Institut für Wirtschaftsforschung Halle
Halle Institute for Economic Research

Testing for Structural Breaks at Unknown Time: A Steeplechase

Makram El-Shagi
Sebastian Giesen

September 2010 No. 19

IWH-Diskussionspapiere
IWH Discussion Papers

Testing for Structural Breaks at Unknown Time: A Steeplechase

Makram El-Shagi
Sebastian Giesen

September 2010 No. 19

Authors: *Makram El-Shagi*
Department of Macroeconomics
Phone: +49 345 7753 835
Email: Makram.El-Shagi@iwh-halle.de

Sebastian Giesen
Department of Macroeconomics
Phone: +49 345 7753 804
Email: Sebastian.Giesen@iwh-halle.de

Editor:

Halle Institute for Economic Research (IWH)
Prof Dr Dr h. c. Ulrich Blum (President), Dr Hubert Gabrisch (Head of Research)
The IWH is member of the Leibniz Association.

Address:	Kleine Märkerstraße 8, D-06108 Halle (Saale)
Postal Address:	P.O. Box 11 03 61, D-06017 Halle (Saale)
Phone:	+49 345 7753 60
Fax:	+49 345 7753 820
Internet:	http://www.iwh-halle.de

Testing for Structural Breaks at Unknown Time: A Steeplechase*

Abstract

This paper analyzes the role of common data problems when identifying structural breaks in small samples. Most notably, we survey small sample properties of the most commonly applied endogenous break tests developed by Brown, Durbin, and Evans (1975) and Zeileis (2004), Nyblom (1989) and Hansen (1992), and Andrews, Lee, and Ploberger (1996). Power and size properties are derived using Monte Carlo simulations. Results emphasize that mostly the CUSUM type tests are affected by the presence of heteroscedasticity, whereas the individual parameter Nyblom test and AvgLM test are proved to be highly robust. However, each test is significantly affected by leptokurtosis. Contrarily to other tests, where skewness is far more problematic than kurtosis, it has no additional effect for any of the endogenous break tests we analyze. Concerning overall robustness the Nyblom test performs best, while being almost on par to more recently developed tests in terms of power.

Keywords: structural breaks, heteroscedasticity, skewness, kurtosis, Monte Carlo study

JEL classification: C12, C15

* This paper was prepared within the research group 'Macroeconomic Forecasting and Macroeconomic Policy' at the Halle Institute for Economic Research, Department of Macroeconomics. The authors are indebted to Herbert Buscher, Katja Drechsel, and Rolf Scheufele for valuable comments.

Endogene Strukturbruchtests: Ein Hindernisparcours*

Zusammenfassung

Das vorliegende Papier untersucht die Bedeutung wichtiger Datenanomalien für die Identifikation von Strukturbrüchen in kleinen Stichproben. Dabei wird insbesondere auf die endogenen Strukturbruchtests eingegangen, die von Brown, Durbin und Evans (1975) und Zeileis (2004), Nyblom (1989) und Hansen (1992) sowie Andrews, Lee und Ploberger (1996) entwickelt wurden. Die Fehler erster und zweiter Art werden mit Hilfe von Monte Carlo Simulationen für verschiedene Szenarien analysiert. Die Resultate legen nahe, dass vor allem CUSUM-Tests durch Heteroskedastizität massiv beeinträchtigt werden, während der so genannte Nyblom-Test und der AvgLM-Test sich als sehr robust erweisen. Alle untersuchten Tests haben erhebliche Probleme im Fall starker Leptokurtosis der Fehlerterme. Die Schiefe der Fehlerverteilung hingegen, die für viele andere statistische Testverfahren problematischer ist, beeinflusst die aufgeführten Tests kaum. Insgesamt erweist sich der Nyblom-Test als der robusteste, wobei er gleichzeitig - bezüglich der Power - durchaus mit neueren Testverfahren mithalten kann.

Schlagwörter: Strukturbruch, Heteroskedastizität, Schiefe, Kurtosis, Monte-Carlo-Studie

JEL-Klassifikation: C12, C15

* Dieses Papier ist im Rahmen des Forschungsschwerpunktes 'Makroökonomische Prognosen und Politikanalysen' am Institut für Wirtschaftsforschung in Halle entstanden. Die Autoren sind Herbert Buscher, Katja Drechsel und Rolf Scheufele für wertvolle Anregungen zu Dank verpflichtet.

1. Introduction

The empirical relation between macroeconomic time series is frequently subject to potential changes caused by the evolution of the political framework. Due to the large number of political decisions it is commonly unclear, which of these decisions have sufficiently strong impact on macroeconomic relations to be considered as structural breaks. Testing for unknown structural breaks, to make sure that these policy changes do not alter the parameter regime significantly, thus is crucial for sound economic analysis; see for instance Stock and Watson (1996), McConnell and Perez-Quiros (2000), Hansen (2001), Zeileis, Shah, and Patnaik (2010), and references therein. In macroeconometrics these tests usually have to be applied to small samples due to the low frequency of the data. Even for the US the typical quarterly time series rarely exceeds 200 observations. Furthermore, heteroscedasticity and nonnormalities frequently obscure a clear view on the true underlying processes. While the properties of the standard tests that are employed to test for structural breaks in the data are well known for large samples with error terms that are Gaussian i.i.d., evidence on their performance under the outlined conditions that commonly prevail in macroeconomic analysis is scarce.

This paper provides a detailed inspection of the size and power properties of frequently used endogenous structural break tests applied to small samples using extensive Monte Carlo simulations. For the sake of comparability we restrict the analysis to tests for a single break.[1] The comparative analysis includes the traditional CUSUM test (Brown, Durbin, and Evans 1975) and its refinements proposed by Ploberger and Krämer (1992) and Zeileis (2004), the tests introduced by Nyblom (1989) and Hansen (1992), and the F type tests by Andrews (1993), Andrews and Ploberger (1994), and Andrews, Lee, and Ploberger (1996).[2] A Monte Carlo analysis of the power properties of endogenous structural break tests is provided most notably byAndrews, Lee, and Ploberger (1996). Diebold and Chen (1996) add evidence on the performance in small samples.

[1] An approach to the analysis of data that contains multiple structural changes in a linear regression setup is for instance presented in Zeileis, Kleiber, et al. (2003).

[2] There is a related estimation technique for the determination of structural break dates, that relies on the minimization of the sum of squared residuals (see e.g. Bai (1997) and Bai and Perron (1998)). Since the methodology is very similar to the F type tests, we do not consider these estimators separately. A survey that covers both structural break tests and structural break estimation is found in Perron (2006).

We contribute to this strand of literature by investigating various violations of the normality and i.i.d. assumptions to the simulation setups. More precisely, the simulations include heteroscedasticity (more specifically persistent regime switches in the variance of residuals and autoregressive conditional heteroscedasticity (ARCH)) as well as leptokurtosis and skewness. The simulation results show that - contrarily to other tests - structural break tests are highly vulnerable to excess kurtosis while being fairly robust to skewness.

The remainder of the paper is structured as follows. Section 2 outlines the tests that we will examine. Section 3 describes the simulation setup that we use for our Monte Carlo simulations and the results for a baseline model. In Section 4 we add heteroscedasticty, as well as kurtosis and skewness to the residual component. Section 5 concludes.

