An over-rejection puzzle of bootstrap average tests for the no-threshold-effect hypothesis

Kaiji Motegi[*] and John W. Dennis[†]

Kobe University IDA

April 1, 2024

Abstract

When testing the null hypothesis of no threshold effects based on threshold autoregressive models, wild-bootstrap supremum, average, and exponential tests are routinely used to handle an identification issue under the null. In this note, we demonstrate via Monte Carlo simulations that the bootstrap average tests lose control for the type-I error rate when the threshold variable is persistent and the delay parameter is chosen from more than a handful of choices. In some cases, the average tests reject the correct null hypothesis with probability exceeding nominal size by more than 10%. The size distortion is present even in large samples, indicating the average tests may not converge to the intended asymptotic null distribution. Supremum and exponential tests achieve correct type-I error rates, posing a puzzle why only the average tests suffer from over-rejections.

JEL codes: C12, C22.

Keywords: Asymmetry, nonlinear time series analysis, size distortion, Threshold Autoregression (TAR), type-I error, wild bootstrap.

[*] *Corresponding author.* Graduate School of Economics, Kobe University. 2-1 Rokkodai-cho, Nada, Kobe, Hyogo 657-8501 Japan. E-mail: motegi@econ.kobe-u.ac.jp

[†] Institute for Defense Analyses (IDA). Research results and conclusions expressed are those of the authors and do not necessarily reflect the views of IDA. E-mail: jay.dennis@alumni.unc.edu

1 Introduction

In the literature of nonlinear time series analysis, it is well known that economic and financial indicators sometimes exhibit *threshold effects*: a target variable y has asymmetric structures when a threshold variable x is below versus above a threshold μ. A variety of threshold time series models, including Tong's (1978) *Threshold Autoregression (TAR)*, have been proposed in the literature. Whichever threshold model is used, a major concern lies in testing the null hypothesis of no threshold effects H_0. Under H_0, several parameters such as the delay d and threshold μ are unidentified. Hansen (1996) provides a well-known solution to this issue via wild-bootstrap tests, where common test statistics are supremum (sup-), average (ave-), and exponential (exp-) Wald statistics as well as their Lagrange Multiplier (LM) counterparts.[1]

In this note, we demonstrate via Monte Carlo simulations that the bootstrap ave-Wald and ave-LM tests exhibit distortions for the type-I error rate if the threshold variable x is persistent and the space of delay parameter d contains several choices. In some cases, the average tests reject the correct no-threshold-effect hypothesis with probability exceeding nominal size by more than 10%. This is particularly puzzling as x and d do not play any role under the true data generating process. The size distortion exists even in large samples, indicating that the average tests may not converge to the intended asymptotic null distribution. Supremum and exponential tests do not appear to exhibit size distortions in the same scenario. This puzzle is not documented in the existing literature, and reasons for the over-rejections are unknown. While using sup-tests or exp-tests is a practical solution, it would be an interesting future task to explain why only the average tests appear to produce these distortions.

The remainder of this note is organized as follows. Section 2 sets up a simulation study. We present our simulation results in Section 3, and give some concluding remarks in Section 4. Omitted technical details and complete Monte Carlo simulations are collected in a separate supplemental material. We use the following notation throughout: $\mathbb{R}$ is the set of real numbers, $\mathbb{N}$ is the set of natural numbers, $\lfloor a \rfloor$ is the largest integer not larger than $a \in \mathbb{R}$, $\#A$ is the number of elements of set A, and $A \times B$ is the Cartesian product of sets A and B.

[1] Extensive surveys on TAR are provided by Tong (2015) and Tsay and Chen (2019), among others. Recent empirical applications of threshold time series models and the wild-bootstrap tests include Chen, Qiao, and Zhang (2022) and Motegi and Hamori (2023); they detected significant threshold effects of stock market realized volatilities on crude oil realized volatilities. Andrews and Ploberger (1994) discuss use of the wild-bootstrap tests when a nuisance parameter is present only under the alternative hypothesis.

