we find that the associated threshold effect is smaller. In the absence of outliers, the frequency of rejections is smaller in this case, even as the sample size increase. However, as outliers contaminate the process, the power of the test increases as $(\phi_1 - \phi_2)$ approaches -0.4 and, when $\zeta = 3$, it gets closer to 100 percent, similar to the case where $(\phi_1 - \phi_2)$ approaches 1.4.

The final case is for the scenario where the persistence is high in the first regime ($\phi_1 = 0.9$), and we vary the persistence of the second regime, ϕ_2. As $(\phi_1 - \phi_2)$ approaches the maximum difference of 1.8, both regimes are very persistent and the associated threshold effect increases. Therefore, the power of the test quickly raises close to 100 percent, even for small samples. By contrast, when $(\phi_1 - \phi_2)$ decreases near 0, the associated threshold effect becomes smaller and, in the absence of outliers, the power of the test remains very low. However, once the series is contaminated with them, the frequency of rejections increases with ζ. Notably, for $\zeta = 3$, the power of the test increases closer to 100 percent, even for small samples, and even when $(\phi_1 - \phi_2)$=0. In this case, while the persistence parameters are identical in both regimes, the intercepts still switch regimes and the presence of outliers seems to help the linearity test better identify the regime-switching.[7]

Overall, the findings here are contrary to what conventional wisdom might indicate. They support the premise that large outliers may help the identification of threshold-type nonlinearity, particularly in the case of persistent dynamics in regimes.

[7] We also repeat the exercise in a scenario where there are regime specific outliers. We do not present the results here since the conclusions are similar to those found above.

4 Concluding Remarks

We have studied the effects of outlier observations on a bootstrap test of linearity against threshold autoregressions when the DGP is described by an AR or a TAR process for different sample sizes, persistence parameters and magnitude of outliers. The Monte Carlo evidence suggests that the empirical performance of Hansen (1996)'s bootstrap linearity test does not suffer from large size distortions or much loss of power in the case of series that are not contaminated. However, their performance can be distorted by the presence of outlier observations. Specifically, the size of the test is more distorted in more persistent AR processes, especially in large sample environments. At the same time, we find that the power of the test increases with the sample size, the magnitude of the threshold effect, the magnitude of the outlier observation and the persistence of the process.

Interestingly, and contrary to our priors, we also find that the magnitude of the outlier observations can help the bootstrap test to better identify the change in regimes, especially in the case of more persistent series. This result is puzzling as we would expect that series that are more highly contaminated would distort the ability of test statistics to correctly identify nonlinear processes as such. Intuitively, this could be explained by the fact that the distribution of observations in more persistent processes is narrower and, consequently, the cloud of points in a scatter plot is tighter. When outliers are relatively large in such settings, then, those observations will lie further away from the cloud of points, making the identification of regimes easier for tests statistics.

It should be mentioned, however, that our results do not suggest that all nonlinearity is caused by outliers. Rather, the results in this study are meant to guide researchers in being careful when making inferences about the presence or absence of nonlinear properties in time series. More importantly, our results seem to suggest that the persistence of time series, and their volatility (associated with large outlier observations), could provide insightful information to develop an outlier-robust version of the bootstrap test. This is left for future research.

References

Ahmad, Y. S. (2008), "The Effects of Small Sample Bias in Threshold Autoregressive Models," *Economics Letters*, 101, 6–8.

Ahmad, Y. S. and Glosser, S. (2011), "Searching for Nonlinearities in Real Exchange Rates," *Applied Economics*, 43, 1829–1845.

Atanasova, C. (2003), "Credit Market Imperfections and Business Cycle Dynamics: A Nonlinear Approach," *Studies in Nonlinear Dynamics and Econometrics*, 7(4), Article 5.

Auerbach, A. J. and Gorodnichenko, Y. (2012), "Measuring the Output Responses to Fiscal Policy," *American Economic Journal: Economic Policy*, 4(2), 1–27.

— (2013), "Fiscal Multipliers in Recession and Expansion," in *Fiscal Policy After the Financial Crisis*, eds. Alesina, A. and Giavazzi, F., University of Chicago Press, pp. 63–98.

Balke, N. S. (2000), "Credit and Economic Activity: Credit Regimes and Nonlinear Propagation of Shocks," *The Review of Economics and Statistics*, 82, 344–349.