2. Methods for Detecting Structural Changes

In the last decades a large number of tests has been developed, to detect structural breaks at unknown points in time. Following Zeileis (2005), these "endogenous structural break tests" can be subdivided in three categories:

The first category, that is commonly referred to as "residual based test" or "fluctuation test", directly relies on the properties of the residual series under the null hypothesis of a constant parameter regime without having an explicit alternative hypothesis. These tests most notably include the original CUSUM test (Brown, Durbin, and Evans 1975) and its refinements (Ploberger and Krämer 1992). For our simulation study, we used the traditional CUSUM and the CUSUM-OLS test with alternative boundaries provided by Zeileis (2004). The small sample correction for these boundaries are proposed by El-Shagi (2010).

The second category of tests builds on the traditional exogenous structural break tests, like the F test that has been proposed by Chow (1960) for this purpose. To identify the most likely break point these tests use the supremum of the F statistic. However, the more recent versions of this test use improved statistics to test whether the null hypothesis of a constant parameter regime should be rejected. We analyze both, the original version of the test proposed by Andrews (1993) and the refinements proposed by Andrews, Lee, and Ploberger (1996).

The third category of tests is based on ML scores. The first test of this type has been developed by Nyblom (1989) for nonlinear models. In our paper we evaluate the alternative version developed by Hansen (1992), that is meant for linear regression models.

All the tests that we consider in this study are applied to test for structural breaks in the parameter regime of a standard linear model:

$$y_t = x_t'\beta + \varepsilon_t, \qquad t = [1, 2, ..., T], \tag{1}$$

where y_t is the dependent variable at time t, x_t the corresponding $(k \times 1)$ vector of exogenous variables and ε_t the residual. The estimated parameters will be marked with a hat in the following.

CUSUM Test: Brown, Durbin, and Evans (1975) proposed a test known as CUSUM test, which is based on the cumulative sum of the recursive residuals. The test statistic W_t is given by:

$$W_t = \frac{1}{\hat{\sigma}\sqrt{T-k}} \sum_{i=k+1}^{t} \hat{\varepsilon}_i, \tag{2}$$

where $\hat{\sigma}$ is defined as:

$$\hat{\sigma} = \sqrt{\frac{\sum_{t=k+1}^{T} \hat{\varepsilon}_t}{T-k}}. \tag{3}$$

In the traditional CUSUM test $\hat{\varepsilon}$ is given by the series of recursive errors that are adjusted for the size distortion:

$$\hat{\varepsilon}_t = \frac{y_t - x_t'\hat{\beta}_{t-1}}{\sqrt{1 + x_t'(X_{t-1}'X_{t-1})^{-1}x_t}}, \tag{4}$$

where $\hat{\beta}_{t-1}$ is the estimate of β using data up to point $t-1$ and X_{t-1} is the corresponding matrix of exogenous variables.

Ploberger and Krämer (1992) introduced an alternative version based on OLS residuals. In here, $\hat{\varepsilon}$ is defined as the common OLS residual i.e.:

$$\hat{\varepsilon}_t = y_t - x_t'\hat{\beta}, \tag{5}$$

If the (estimated) error terms were Gaussian i.i.d., the cumulated sum of errors as given in equation 2 could thus be seen as a standard Brownian motion in the case of the conventional CUSUM test. In the case of Ploberger and Krämer the errors add up to zero by construction such that the cumulated sum of errors can be seen as a Brownian bridge; i.e. as "tied down Brownian motion" for which the starting value and the final value is the same with probability one.[3]

The probability p that W_t exceeds a predefined boundary b_t is constant, if the process b follows the variance of the process that describes the test statistic under the null hypothesis. Therefore, Zeileis (2004) recommends to use boundaries that follow the theoretically derived variances of a Brownian motion for the original CUSUM test and a Brownian bridge for the OLS based test. This is because the variance of a Brownian motion is smaller than the variance of a Brownian bridge.[4]

Thus, a boundary that is passed by the test statistic with a predefined possibility under the null hypothesis is given by:

$$b_t = \lambda\sigma_t, \tag{6}$$

where σ_t is the variance of the relevant stochastic process. These variances are given by:

$$\sigma_t^{CUSUM} = \sqrt{q_t}, \tag{7}$$

in the case of the Brownian motion underlying the original CUSUM test, and by:

$$\sigma_t^{CUSUM-OLS} = \sqrt{q_t(1-q_t)}, \tag{8}$$

in the case of the Brownian bridge underlying the CUSUM-OLS test. In both cases $q_t = t/(T-k)$, i.e. q_t is a normalization of the time to the interval $[0,\ 1]$. The CUSUM tests can then be rewritten as:

[3] See for instance Karatzas and Shreve (1991) for a more detailed description of these processes.
[4] See Hassler (2007).

$$\sup \frac{|W_t|}{b_t} < \lambda, \tag{9}$$

where λ determines the probability that the boundary of interest is crossed at least once.

Zeileis (2004) provides asymptotic estimates of λ for a commonly used set of p values.

Nyblom Test: The Nyblom test (Nyblom 1989, Hansen 1992) describes a simple yet powerful test for parameter instability for a fairly general class of time series models. The null hypothesis of constant parameters is tested against the alternative that the parameters follow a martingale process.[5] It is based on a cumulative sum of the least squares residuals. From the least squares normal equations we can derive:

$$\sum_{t=1}^{T} x_{it}\varepsilon_t = 0 \quad \text{for } i = 1, ..., k \quad \text{and} \quad \sum_{t=1}^{T}(\varepsilon_t^2 - \hat{\sigma^2}) = 0. \tag{10}$$

Following Hansen (1992), we define a $(1 \times (k+1))$-vector f_t for each point in time, where:

$$f_{it} = \begin{cases} x_{it}\hat{\varepsilon}_t, i = 1, ..., k \\ \hat{\varepsilon}_t^2 - \hat{\sigma}, i = k+1. \end{cases} \tag{11}$$

Defining S_{it} as the sum of f_{it} over time

$$S_{it} = \sum_{t=1}^{T} f_{it}, \tag{12}$$

and defining the vectors:

$$f_t = (f_{1t}, ..., f_{m+1t}), \tag{13}$$

$$S_t = (S_{1t}, ..., S_{m+1t}), \tag{14}$$

[5] The test statistic described here is very similar to unit root tests proposed by Kwiatkowski, Phillips, et al. (1992) and Breitung (2002), for instance.

the test statistic L_j can be written as:

$$L_j = \frac{1}{T} \sum_{t=1}^{T} S_t' V^{-1} S_t, \tag{15}$$

with $V = f_t f_t'$. Since S_{k+1} holds the cumulated deviations of squared residuals from average variance of residuals, the Nyblom test does not only respond to changes in the parameters but to changes in the variance of errors as well.

The corresponding test statistic for the individual parameter Nyblom test is given by:

$$L_i = \frac{1}{T V_i} \sum_{t=1}^{T} S_{it}, \tag{16}$$

with $V_i = \sum_{t=1}^{T} f_{it}^2$ for all $i = 1, ..., k+1$. The corresponding nonstandard asymptotic distributions can be found in Nyblom (1989) and Hansen (1990). Since these authors only report critical values for some standard significance criteria, this study employs a bootstrapped distribution of the Nyblom test statistic, to allow testing at any significance level.

