# MONITORING STRUCTURAL CHANGE IN DYNAMIC ECONOMETRIC MODELS

Achim Zeileis[a*] Friedrich Leisch[b] Christian Kleiber[c] Kurt Hornik[a]

[a] *Institut für Statistik und Mathematik, Wirtschaftsuniversität Wien, Austria*

[b] *Institut für Statistik & Wahrscheinlichkeitstheorie, Technische Universität Wien, Austria*

[c] *Institut für Wirtschafts- und Sozialstatistik, Universität Dortmund, Germany*

SUMMARY

The classical approach to testing for structural change employs retrospective tests using a historical data set of a given length. Here we consider a wide array of fluctuation-type tests in a monitoring situation – given a history period for which a regression relationship is known to be stable, we test whether incoming data are consistent with the previously established relationship. Procedures based on estimates of the regression coefficients are extended in three directions: we introduce (a) procedures based on OLS residuals, (b) rescaled statistics and (c) alternative asymptotic boundaries. Compared to the existing tests our extensions offer ease of computation, improved size in finite samples for dynamic models and better power against certain alternatives, respectively. We apply our methods to three data sets, German M1 money demand, U.S. labor productivity and S&P 500 stock returns.

**Keywords:** Online monitoring, empirical fluctuation process, CUSUM, MOSUM, moving estimates, recursive estimates.

**JEL classification:** C22, C52.

## 1. INTRODUCTION

Structural stability is of prime importance in applied time series econometrics. Estimates derived from unstable relationships erroneously considered as stable are not meaningful, inferences can be severely biased, and forecasts lose accuracy. In a comprehensive study using a sample of 76 representative US monthly time series and several thousand forecasting relations derived from these, Stock and Watson (1996) found evidence for parameter instability in a substantial fraction of their models.

The by now standard approach to the detection of structural changes attempts to detect breaks ex post, see Hansen (2001) for a state of the art survey. Starting with the pioneering work of Chu, Stinchcombe and White (1996) a second line of research has emerged: given that in the real world new data arrive steadily it is frequently more natural to check whether incoming data are consistent with a previously established relationship, i.e., to employ a monitoring approach.

While such problems have been considered in statistical quality control for a long time, similar questions also arise in econometric regression models. Below we adopt a sequential testing approach in the context of dynamic regression models and implement a variety of procedures for the monitoring situation.

*Correspondence to: Achim Zeileis, Institut für Statistik und Mathematik, Wirtschaftsuniversität Wien, Augasse 2–6, A-1090 Wien, Austria. E-mail: Achim.Zeileis@wu-wien.ac.at

This is a preprint of an article published in *Journal of Applied Econometrics* **20**: 99–121 (2005).

As of today, there are essentially two theoretical papers about monitoring in the econometrics literature (Chu et al., 1996, Leisch, Hornik and Kuan, 2000). When we wrote the first version of this paper, we believed it was the first to apply monitoring methodology in dynamic econometric models. However, when our paper was under review Carsoule and Franses (2003) appeared which contains a brief illustration utilizing a U.S. industrial production index. In this paper, we demonstrate the usefulness of these methods in applications to a wide array of data sets: German M1 money demand (Lütkepohl, Teräsvirta and Wolters, 1999) – where one might suspect a structural shift following the German monetary union in 1990 – U.S. labor productivity (Hansen, 2001), and S&P 500 stock returns. In contrast to Carsoule and Franses (2003), we consider various modifications and extensions of the "plain" Chu et al. (1996) and Leisch et al. (2000) methods which, in our view, are crucial in empirical work.

On the methodological side, tests for structural change suffer from a huge alternative as there are infinitely many conceivable ways of deviation from the null hypothesis of structural stability. In econometric literature, two classes of structural change tests emerged that deal with this problem in different ways: $F$ tests are designed for a single-shift (of unknown timing) alternative and enjoy certain (weak) optimality properties in this setup (Andrews, 1993, Andrews and Ploberger, 1994). Fluctuation tests, on the other hand, do not assume a particular pattern of structural change. The generalized fluctuation test framework "includes formal significance tests but its philosophy is basically that of data analysis as expounded by Tukey .... Essentially, the techniques are designed to bring out departures from constancy in a graphic way instead of parameterizing particular types of departure in advance and then developing formal significance tests intended to have high power against these particular alternatives." (Brown, Durbin and Evans, 1975, pp. 149–150). To be able to capture different types of structural changes, fluctuation tests can either be based on sequences (so-called "(empirical) fluctuation processes") of estimates of the regression coefficients or on regression residuals (recursive or OLS), both from a widening data window or from a moving window of fixed size. Due to the broad alternative this rich variety of tests proved very useful in practice (Stock and Watson, 1996), e.g., contrary to $F$ statistics the moving window tests are able to capture certain double structural changes where parameters temporarily deviate from a "normal" level (Chu, Hornik and Kuan, 1995). The probably best-known test from the fluctuation test framework is the recursive (or standard) CUSUM test introduced by Brown et al. (1975), later extended by Krämer, Ploberger and Alt (1988) to dynamic models. A unifying view on fluctuation-type tests in historical samples is provided by Kuan and Hornik (1995). These tests are commonly used to detect structural change *ex post* and are here referred to as historical tests. The class of fluctuation tests can be extended to the monitoring of structural changes, i.e., for detecting shifts *online*. Chu et al. (1996) introduced the first fluctuation test for monitoring by extending the recursive estimates test of Ploberger, Krämer and Kontrus (1989). Leisch et al. (2000) generalized these results and established a class of estimates-based fluctuation tests for monitoring.