Beaudry, P. and Koop, G. (1993), "Do Recessions Permanently Change Output?" *Journal of Monetary Economics*, 31, 149–63.

Bec, F., Ben-Salem, M., and Carrasco, M. (2004), "Tests for Unit Root versus Threshold Specification with an Application to the Purchase Power Parity Relationship," *Journal of Business and Economic Statistics*, 22, 382–395.

Donayre, L. (2014), "Estimated Thresholds in the Response of Output to Monetary Policy: Are Large Policy Changes Less Effective?" *Macroeconomic Dynamics*, 18(1), 41–64.

Enders, W., Falk, B., and Siklos, P. (2007), "A Threshold Model of U.S. Real GDP and the Problem of Constructing Confidence Intervals in TAR Models," *Studies in Nonlinear Dynamics and Econometrics*, 11(3), Article 4.

Fazzari, S., Morley, J. C., and Panovska, I. (2013), "State-Dependent Effects of Fiscal Policy," Working Paper, Lehigh University.

Galbraith, J. W. (1996), "Credit Rationing and Threshold Effects in the Relation between Money and Output," *Journal of Applied Econometrics*, 12, 419–429.

Gonzalo, J. and Pitarakis, J.-Y. (2013), “Estimation and Inference in Threshold Type Regime Switching Models,” in *Handbook of Research Methods and Applications in Empirical Macroeconomics*, eds. Hashimzade, N. and Thornton, M. A., Cheltenham: Edward Elgar Publishing Limited, pp. 189–204.

Gospodinov, N. (2005), “Testing for Threshold Nonlinearity in Short-term Interest Rates,” *Journal of Financial Econometrics*, 3, 344–371.

Hansen, B. and Seo, B. (2002), “Testing for Two-Regime Threshold Cointegration in Vector Error Correction Models,” *Journal of Econometrics*, 110, 293–318.

Hansen, B. E. (1996), “Inference when a Nuisance Parameter is not Identified under the Null Hypothesis,” *Econometrica*, 64, 413–430.

— (1997), “Inference in TAR Models,” *Studies in Nonlinear Dynamics and Econometrics*, 2(1), 1–14.

— (2011), “Threshold Autoregression in Economics,” *Statistics and Its Interface*, 4, 123–127.

Juvenal, L. and Taylor, M. P. (2008), “Threshold Adjustment of Deviations from the Law of One Price,” *Studies in Nonlinear Dynamics and Econometrics*, 12(3), 1–46.

Koop, G. and Potter, S. (2001), “Are Apparent Findings of Nonlinearity due to Structural Instability in Economic Time Series?” *Econometrics Journal*, 4, 37–55.

— (2004), “Dynamic Asymmetries in U.S. Unemployment,” *Journal of Business and Economic Statistics*, 17(3), 298–312.

López Villavicencio, A. (2008), “Nonlinearities or Outliers in Real Exchange Rates?” *Economic Modelling*, 25, 714–730.

Lucas, A., Franses, P. H., and van Dijk, D. (2002), *Outlier Robust Analysis of Economic Time Series*, Oxford, U.K.: Oxford University Press Inc.

Martin, R. D. and Yohai, V. J. (1986), “Influence Functionals for Time Series,” *The Annals of Statistics*, 14, 781–818.

Pesaran, H. and Potter, S. (1997), “A Floor and Ceiling Model of U.S. Output,” *Journal of Economic Dynamics and Control*, 21(4-5), 661–695.

Potter, S. (1995), “A Nonlinear Approach to U.S. GNP,” *Journal of Applied Econometrics*, 10, 109–25.

Sander, H. and Kleimeier, P. (2004), "Convergence in Euro-zone Retail Banking? What Interest Rate Pass-through Tells Us About Monetary Policy Transmission, Competition and Integration," *Journal of International Money and Finance*, 23, 461–491.

Sarno, L., Taylor, M., and Chowdhury, I. (2004), "Nonlinear Dynamics in Deviations from the Law of One Price: A Broad-Based Empirical Study," *Journal of International Money and Finance*, 23(1), 1–25.

Taylor, A. (2001), "Potential Pitfalls for the Purchasing Power Parity Puzzle? Sampling and Specification Tests of the Law of One Price," *Econometrica*, 69, 473–498.

Tong, H. (2011), "Threshold Models in Time Series Analysis - 30 years on," *Statistics and Its Interface*, 4, 107–136.