SupF Type Tests: SupF type tests are constructed for unknown breakpoints t_b like the tests described above. However, contrarily to the latter, they allow to determine the most likely position of t_b. Therefore, the testing procedure is nonstandard because t_b appears only under the alternative and not under the null hypothesis. How to deal with such a framework is described by Davies (1977) and Hawkins (1987). They proposed the supremum statistics of a Wald test, a likelihood ratio test and a Lagrange multiplier test (SupW, SupLR, and SupLM) to test for structural breaks. Asymptotically, these tests are equivalent. Numerical approximations to the asymptotic distribution (examined by Andrews (1993) and Andrews and Ploberger (1994) for a related class of tests) are given in Hansen (1997). In our paper we follow Hansen and analyze the properties of the SupLM and two more recent extensions that use statistics based on the Lagrange multiplier (AvgLM and ExpLM).

$$\text{SupLM} = \sup_{\pi_1 < t_b < \pi_2} \text{LM}_t(t_b), \tag{17}$$

where t_b denotes the date of the structural change which lies between π_1 and π_2. The corresponding Andrews, Lee, and Ploberger (1996) test statistics are:

$$\text{AvgLM} = \frac{1}{\pi_2 - \pi_1 + 1} \sum_{t=\pi_1}^{\pi_2} \text{LM}_t(t_b), \tag{18}$$

$$\text{ExpLM} = \ln\{\frac{1}{\pi_2 - \pi_1 + 1} \sum_{t=\pi_1}^{\pi_2} \exp(\frac{1}{2}\text{LM}_t(t_b))\}. \tag{19}$$

The corresponding asymptotic distribution and tabulated asymptotic critical values are given in Andrews, Lee, and Ploberger (1996) and Andrews (2003). This paper relies on the approximation provided by Hansen (1997).

3. Simulation Setup

3.1. The baseline model

All tests are applied to a simple linear model with a break at time t_b. The model takes the form:

$$y_t = x_t'\beta_t + \varepsilon_t, \qquad t = [1, 2.., T], \;\; \varepsilon_t \sim N(0, 0.01), \tag{20}$$

where

$$\beta_t = \begin{cases} \beta_t = [0 \;\; 0.5] & \forall t < t_b \\ \beta_t = [0 \;\; 0.5 + \Delta\beta] & \forall t \geq t_b. \end{cases} \tag{21}$$

Table 1: Break intensities Θ and the corresponding values of $\Delta\beta$

Θ	$\Delta\beta$	Θ	$\Delta\beta$
0.005	-0.040	0.055	-0.027
0.010	-0.037	0.060	-0.027
0.015	-0.035	0.065	-0.026
0.020	-0.033	0.070	-0.026
0.025	-0.032	0.075	-0.025
0.030	-0.031	0.080	-0.025
0.035	-0.030	0.085	-0.024
0.040	-0.029	0.090	-0.024
0.045	-0.029	0.095	-0.024
0.050	-0.028	0.100	-0.023

The exogenous time series is given by $[1, x_1]'$, where x_1 is normally distributed with $x_1 \sim N(1, 1)$.[6]

For our simulation we use a model where the constant term equals zero (both, before and after the break). The structural break thus affects the simulated time series correlation of x_1 and y. Albeit the true process has no constant term the tests are performed allowing for a constant. This is especially important for the CUSUM type tests that produce biased results if applied to a model without constant.

In our baseline simulations the break occurs exactly in the middle of the sample. We test a broad range of break intensities, where we understand the intensity Θ of a break $\Delta\beta$ as one minus the significance level of a two-sided t test that compares the parameter regime before and after the break, if the break was known. Table 1 summarizes selected break intensities and the corresponding values of $\Delta\beta$.

The Monte Carlo analysis includes 10'000 simulations for each of the 100 break intensities Θ, that are equally distributed over the interval [0.9 0.999]. That is, we only include breaks, where the null hypothesis of no break could be rejected with at least 90% probability, if the break point was known.

[6] Note that the exogenous variable has a nonzero mean. This guarantees that angle ψ between the average exogenous vector and the shift of the parameter coefficient $\Delta\beta$ differs from 90°. This is important, since the CUSUM family of tests, is not able to detect breaks that do not fulfill this condition.

To test whether the empirical size matches the nominal size of the test[7], further 100'000 repetitions with no break are simulated.

All tests are made for 100 observations, where we look for breaks between the $16th$ and $85th$ observation.

3.2. Power and size of structural break tests in the baseline model

Under Gaussian i.i.d. errors the empirical size of most tests equals nominal size. In small samples the CUSUM type tests with alternative boundaries commonly find less breaks (under the null hypothesis) than suggested by the nominal p value. This size distortion is mitigated by using the size adjusted critical values proposed by El-Shagi (2010).

The empirical size of the Nyblom test equals nominal size for both joint test and individual parameter test.

The empirical size of the F type tests also generally matches nominal size. The only exception is the original SupLM test. The null hypothesis is falsely rejected with a probability that is roughly equal to two thirds of nominal size. This corresponds to the results of Diebold and Chen (1996), who find a tendency to underreject for this type of test in an AR(1) setup. Contrarily to the SupLM test, AvgLM and ExpLM both reject the null with the expected probability.

Figure B.1 in the appendix shows the power properties of the CUSUM, CUSUM-OLS, Nyblom, SupLM, AvgLM and ExpLM tests for a range of break sizes.

Unsurprisingly, the CUSUM type tests have a very low power, although the OLS type tests already constitute a substantial power improvement compared to the traditional recursive residual setup. Although our setup guarantees that the break (i.e. the vector $\Delta\beta$) is orthogonal to the average exogenous vector and thus maximizes the power of CUSUM type tests, breaks are mostly not recognized as such. Even the strongest breaks that are included in our Monte Carlo experiment are found with a probability of only 25% with a CUSUM test using a 10 % significance criterion. The

[7] That is, whether the probability of a type 1 error matches the probability that is defined by the required significance level.

power of the CUSUM-OLS test to find breaks of this size is about 50%. While this doubles the power of the CUSUM test, the CUSUM-OLS test still is the second worst of all tests included. The power of traditional CUSUM tests, that do not rely on the alternative boundaries, is higher if the break occurs in the middle of the sample. However, this additional power is bought at the cost of a size distortion.

The joint parameter Nyblom test performs substantially worse than the joint parameter AvgLM and ExpLM test, that perform best for most scenarios. On average the power difference is about 10 percentage points. However, this is mostly due to the Nyblom test including variance stability in the joint parameter version. This adds an additional level of uncertainty that in turn reduces power. Accordingly, the power of an individual parameter Nyblom test almost matches the power of the AvgLM and ExpLM tests.[8] However, the individual parameter Nyblom test still performs worse if very rigid significance criteria are employed. Especially strong breaks are rarely identified compared to other tests if the significance level is 1% or less.

The F type tests perform quite well. However, the SupF test has substantially lower power than the AvgLM and ExpLM tests. This lower power, that roughly corresponds to the power of the joint Nyblom test, is mostly due to the size distortion caused by the small sample size. Scaled to empirical size instead of nominal size, the power of the SupF test is more or less on par with the other F type tests.

The ExpLM test outperforms the AvgLM test if very strict significance criteria are employed. However, the AvgLM test performs consistently better if the critical p value is larger than one percent. That means, that in small samples the AvgLM test should be favored given the traditionally employed significance criterion of 5%. Anyhow, the power difference only is about 2% in favor of the AvgLM test.