2 Simulation design

Suppose that the true data generating process (DGP) is as follows.

$$\begin{bmatrix} y_t \\ x_t \end{bmatrix} = \begin{bmatrix} \phi_0 & 0 \\ 0 & \psi_0 \end{bmatrix} \begin{bmatrix} y_{t-1} \\ x_{t-1} \end{bmatrix} + \begin{bmatrix} \epsilon_t \\ \nu_t \end{bmatrix}, \qquad \begin{bmatrix} \epsilon_t \\ \nu_t \end{bmatrix} \overset{i.i.d.}{\sim} \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right). \tag{1}$$

The target variable y and the threshold variable x follow mutually independent single-regime AR(1) processes. Fix the AR(1) parameter for y at $\phi_0 = 0.6$.[2] Consider low, medium, or high persistence in x: $\psi_0 \in \{0.3, 0.6, 0.9\}$. The joint standard normality of the true errors is a conventional assumption which simplifies analysis. Generate $J = 1000$ Monte Carlo samples of size $n \in \{125, 250, 500, 1000\}$ from DGP (1).

For each sample generated from (1), fit a two-regime TAR model with lag length $p = 1$:

$$y_t = \begin{cases} \alpha_1 + \phi_1 y_{t-1} + u_t & \text{if} \quad x_{t-d} < \mu, \\ \alpha_2 + \phi_2 y_{t-1} + u_t & \text{if} \quad x_{t-d} \geq \mu, \end{cases} \tag{2}$$

where $\boldsymbol{\beta}_r = (\alpha_r, \phi_r)^\top$ is a vector of regression parameters in regime $r \in \{1, 2\}$; $d \in \mathbb{N}$ is the delay parameter; and $\mu \in \mathbb{R}$ is the threshold parameter.[3] Consider the null hypothesis of no threshold effects $H_0 : \boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$ versus a fixed alternative hypothesis $H_1 : \boldsymbol{\beta}_1 \neq \boldsymbol{\beta}_2$. Since the DGP is given by (1), H_0 is true in our experiment. Under H_0, the nuisance parameters $\boldsymbol{\gamma} = (d, \mu)^\top$ are not identified. To address this identification problem, we perform Hansen's (1996) wild-bootstrap tests with ave-Wald, exp-Wald, ave-LM, and exp-LM test statistics. When computing the actual and bootstrap test statistics, we use a simple covariance matrix which is not robust to heteroscedasticity. This is a correct choice since the error term u in model (2) is indeed homoscedastic given the true DGP (1). Nominal size is set to be $a = 0.05$, and the number of bootstrap iterations is $B = 500$.

To implement the bootstrap tests, one needs to specify the choice space of nuisance parameter $\boldsymbol{\gamma}$. We set the choice space of d to be $D = \{1, \ldots, \bar{d}\}$ with the upper bound being $\bar{d} \in \{1, 2, 4, 8\}$. Specifically, the choice space D is either $\{1\}$, $\{1, 2\}$, $\{1, 2, 3, 4\}$, or $\{1, \ldots, 8\}$. For the choice space of μ, let $x_{[1]} \leq \cdots \leq x_{[n]}$ be a sorted version of x.

[2] Simulation results are similar for $\phi_0 \in \{0.3, 0.6, 0.9\}$, hence we focus on the middle value here. Complete results are reported in the supplemental material.

[3] Previous results in the literature, including Hansen (1996), ensure uniform consistency of the estimator for $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$.

In general, the space of μ is specified as

$$\mathcal{X}_{\kappa,n} = \left\{ x_{[\lfloor 0.5(1-\kappa)n \rfloor]}, \ \ldots, \ x_{[\lfloor \{1-0.5(1-\kappa)\}n \rfloor]} \right\}, \tag{3}$$

where $\kappa \in [0, 1)$ signifies the fraction of $\#\mathcal{X}_{\kappa,n}$ to n. We pick $\kappa = 0.7$, following a well-known suggestion of Andrews (1993). This means that $\mathcal{X}_{\kappa,n}$ consists of observations above the 15 percentile and below the 85 percentile of x. The space of $\boldsymbol{\gamma}$ is given by $\Gamma_{\kappa,n} = D \times \mathcal{X}_{\kappa,n}$. Compute a conditional test statistic given each $\boldsymbol{\gamma} \in \Gamma_{\kappa,n}$, and aggregate the $\#\Gamma_{\kappa,n}$ outcomes into a single test statistic by either the average or exponential transformation.[4]

In finite samples, there is expected to be a bias-variance trade-off on the cardinality of $\Gamma_{\kappa,n}$. Under H_0, the finite sample performance of the bootstrap tests is expected to be worse due to larger variance with larger $\#\Gamma_{\kappa,n}$. Under H_1, asymptotic bias arises if a true value of $\boldsymbol{\gamma}$ is not included in $\Gamma_{\kappa,n}$. In our current set-up, threshold effects are absent and H_0 is true given DGP (1). Hence, the empirical size of the bootstrap tests should be closest to the nominal size 5% for $\bar{d} = 1$ and farthest for $\bar{d} = 8$, particularly for small sample sizes. When the sample size is large enough (say $n = 1000$), the type-I error rate should be sufficiently close to 5% for any $\bar{d} \in \{1, 2, 4, 8\}$. Keeping this conjecture in mind, we report rejection frequencies in the next section.