Tsay, R. S. (1998), "Testing and Modeling Multivariate Threshold Models," *Journal of the American Statistical Association*, 93, 1188–1202.

van Dijk, D., Franses, P. H., and Lucas, A. (1999), "Testing for Smooth Transition Nonlinearity in the Presence of Outliers," *Journal of Business and Economics Statistics*, 17(2), 217–235.

van Dijk, D., Teräsvirta, T., and Franses, P. H. (2002), "Smooth Transition Autoregressive Models: A Survey of Recent Developments," *Econometric Reviews*, 21, 1–47.

Table 1: Size of Hansen (1996)'s bootstrap test in the presence of outlier observations

		Sample size	40				80				160				320				640			
π	ϕ	ζ	0	1	2	3	0	1	2	3	0	1	2	3	0	1	2	3	0	1	2	3
0.05	-0.9		5.7	8.0	17.1	35.0	5.3	8.0	22.1	51.4	5.4	8.6	32.0	81.3	5.0	9.8	51.5	98.2	4.5	9.8	79.4	99.9
	-0.5		4.8	5.1	5.3	9.9	5.2	4.1	7.3	12.5	4.5	4.6	8.8	23.5	5.1	4.5	12.9	45.9	5.5	4.9	21.5	77.9
	-0.1		4.7	4.9	4.8	5.3	5.6	4.8	4.9	5.9	4.7	5.1	5.6	4.4	4.6	5.3	3.6	5.1	4.6	5.8	4.7	8.1
	0.1		4.2	4.8	4.5	5.5	4.7	4.7	4.4	5.7	5.2	4.9	6.6	5.1	4.2	5.8	5.2	7.8	5.0	4.7	4.0	7.3
	0.5		5.2	3.0	5.8	8.6	5.1	4.0	8.3	14.4	4.4	4.4	9.4	22.6	4.5	5.0	13.5	46.8	5.3	5.7	22.6	76.4
	0.9		2.9	6.1	16.1	27.8	3.3	12.1	24.0	53.5	5.2	11.5	42.8	82.9	4.7	13.7	65.3	98.4	4.7	13.0	85.7	100.0

Rejection frequencies of the null hypothesis of linearity, using Hansen (1996)'s bootstrap linearity test, based on 1,000 Monte Carlo simulations. The DGP is generated according to equation (5) with $\alpha = 1.2$ and different values of the magnitude of the outliers, ζ, sample size, T, and level of persistence ϕ. Outliers occur with probability $\pi = 0.05$.

Table 2: Power of Hansen (1996)'s bootstrap test in the presence of outlier observations

		Sample size	40				80				160				320				640			
π	α_2	ζ	0	1	2	3	0	1	2	3	0	1	2	3	0	1	2	3	0	1	2	3
0.05	0.1		9.4	8.4	10.1	11.8	17.9	15.7	15.8	20.7	40.6	35.2	30.8	37.7	74.6	62.7	61.0	70.3	96.7	94.5	92.1	96.6
	0.2		8.7	8.5	7.1	12.4	19.3	17.9	14.7	19.9	38.2	36.3	33.2	42.9	73.7	67.4	63.9	73.1	98.6	95.1	93.2	98.1
	0.3		9.8	8.9	11.0	13.1	22.2	17.7	16.9	24.2	46.8	39.9	39.3	47.4	81.5	72.2	72.2	80.6	99.0	97.6	96.5	98.9
	0.4		9.9	12.4	9.8	14.7	24.4	22.7	19.5	29.8	53.1	47.7	43.7	54.6	85.8	80.8	77.9	88.2	99.3	99.3	98.1	99.5
	0.5		10.7	12.8	11.9	15.5	33.9	24.0	25.8	31.8	63.2	54.3	50.8	59.5	93.1	87.1	86.4	93.1	99.9	99.9	99.9	99.9
	0.6		15.1	13.6	14.3	16.9	36.7	31.5	30.7	33.0	73.2	63.1	60.4	69.4	97.5	94.8	92.6	95.5	100.0	100.0	99.9	100.0

Rejection frequencies of the null hypothesis of linearity, using Hansen (1996)'s bootstrap linearity test, based on 1,000 Monte Carlo simulations. The DGP is generated according the threshold process described in equation (9) with $d = 1$, $\gamma = \alpha_1 = 0$, $\phi_1 = 0.6$ and different values of the magnitude of the outliers, ζ, sample size, T, and threshold effect α_2. Outliers occur with probability $\pi = 0.05$.