Even the ExpLM and AvgLM tests detect a break of intensity Θ with a probability of less than 40% on the corresponding significance level $p_{crit} = (1 - \Theta)$. This lack of power shows the general difficulties of testing for breaks if the break point is unkown.

[8] Since our Monte Carlo setup includes only one regression parameter and the F type tests do not test for additional parameters of the setup, the joint and individual parameter F type tests are equivalent. This allows direct comparison of the individual parameter Nyblom test and the joint parameter F type tests in our setup.

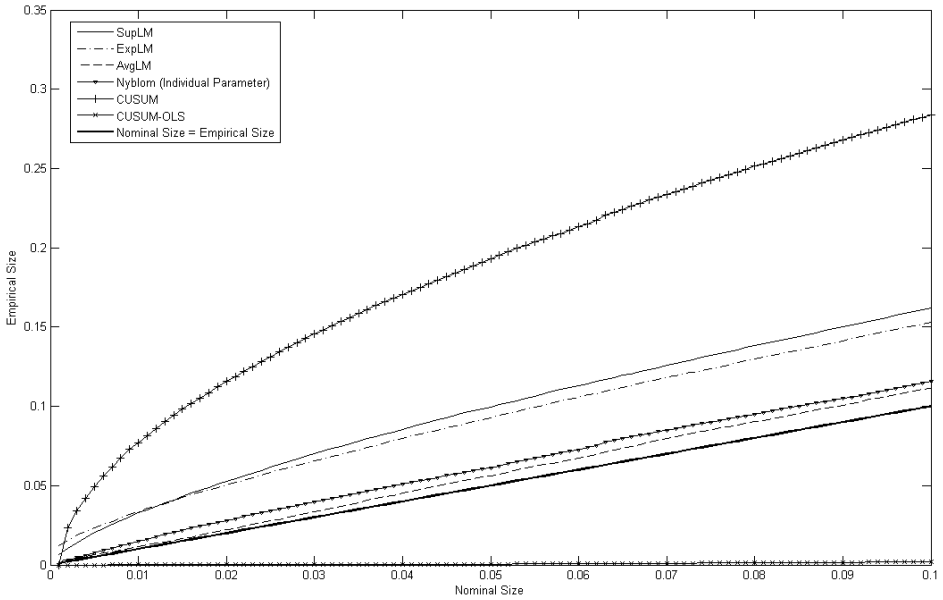


Figure 1: Size of structural break tests with break in variance

4. Simulation under common data and model problems

4.1. Heteroscedasticity

The analysis comprises two kinds of heteroscedasticity: first, a model with a single break in the variance, and secondly, a model with autoregressive conditional heteroscedasticity (ARCH).

We compare SupLM, AvgLM, ExpLM, individual parameter Nyblom and the CUSUM type tests concerning their power and size properties in the presence of heteroscedasticity. Since the joint parameter Nyblom test is designed to capture changes in the variance of errors, it strongly reacts to heteroscedasticity by construction. This is intentional rather than a sign of distorted size. Therefore, the joint parameter Nyblom test is not included in this section.

4.1.1. Variance regimes

The most basic form of heteroscedasticity is regime change in the variance of the idiosyncratic error term. Corresponding to our structural break setup, this break occurs in the middle of the sample in the simulations.

More precisely, the simulation uses the following setup:

$$y_t = x_t'\beta + \varepsilon_t \tag{22}$$
$$\varepsilon_t \sim N(0, \sigma_t) \begin{cases} \sigma_t = 0.01 & \forall t < t_b \\ \sigma_t = 0.19 & \forall t \geq t_b. \end{cases}$$

Surprisingly, the F type tests exhibit strong differences in their robustness to heteroscedasticity as can be seen in figure 1. While the rate of false rejects increases substantially in the case of the SupLM and the ExpLM tests, the empirical size of the AvgLM test remains closer to the nominal size. The individual parameter Nyblom test performs very similar to AvgLM. Both versions of the CUSUM test react heavily to heteroscedasticity. While the rate of type one errors increases to three times the nominal size in the case of the OLS CUSUM, the size (and correspondingly the power) of the standard CUSUM test is reduced almost to zero. However, the reaction of the latter test (standard CUSUM) depends strongly on the specific model setup. This is discussed in more detail in the technical appendix.

4.1.2. Autoregressive conditional heteroscedasticity

Frequently, a change in the variance of the error term is not due to a permanent change. Rather, periods of high volatility induce further volatility; therefore times of high volatility alternate with fairly stable times. This is mostly captured using models of autoregressive conditional heteroscedasticity.

The model used in this paper is a standard ARCH(1) model, where the conditional variance of the error term in t only depends on the error term in $t-1$.:

$$\begin{aligned} y_t &= x_t'\beta + \varepsilon_t, \\ \sigma^2_{\varepsilon,t} &= \alpha_1 + \alpha_2 \varepsilon^2_{t-1}. \end{aligned} \tag{23}$$

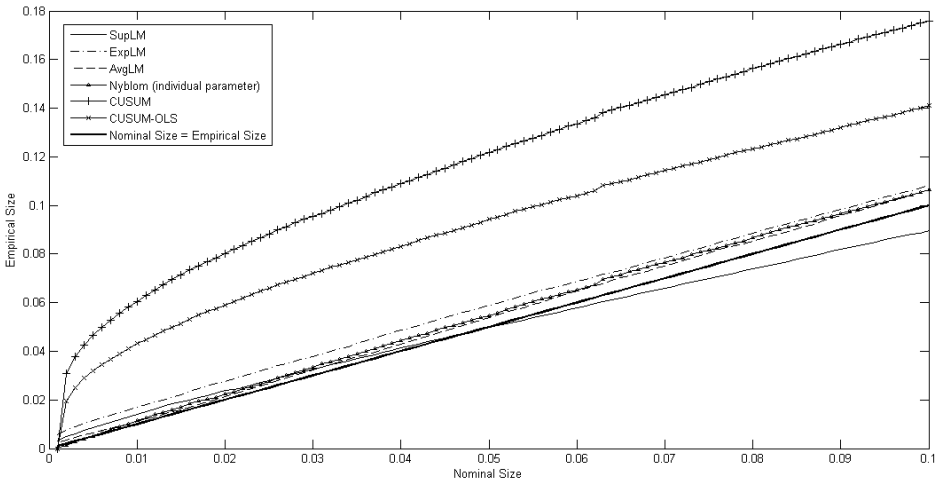


Figure 2: Size of structural break tests with ARCH

Except the CUSUM type tests the structural break tests perform reasonably well in the presence of ARCH effects. While the empirical size increases slightly for the other tests (to about 10.6% at a 10% significance level) with a corresponding change in power, this should still be feasible for practical purposes. Figure 2 provides a visual inspection of empirical size for all tests.

4.2. Non normal error terms

Standard econometric tests mostly rely on the assumption of Gaussian errors. For practical purposes this assumption is usually reduced to the requirement that the kurtosis and skewness of the empirical error distribution do not differ significantly from the respective moments of the normal distribution. However, Monte Carlo simulations suggest that valid statistical inference is far more sensitive to skewness than to excess kurtosis (see e.g. Jarque and Bera (1980) and Bai and Ng (2005)).

Therefore, we will first analyze the robustness of endogenous structural break tests to error terms that exhibit excess kurtosis. In a second subsection we will draw the simulated error terms from a distribution that is skewed.