3 Simulation results

In Table 1, we report rejection frequencies of the bootstrap average tests for the no-threshold-effect hypothesis H_0 based on the TAR model. We maintain focus on the ave-LM test, as the ave-Wald test results are similar. When the threshold variable x has low persistence (i.e., $\psi_0 = 0.3$), the empirical size of the ave-LM test is sufficiently close to the nominal size $a = 0.05$ for any $\bar{d} \in \{1, 2, 4, 8\}$. Taking $(\psi_0, \bar{d}, n) = (0.3, 1, 1000)$ as an example, the empirical size of the ave-LM test is 0.055.

When x has medium persistence (i.e., $\psi_0 = 0.6$), size distortions emerge as $\bar{d}$ increases. The empirical size of the ave-LM test with $(\bar{d}, n) = (8, 1000)$ is 0.108 (Table 1). When x has high persistence (i.e., $\psi_0 = 0.9$), the tendency to over-reject H_0 becomes more salient. The average tests lose control for the type-I error rate as $\bar{d}$ grows. The empirical size of the ave-LM test with $n = 1000$ is $\{0.051, 0.086, 0.132, 0.163\}$ for

[4] Complete procedures of the bootstrap tests are described in the supplemental material.

Table 1: Empirical size of the bootstrap average tests for the no-threshold-effect hypothesis based on the TAR model (nominal size 5%)

		$n = 125$		$n = 250$		$n = 500$		$n = 1000$	
ψ_0	$\bar{d}$	Wald	LM	Wald	LM	Wald	LM	Wald	LM
0.3	1	0.057	0.041	0.085	0.071	0.055	0.041	0.058	0.055
0.3	2	0.074	0.046	0.063	0.046	0.055	0.050	0.070	0.068
0.3	4	0.073	0.043	0.081	0.049	0.072	0.058	0.070	0.062
0.3	8	0.080	0.037	0.061	0.050	0.067	0.059	0.057	0.055
0.6	1	0.075	0.045	0.047	0.038	0.050	0.045	0.049	0.049
0.6	2	0.098	0.061	0.084	0.065	0.063	0.053	0.069	0.066
0.6	4	0.104	0.067	0.087	0.068	0.083	0.074	0.097	0.092
0.6	8	0.125	0.073	0.104	0.081	0.096	0.083	0.117	0.108
0.9	1	0.075	0.049	0.071	0.062	0.059	0.050	0.051	0.051
0.9	2	0.120	0.088	0.094	0.082	0.082	0.072	0.089	0.086
0.9	4	0.152	0.128	0.142	0.122	0.124	0.115	0.136	0.132
0.9	8	0.193	0.141	0.138	0.122	0.138	0.131	0.172	0.163

DGP: $y_t = 0.6y_{t-1} + \epsilon_t$, $x_t = \psi_0 x_{t-1} + \nu_t$, $\psi_0 \in \{0.3, 0.6, 0.9\}$, $(\epsilon_t, \nu_t)^\top \overset{i.i.d.}{\sim} \mathcal{N}(\mathbf{0}, \boldsymbol{I})$. Sample size: $n \in \{125, 250, 500, 1000\}$. TAR Model: $y_t = \alpha_1 + \phi_1 y_{t-1} + u_t$ if $x_{t-d} < \mu$ and $y_t = \alpha_2 + \phi_2 y_{t-1} + u_t$ if $x_{t-d} \geq \mu$. We perform the bootstrap ave-Wald and ave-LM tests for the no-threshold-effect hypothesis $H_0 : (\alpha_1, \phi_1) = (\alpha_2, \phi_2)$. The choice space of delay parameter, D, is either $\{1\}$, $\{1, 2\}$, $\{1, 2, 3, 4\}$, or $\{1, \ldots, 8\}$ (i.e., $\bar{d} \in \{1, 2, 4, 8\}$). The choice space of μ is $\mathcal{X}_{0.7,n} = \{x_{[\lfloor 0.15n \rfloor]}, \ldots, x_{[\lfloor 0.85n \rfloor]}\}$, where $x_{[1]} \leq \cdots \leq x_{[n]}$ is a sorted x. The nominal size is $a = 0.05$, and the number of bootstrap samples is $B = 500$. This table reports the empirical size of the tests across $J = 1000$ Monte Carlo samples.