In view of the broad alternative it would seem that what the practitioner requires is (1) a wide variety of tests and (2) tools that are helpful to understand the type of deviation from the null hypothesis of stability. Regarding (1), we extend several ideas from the historical to the monitoring framework. For instance, we consider processes based on OLS residuals as these are not only easy to compute and interpret but also able to capture various types of structural changes. Furthermore, we consider rescaled estimates-based processes in order to improve empirical size which is especially important for practical purposes. Finally we suggest alternative boundaries for the Brownian bridge in order to improve power against certain alternatives. Regarding (2), we show how to combine formal significance tests with a graphical analysis of empirical fluctuation processes. As argued by (Cleveland, 1993, pp. 12–14), visualization and probabilistic inference can supplement each other: "Visualization—with its two components, graphing and fitting—. . . stresses a penetrating look at the structure of the data." Empirical fluctuation processes are eminently suitable for this purpose.

All applications are carried out in the statistical software package R (R Development Core Team, 2003, see http://www.R-project.org/), which offers a growing collection of methods useful to econometricians and therefore finds more and more attention in the econometrics community (Cribari-Neto and Zarkos, 1999, Racine and Hyndman, 2002). All methods introduced are available in the package `strucchange` (Zeileis, Leisch, Hornik and Kleiber, 2002), which reflects the common features of the procedures and offers a modern computational and graphical approach to testing for structural change. Empirical fluctuation processes as well as $F$ statistics can be fitted and graphed and formal significance tests can be carried out.

The rest of this paper is structured as follows. Section 2 presents the model. We briefly summarize the results of Chu et al. (1996) and Leisch et al. (2000) at the beginning of Section 3 and then extend the class of fluctuation tests for monitoring in three directions as mentioned above. In Section 4 we apply the methods introduced as well as some historical tests to three data sets. Our conclusions are summarized in Section 5.

## 2. THE MODEL

Consider the standard linear regression model

$$y_i = x_i^\top \beta_i + u_i \qquad (i = 1, \dots, n, n+1, \dots), \tag{1}$$

where at time $i$, $y_i$ is the observation of the dependent variable, $x_i = (1, x_{i2}, \dots, x_{ik})^\top$ is a $k \times 1$ vector of regressors, with the first component usually equal to unity, and $\beta_i$ is the $k \times 1$ vector of regression coefficients.

We refer to the data from $i = 1, \dots, n$ as the history period, where the regression coefficients are assumed to be constant, i.e. $\beta_i \equiv \beta_0$, $i = 1, \dots, n$, and we want to monitor new data from time $n+1$ onwards to test whether any structural change occurs in this monitoring period. Thus, tests for monitoring are concerned with the hypothesis that

$$\beta_i \quad = \quad \beta_0 \qquad (i > n) \tag{2}$$

against the alternative that at some point in the future the coefficient vector $\beta_i$ changes.

The results in this paper are valid under fairly general assumptions on regressors and disturbances; basically, they have to be such that a functional central limit theorem holds. One possible set of assumptions is given by Krämer et al. (1988), who make one assumption about the disturbances and one about the regressors:

(A1) $\{u_i\}$ is a homoskedastic martingale difference sequence with respect to $\mathcal{A}_i$, the $\sigma$-field generated by $\{y_s, x_s, u_s | s < i\}$, with $\mathrm{E}[u_i^2|\mathcal{A}_i] = \sigma^2$.

(A2) $\{x_i\}$ is such that $\limsup_{n\to\infty} \frac{1}{n}\sum_{i=1}^n ||x_i||^{2+\delta} < \infty$ for some $\delta > 0$ and $||\cdot||$ the Euclidean norm; and furthermore that

$$\frac{1}{n}\sum_{i=1}^n x_i x_i^\top \xrightarrow{\text{p}} Q$$

for some finite regular nonstochastic matrix $Q$.