Following Mantalos and Shukur (2007), who analyze the properties of the RESET test under nonnormality, we employ distributions of the Generalized-Tukey-Lambda (GTL) family. Based on Freimer, Kollia, et al. (1988) we use the specification:

Table 2: Standardized Moments of the Error Distributions

	N(0,1)	$\Psi_{-0.2,-0.2}$	$\Psi_{-0.7,-0.7}$	$\Psi_{0.7,-0.1325}$
μ	0	0	0	0
σ^2	1	1	1	1
Skewness	0	0	0	2.04
Kurtosis	3	10.9	735	10.9

$$Q(U) = \lambda_4 + \left[\frac{U^{\lambda_1} - 1}{\lambda_1} - \frac{(1-U)^{\lambda_2} - 1}{\lambda_2}\right] / \lambda_3, \tag{24}$$

where Q is the quantile function and U is a uniform $(0, 1)$ random variable. The parameter combination λ_1 and λ_2 determines skewness and kurtosis of the distribution. The distribution is non skewed if (and only if) $\lambda_1 = \lambda_2$. The additional parameters λ_3 and λ_4 can be used to adjust variance and mean respectively. In our paper, both of these are set to assure that the distribution has a zero mean and variance of 0.01, that is to match the first moments of the error distribution with those of the baseline specification.

For simplicity we refer to the distributions that are defined by the parameter combination λ_1 and λ_2 as $\Psi_{\lambda_1,\lambda_2}$.

4.2.1. Kurtosis

To test the impact of kurtosis on the power and size properties of the structural break tests we use two alternative setups with excess kurtosis, where the errors are drawn from a $\Psi_{-0.2,-0.2}$ and a $\Psi_{-0.7,-0.7}$ respectively.[9] Figure B.2 in the appendix shows the density functions of these distribution scaled to a variance of 1 compared to the standard normal distribution that is given as a reference. Table 2 provides the relevant standardized moments for the respective distributions.

As with skewness we run 100′000 repetitions to test empirical size. We find that the empirical size differs significantly from nominal size for each test. This holds true for both setups, slight and extreme leptokurtosis. An overview of empirical size, compared to nominal size is given in Figure 3 for the extreme setup. Table B.1 in the appendix summarizes the empirical size for standard p-values for both setups.

[9] In general small values of λ_1 and λ_2 lead to a high kurtosis.

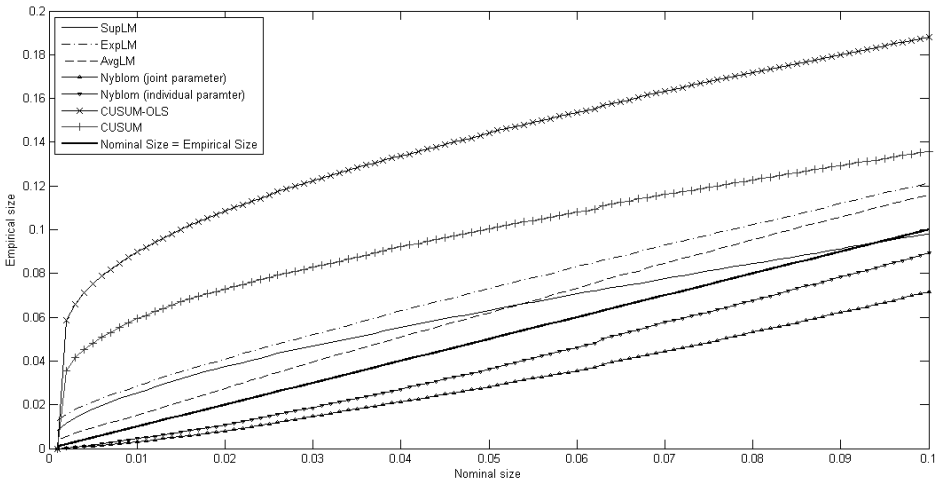


Figure 3: Size of structural break tests with strong excess kurtosis

Interestingly, while the empirical size of SupLM tests and CUSUM type tests increases compared to the baseline specification, the empirical size of Nyblom tests is substantially reduced. Especially the size—and correspondingly the power—of the joint parameter test is strongly affected by kurtosis problems. Even in the setup with slight excess kurtosis the empirical size is reduced by about 15% compared to nominal size. In the extreme scenario, the difference between nominal and emprical size increases to almost 30% with a corresponding loss of power.

Both, AvgLM and ExpLM exhibit an increase in the empirical size. However, this effect is almost negligible in the setup with less excess kurtosis. Since extreme cases of leptokurtosis as given by the $\Psi_{-0.7,-0.7}$ distribution, are scarce, these tests seem sufficiently robust for most practical applications. While the SupLM tests performs quite well, it should be taken into account that empirical size in the baseline model is about two thirds of nominal size. That is, although the empirical size of the SupLM test still is below nominal size, empirical size increases due to leptokurtosis of the error distribution. Thus, the general problem of the test is only incidentally quantitatively compensated quite exactly by this specific violation of the assumptions. Therefore, this should not be interpreted as a signal of robustness to excess kurtosis. Also, the size of the SupLM test is affected extremely strong, at the most common significance level of 5%, making it very hard to interpret the results.

Again, the CUSUM type tests are hit hardest by the violation of the Gaussian i.i.d. assumption. Since they rely on the analysis of the error terms, this is not very

surprising. While the CUSUM type tests might be taken as a very conservative approach, if the Gaussian i.i.d. assumption holds, they are highly unreliable from a practical perspective, where this can be rarely taken to be guaranteed.

4.2.2. Skewness

The errors in the setup that is employed to test the robustness to a skewed error distribution is based on a $\Psi_{0.7,-0.1325}$ distribution, as depicted in Figure B.3 in the appendix. This distribution is chosen to match the moments of the $\Psi_{-0.2,-0.2}$ distribution, that has been used in the last section.[10] The moments are summarized in Table 2.

The results are virtually identical to those achieved with slight kurtosis but without skewness. Thus, while we again find a minor increase in the empirical size of the LM-tests, a decrease in the size of the joint parameter Nyblom test, and excessive size distortions in the case of the CUSUM type tests, we are able to attribute this to kurtosis. Contrarily to other tests, where skewness is far more problematic than kurtosis, it has no additional effect for any of the endogenous break tests that we analyze.

5. Conclusion

In this paper we analyze the role of common data problems when identifying structural breaks in small samples. These data problems involve serveral forms of heteroscedasticity, as well as skewness, and kurtosis being present in the residual series. We survey the most commonly applied endogenous break tests, such as the CUSUM and CUSUM-OLS, the joint and individual parameter Nyblom test, and the Lagrange multiplier tests SupLM, AvgLM, and ExpLM. To investigate power and size properties we used Monte Carlo simulations.

Results emphasize that the structural break tests perform reasonably well in the presence of ARCH effects, except the CUSUM type tests. The individual Nyblom test

[10] There are no suitable GTL distributions that are skewed but match the kurtosis of the normal distribution. To be able to separate kurtosis effects from skewness effects, we thus match the kurtosis of the distribution used in the last section.

and the F type tests are proved to be robust. However, persistent regime switches in the variance of the error term severely compromise the performance of most structural break tests. The Nyblom and AvgLM tests are the notable exceptions, although even these exhibit an increase in the rate of false rejects.