$\bar{d} \in \{1, 2, 4, 8\}$, respectively. The excess rejection rate is 8.2% for $\bar{d} = 4$ and 11.3% for $\bar{d} = 8$, which is substantially large compared with the fact that our DGP and model are quite simple.

Surprisingly, the size distortion of the average tests does not vanish as the sample size grows; the type-I error rate is still above the nominal size even for the largest sample sizes we examined. This indicates that the average tests may not converge to the intended asymptotic distribution in the relevant scenario. We are not aware of any existing work that documents this puzzle. In previous simulation studies such as Hansen (1996, Table II) and Ahmad and Donayre (2016, Table 1), the average tests with persistent x and large upper bound $\bar{d}$ are not covered. This note is likely the first work that inspects the relevant case.

What is even more puzzling is that the sup-Wald, exp-Wald, sup-LM, and exp-LM tests achieve correct type-I error rates. The rejection frequencies of the exp-Wald and exp-LM tests are shown in Table 2. For all cases considered, the empirical size is sufficiently close to the nominal size $a = 0.05$. The exp-Wald test tends to over-reject correct H_0 in small samples such as $n = 125$, but their size quickly converges to the nominal size as n grows. Focus on $(\psi_0, \bar{d}) = (0.9, 8)$, in which case the average tests suffer from the most serious over-rejections. The empirical size associated with $n = 1000$ is 0.056 for the exp-Wald test and 0.050 for the exp-LM test. The rejection frequencies of the supremum tests are similar to those of the exponential tests, and we relegate them to the supplemental material.

Table 2: Empirical size of the bootstrap exponential tests for the no-threshold-effect hypothesis based on the TAR model (nominal size 5%)

		$n = 125$		$n = 250$		$n = 500$		$n = 1000$	
ψ_0	$\bar{d}$	Wald	LM	Wald	LM	Wald	LM	Wald	LM
0.3	1	0.062	0.033	0.073	0.048	0.059	0.049	0.055	0.053
0.3	2	0.068	0.032	0.063	0.040	0.054	0.044	0.067	0.055
0.3	4	0.068	0.023	0.064	0.044	0.059	0.049	0.058	0.056
0.3	8	0.080	0.019	0.055	0.032	0.047	0.034	0.059	0.052
0.6	1	0.079	0.036	0.051	0.037	0.057	0.050	0.049	0.046
0.6	2	0.078	0.030	0.059	0.042	0.048	0.042	0.048	0.048
0.6	4	0.074	0.033	0.062	0.043	0.061	0.044	0.069	0.064
0.6	8	0.086	0.024	0.045	0.031	0.059	0.047	0.064	0.054
0.9	1	0.085	0.045	0.076	0.056	0.057	0.051	0.060	0.056
0.9	2	0.088	0.044	0.059	0.043	0.044	0.038	0.054	0.049
0.9	4	0.076	0.039	0.065	0.048	0.048	0.037	0.057	0.052
0.9	8	0.084	0.025	0.051	0.033	0.059	0.044	0.056	0.050

DGP: $y_t = 0.6 y_{t-1} + \epsilon_t$, $x_t = \psi_0 x_{t-1} + \nu_t$, $\psi_0 \in \{0.3, 0.6, 0.9\}$, $(\epsilon_t, \nu_t)^\top \overset{i.i.d.}{\sim} \mathcal{N}(\mathbf{0}, \boldsymbol{I})$. Sample size: $n \in \{125, 250, 500, 1000\}$. TAR Model: $y_t = \alpha_1 + \phi_1 y_{t-1} + u_t$ if $x_{t-d} < \mu$ and $y_t = \alpha_2 + \phi_2 y_{t-1} + u_t$ if $x_{t-d} \geq \mu$. We perform the bootstrap exp-Wald and exp-LM tests for the no-threshold-effect hypothesis $H_0 : (\alpha_1, \phi_1) = (\alpha_2, \phi_2)$. The choice space of delay parameter, D, is either $\{1\}$, $\{1, 2\}$, $\{1, 2, 3, 4\}$, or $\{1, \dots, 8\}$ (i.e., $\bar{d} \in \{1, 2, 4, 8\}$). The choice space of μ is $\mathcal{X}_{0.7,n} = \{x_{[\lfloor 0.15n \rfloor]}, \dots, x_{[\lfloor 0.85n \rfloor]}\}$, where $x_{[1]} \leq \cdots \leq x_{[n]}$ is a sorted x. The nominal size is $a = 0.05$, and the number of bootstrap samples is $B = 500$. This table reports the empirical size of the tests across $J = 1000$ Monte Carlo samples.