Each test is affected by error terms that are not Gaussian. There are substantial differences in empirical and nominal size for each test when excess kurtosis is present. However, opposed to other tests, where skewness is far more problematic than kurtosis, it has no additional effect for any of the endogenous break tests that we analyze. Especially the AvgLM and ExpLM tests - that perform best on average - tend to produce a high rate of type one errors if these tests are employed. Contrarily, the empirical size of the Nyblom test is below the nominal size under these conditions. Thus, if looking for a conservative while sufficiently powerful test the Nyblom suggests itself.

References

Andrews, D. W. K. (1993): Tests for Parameter Instability and Structural Change with Unknown Change Point. *Econometrica* 61.4, pp. 821–856.

Andrews, D. W. K. (2003): Tests For Parameter Instability and Structural Change With Unknown Change Point: A Corrrigendum. *Econometrica* 71.1, pp. 395–397.

Andrews, D. W. K., I. Lee, and W. Ploberger (1996): Optimal Changepoint Tests for Normal Linear Regression. *Journal of Econometrics* 70.1, pp. 9–38.

Andrews, D. W. K. and W. Ploberger (1994): Optimal Tests When a Nuisance Parameter Is Present Only Under the Alternative. *Econometrica* 62.6, pp. 1383–1414.

Bai, J. (1997): Estimating multiple breaks one at a time. *Econometric Theory* 13.3, pp. 315–352.

Bai, J. and S. Ng (2005): Tests for Skewness, Kurtosis, and Normality for Time Series Data. *Journal of Business and Economic Statistics* 23.1, pp. 49–60.

Bai, J. and P. Perron (1998): Estimating and Testing Linear Models with Multiple Structural Changes. *Econometrica* 66.1, pp. 47–78.

Breitung, J. (2002): Nonparametric Tests for Unit Roots and Cointegration. *Journal of Econometrics* 108.2, pp. 343–363.

Brown, R., J. Durbin, and J. Evans (1975): Techniques for testing the constancy of regression relationships over time. *Journal of the Royal Statistical Society* B 37, pp. 149–163.

Chow, G. C. (1960): Tests of Equality Between Sets of Coefficients in Two Linear Regressions. *Econometrica* 28.3, pp. 591–605.

Davies, R. B. (1977): Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika* 64.2, pp. 247–254.

Diebold, F. X. and C. Chen (1996): Testing structural stability with endogenous breakpoint: A size comparison of analytic and bootstrap procedures. *Journal of Econometrics* 70.1, pp. 221–241.

El-Shagi, M. (2010): "Small Sample Correction for the Alternative CUSUM-Tests." mimeo.

Freimer, M., G. Kollia, G. S. Mudholkar, and C. T. Lin (1988): A study of the generalized tukey lambda family. *Communications in Statistics-Theory and Methods* 17.10, pp. 3547–3567.

Hansen, B. E. (1990): "Lagrange Multiplier Tests for Parameter Instability in Non-Linear Models." University of Rochester.

Hansen, B. E. (1992): Testing for parameter instability in linear models. *Journal of Policy Modeling* 14.4, pp. 517–533.

Hansen, B. E. (1997): Approximate asymptotic p-values for structural change tests. *Journal of Business and Economic Statistics* 15.1, pp. 60–67.

Hansen, B. E. (2001): The New Econometrics of Structural Change: Dating Breaks in U.S. Labor Productivity. *The Journal of Economic Perspectives* 15.4, pp. 117–128.

Hassler, U. (2007): Stochastische Integration und Zeitreihenmodellierung. Ed. by Holger Dette and Wolfgang Härdle. Springer.

Hawkins, D. (1987): A test for a change point in a parametric model based on a maximal Wald-type statistic. *Sankhya: The Indian Journal of Statistics* 49.Series A, pp. 368–376.

Jarque, C. M. and A. K. Bera (1980): Efficient Tests for Normality, Homoscedasticty, and Serial Independence of Regression Residuals. *Economics Letters* 6.3, pp. 255–259.

Karatzas, I. and S. E. Shreve (1991): Brownian Motion and Stochastic Calculus. Ed. by S. Axler and F.W. Gehring and K.A. Ribet. Springer.

Kwiatkowski, D., P. C. B. Phillips, P. Schmidt, and Y. Shin (1992): Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root?: *Journal of Econometrics* 54.1-3, pp. 159–178.

Mantalos, P. and G. Shukur (2007): The Robustness of the RESET Test to Non-Normal Error-Terms. *Computational Economics* 30.4, pp. 393–408.

McConnell, M. M. and G. Perez-Quiros (2000): Output Fluctuations in the United States: What Has Changed Since the Early 1980's?. *The American Economic Review* 90.5, pp. 1464–1476.

Nyblom, J. (1989): Testing for the constancy of parameters over time. *Journal of the American Statistical Association* 84.405, pp. 223–230.

Perron, P. (2006): "Dealing with Structural Breaks." Palgrave Handbook of Econometrics. Ed. by T. Mills and K. Patterson. Vol. 1. Palgrave Macmillan, pp. 278–352.

Ploberger, W. and W. Krämer (1992): The CUSUM test with OLS residuals. *Econometrica* 60.2, pp. 271–285.

Stock, J. H. and M. W. Watson (1996): Evidence on Structural Instability in Macroeconomic Time Series Relations. *Journal of Business & Economic Statistics* 14.1, pp. 11–30.

Zeileis, A. (2004): Alternative Boundaries for CUSUM Tests. *Statistical Papers* 45.1, pp. 123–131.

Zeileis, A. (2005): A Unified Approach to Structural Change Tests Based on ML Scores, F Statistics, and OLS Residuals. *Econometric Reviews* 24.4, pp. 445–466.

Zeileis, A., C. Kleiber, W. Krämer, and K. Hornik (2003): Testing and Dating of Structural Changes in Practice. *Computational Statistics & Data Analysis* 44.1–2, pp. 109–123.

Zeileis, A., A. Shah, and I. Patnaik (2010): Testing, monitoring, and dating structural changes in exchange rate regimes. *Computational Statistics & Data Analysis* 54.6, pp. 1696–1706.

A. Technical Appendix

Heteroscedasticity and the CUSUM Test

The CUSUM test is highly sensitive to the order of variance regimes. While a sequence where a high variance regime follows a low variance regime can hide structural breaks (i.e. reduce the power of the test), a sequence where a high variance regime precedes a low variance regime is frequently erroneously mistaken as a structural break in the parameter regime.

This can easily be seen in figure A.1. The solid line is the 90% quantile of the absolutes of a Brownian motion with a variance $\sigma = 1$. Normalizing the residual size to one, the CUSUM test (using the alternative boundaries that are provided by Zeileis (2004)) rejects the null, if the absolute of the cumulative sum of recursive residuals passes this line.

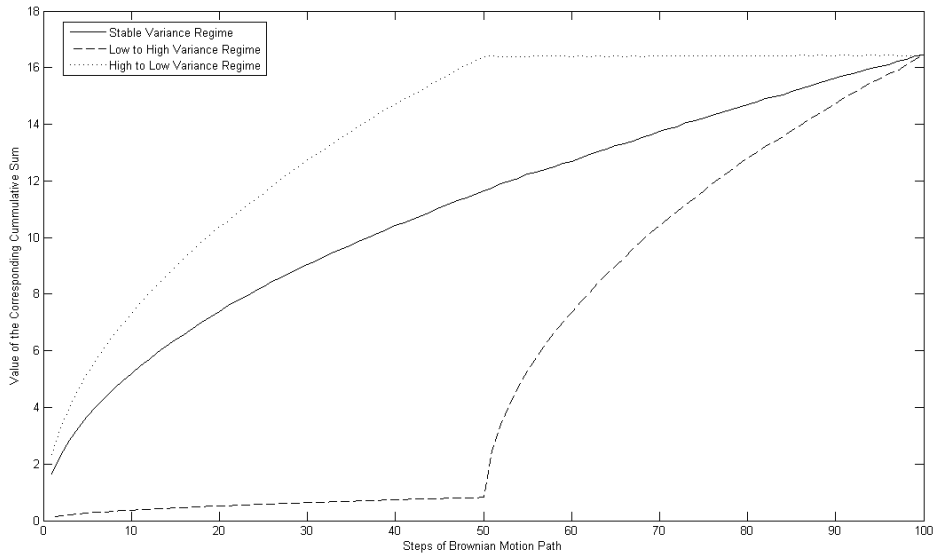


Figure A.1: 90% Quantiles of Brownian Motions with different variance regimes

The upper, dotted line is the corresponding 90% quantile of the absolutes of a Brownian motion with a reduction in the variance after the 50^th step. The lower, dashed line gives the same for a Brownian moten with the reverse order of variance regimes, i.e. high variance follows low variance. Both Brownian motions use an average variance of $\sigma = 1$ of the entire sample of 100 steps.

Given the identical variance all three plots reach the same point after 100 steps. However, the pathes towards this joint target differ drastically. If the variance starts on a low level, there is virtually no chance that the Brownian motion passes the threshold defined by the solid line in the initial steps. After 50 steps the distance to this threshold is so very large, that it is highly unlikely that the high variance in the following steps suffices to drive the Brownian motion far enough to surpass the critical value.

If however, the high variance regime precedes the low variance regime, it is very likely that the threshold is surpassed very quickly. After all, this ordering effectively means that the high variance regime has to face the critical values defined by a medium variance regime, without starting at a lower level.

All other tests that we analyze in this paper perform similarly under both setups. This can be seen when comparing figure 1 with figure A.2.

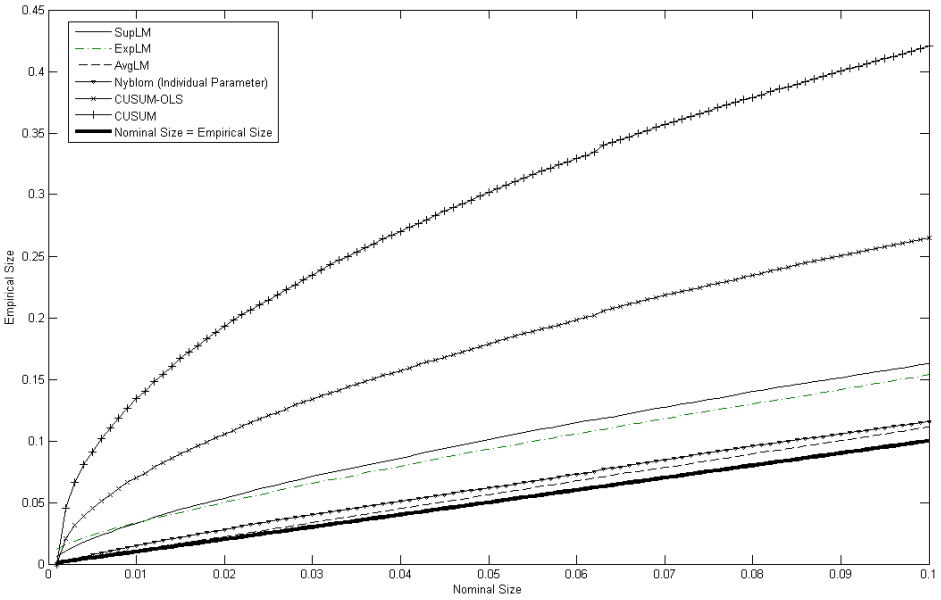
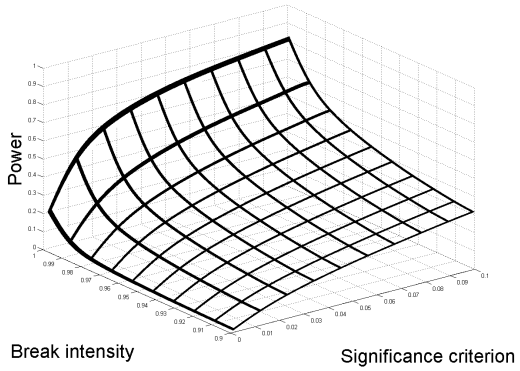


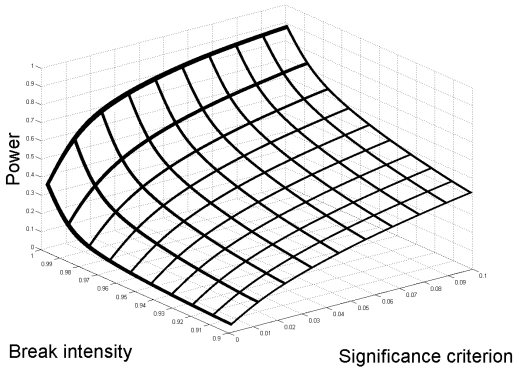
Figure A.2: Size of structural break tests with break in variance

B. Graphics and Tables

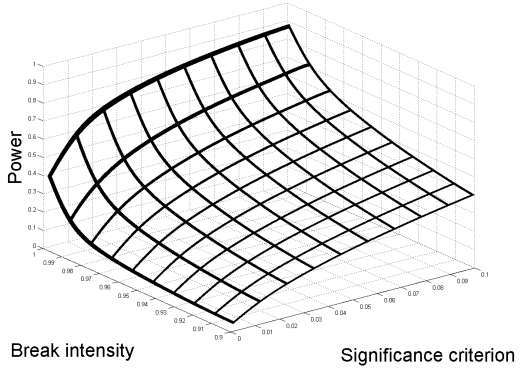
(a) SupLM-Test



(b) AvgLM-Test



(c) ExpLM-Test



(d) Joint Nyblom-Test

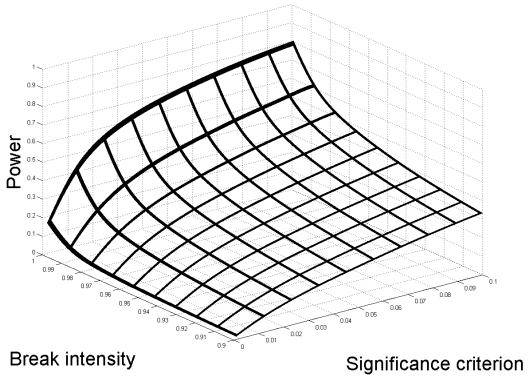
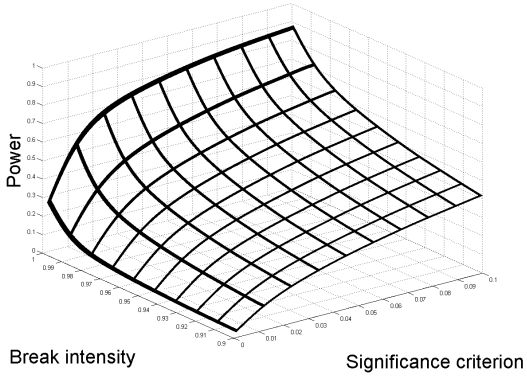
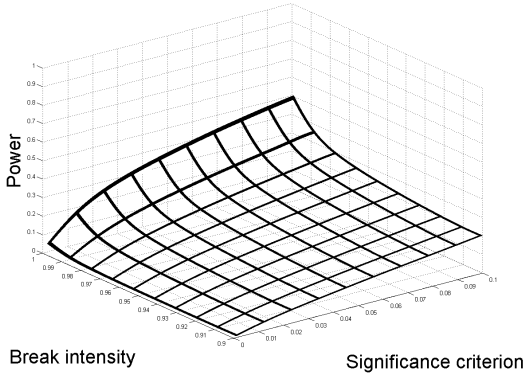