Summarizing Tables 1-2, only the average tests over-reject the true no-threshold-effect hypothesis H_0 in large samples. Reasons for the over-rejections are unknown. A practical solution is to use a sup-LM or exp-LM test, as these lead to accurate empirical size in both small and large samples. It is an interesting future task, however, to explain what distorts the type-I error rate of the average tests.

To further characterize the over-rejection puzzle of the bootstrap average tests, we report additional simulation evidence. Recall that D signifies the choice space of the delay parameter d. In the previous simulation, we always fixed the lower bound of D at 1. Consequently, an exact role of D remains unclear. Which distorts the size of the average tests: a large cardinality of D or a large candidate value for d? To answer this question, we consider some additional specifications for D with various lower and upper bounds. In view of the resulting rejection frequencies, it is the cardinality of D, *not* a specific value for candidate d, that plays a key role in the over-rejection puzzle. Excess rejection rates of the average tests increase as $\#D$ grows. Conditional on the value of $\#D$, we observe similar rejection rates irrespective of the specific candidate values contained in D. More detailed results are presented in the supplemental material.

4 Discussion and conclusion

We have presented simulation evidence that the wild-bootstrap average tests for the no-threshold-effect hypothesis $H_0 : \boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$ fail to control the type-I error rate when the threshold variable x is sufficiently persistent and the choice space of the delay parameter d is sufficiently large. In these cases, the type-I error rate far exceeds the nominal size of 5%. Moreover, the size distortion does not diminish as the sample size grows, which indicates that the average tests may converge to an unintended null distribution. The supremum and exponential tests have correct type-I error rates, posing a puzzle why only the average tests suffer from over-rejections.

In additional simulations, we consider a number of alternative scenarios to confirm the robustness of our finding. First, we try larger sample sizes, such as $n = 2000$. Second, the nominal size is determined from $a \in \{0.01, 0.05, 0.10\}$. Third, various degrees of persistence in y are considered: $\phi_0 \in \{0.3, 0.6, 0.9\}$. Fourth, the choice space of μ is more tightly restricted by choosing $\kappa \in \{0, 0.35, 0.7\}$ in (3). Fifth, the simple covariance matrix is replaced with a heteroscedasticity-robust covariance matrix when computing the actual and bootstrap test statistics. Sixth, the TAR model (2) is replaced with a self-exciting version (SETAR) by using y itself as a threshold variable. Our findings are the same for all these scenarios, and most of the additional simulation results are shown in the supplemental material.

Acknowledgements

We thank Shigeyuki Hamori, Seok Young Hong, Ekaterina Kazak, and Tatevik Sekhposyan for helpful comments and discussions. The first author is grateful for the financial support of JSPS KAKENHI Grant-in-Aid for Challenging Research (Exploratory) with Grant Number 23K17555, Ishii Memorial Securities Research Promotion Foundation, Japan Securities Scholarship Foundation, Murata Science and Education Foundation, and Nomura Foundation.

References

Ahmad, Y., and L. Donayre (2016): "Outliers and persistence in threshold autoregressive processes," *Studies in Nonlinear Dynamics & Econometrics*, 20, 37–56.

Andrews, D. W. K. (1993): "Tests for Parameter Instability and Structural Change with Unknown Change Point," *Econometrica*, 61, 821–856.

Andrews, D. W. K., and W. Ploberger (1994): "Optimal tests when a nuisance parameter is present only under the alternative," *Econometrica*, 62(6), 1383–1414.

Chen, Y., G. Qiao, and F. Zhang (2022): "Oil price volatility forecasting: Threshold effect from stock market volatility," *Technological Forecasting & Social Change*, 180, #121704.

Hansen, B. E. (1996): "Inference when a nuisance parameter is not identified under the null hypothesis," *Econometrica*, 64(2), 413–430.

Motegi, K., and S. Hamori (2023): "Conditional threshold effects of stock market volatility on crude oil market volatility," SSRN Working Paper #4512310.

Tong, H. (1978): "On a threshold model," in *Pattern Recognition and Signal Processing*, ed. by C. H. Chen. Sijthoff and Noordhoff, Amsterdam.

——— (2015): "Threshold models in time series analysis—Some reflections," *Journal of Econometrics*, 189, 485–491.

Tsay, R. S., and R. Chen (2019): *Nonlinear Time Series Analysis*. John Wiley & Sons, Inc.