Figure B.1: Power of structural break tests for different break sizes

(e) Individual Nyblom-Test



(f) CUSUM-OLS



(g) CUSUM

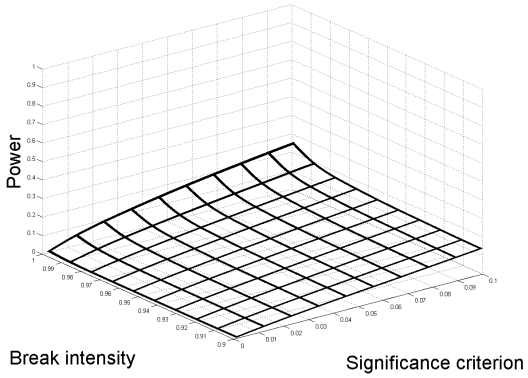


Figure B.1(cont.): Power of structural break tests for different break sizes

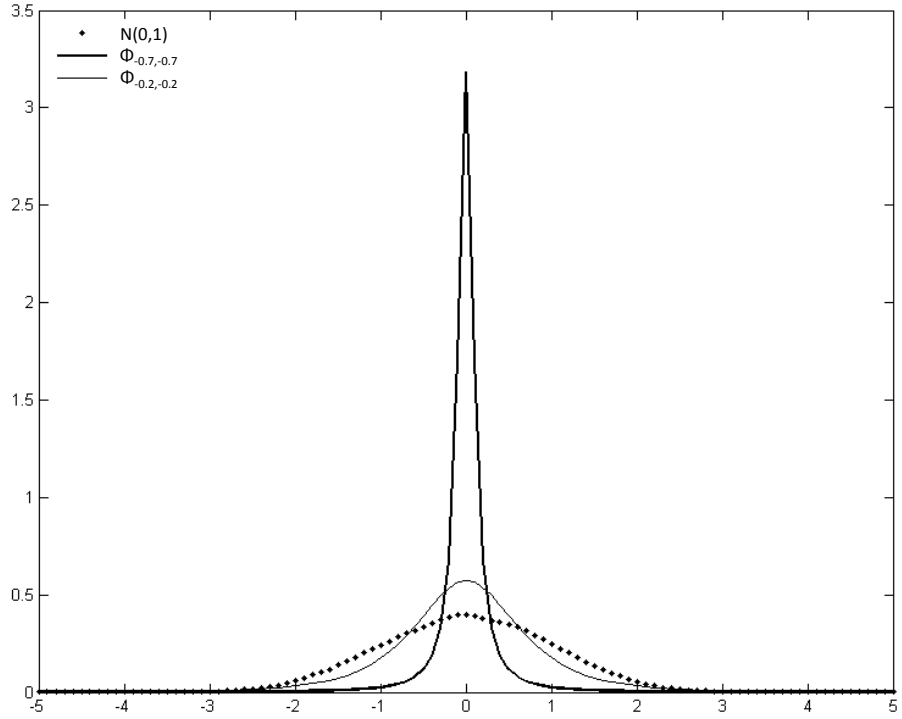


Figure B.2: Generalized Tukey Lambda Distributions with Kurtosis

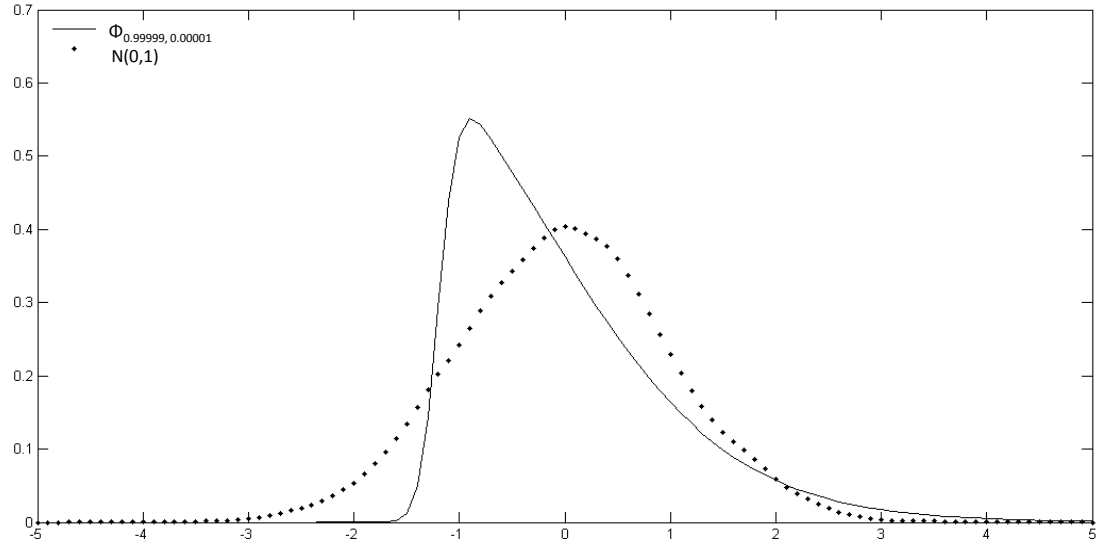


Figure B.3: Generalized Tukey Lambda Distribution with Skewness

Table B.1: Empirical size of endogenous break tests under excess kurtosis

$\Psi_{-0.2,-0.2}$	SupLM	ExpLM	AvgLM	JNyblom	INyblom	CUSUM-OLS	CUSUM
0.100	0.07392	0.10068	0.10221	0.09091	0.09907	0.13103	0.11801
0.050	0.03758	0.05160	0.04978	0.04122	0.04858	0.08027	0.07209
0.025	0.01878	0.02658	0.02368	0.01859	0.02328	0.05199	0.04657
0.010	0.00790	0.01165	0.00914	0.00634	0.00870	0.03099	0.02683
$\Psi_{-0.7,-0.7}$							
0.100	0.09796	0.12140	0.11589	0.07156	0.08938	0.18801	0.13560
0.050	0.06308	0.07303	0.06195	0.02827	0.03632	0.14414	0.10032
0.025	0.04229	0.04668	0.03369	0.01111	0.01472	0.11575	0.07827
0.010	0.02552	0.02847	0.01536	0.00299	0.00449	0.08974	0.05923

Note: $\Psi_{-0.2,-0.2}$ and $\Psi_{-0.7,-0.7}$ denote the two alternative excess kurtosis setups we use. The corresponding significance levels are 10%, 5%, 2.5%, 1%. Power is given as a share of correctly identified breaks.