Assumption (A2) allows for dynamic models, provided the regressors are (almost) stationary. These assumptions can be modified to some extent without affecting the asymptotics, see e.g. Bai (1997) for a discussion of some variations.

In what follows, $\hat\beta^{(i,j)}$ is the ordinary least squares (OLS) estimate of the regression coefficients based on the observations $i+1, \dots, i+j$, similarly matrices $Q_{(i,j)}$ indexed with $(i,j)$ are composed using the observations from the same data window. Analogously, $\hat\beta^{(i)} \equiv \hat\beta^{(1,i)}$ denotes the OLS estimate based on all observations from 1 through $i$, and $Q_{(i)}$ is shorthand for $Q_{(1,i)}$. The OLS residuals are denoted as $\hat u_i = y_i - x_i^\top \hat\beta^{(n)}$ and $\hat\sigma^2$ is some suitable estimator of the disturbance variance, e.g., $\hat\sigma^2 = \frac{1}{n-k}\sum_{i=1}^n \hat u_i^2$.

The general idea behind all of the procedures we consider is to derive a process that captures the fluctuation either in estimates or in residuals of a regression model and to reject the null hypothesis of stability whenever there is excessive fluctuation in these processes, as assessed against asymptotic boundaries that the limiting processes are known to cross with a given probability. The following section presents a class of such fluctuation-type tests and extends them in several directions.

## 3. THE GENERALIZED FLUCTUATION TEST FOR MONITORING

### 3.1 Estimates-based processes

Chu et al. (1996) were the first to extend a fluctuation test, namely the RE (recursive estimates) test, to the monitoring case. They suggested to employ the recursive estimates process

$$Y_n(t) = \frac{i}{\hat{\sigma}\sqrt{n}} \cdot Q_{(n)}^{\frac{1}{2}} \left(\hat{\beta}^{(i)} - \hat{\beta}^{(n)}\right), \tag{3}$$

where $Q_{(n)} = X_{(n)}{}^\top X_{(n)}/n$ and $i = \lfloor k + t(n-k) \rfloor$ and $t \geq 0$ is standardized time relative to the history sample (i.e., $t = 1$ corresponds to $i = n$), and to reject the null hypothesis whenever (one component of) the process $Y_n(t)$ crosses the boundary $\pm b_1(t)$ where

$$b_1(t) \quad = \quad \sqrt{t(t-1)\left[\lambda^2 + \log\left(\frac{t}{t-1}\right)\right]} \tag{4}$$

in the monitoring period $1 < t < T$ and $\lambda$ determines the significance level of this procedure, or equivalently when $\max_i |Y_{in}(t)|$, $i = 1, \dots, k$, crosses $b_1(t)$. Both $1 < t < T$ and $n < i < Tn$ are referred to as the monitoring period as they correspond to the same observations.

Leisch et al. (2000) introduced the generalized fluctuation test for monitoring, which contains the test of Chu et al. (1996) as a special case. Specifically, they considered processes that reflect the fluctuation within estimates of the regression coefficients to detect structural changes. Another special case of this class of tests is the ME (moving estimates) test which uses estimates from a moving data window of fixed width, i.e.

$$Z_n(t|\,h) = \frac{\lfloor nh \rfloor}{\hat{\sigma}\sqrt{n}} \cdot Q_{(n)}^{\frac{1}{2}} \left(\hat{\beta}^{(\lfloor nt \rfloor - \lfloor nh \rfloor, \lfloor nh \rfloor)} - \hat{\beta}^{(n)}\right), \qquad (t \geq h) \tag{5}$$

and rejects the null hypothesis if (one component of) the process crosses the boundary $\pm c(t)$, where

$$c(t) \quad = \quad \lambda \cdot \sqrt{\log_+ t}, \tag{6}$$

in the monitoring period $1 < t < T$; here $\log_+ t$ is 1 for $t \leq e$ and $\log t$ otherwise. In theory the end of the monitoring period $T$ may be infinity but in many applications using a finite $T$ is more natural because the monitoring period, or at least a reasonable upper bound for it, is known in advance. In that way no size is lost for an infinite monitoring period on $[T, \infty)$. In addition, Leisch et al. (2000) consider tests based on the same processes but capturing the fluctuation with the range instead of the maximum of the deviation of the estimates.

### 3.2 Residual-based processes

As for tests for structural change in the history period, fluctuation tests for monitoring can not only be based on the differences of estimates of the regression coefficients but also on residuals; this was already considered by Chu et al. (1996) although they focused on the recursive estimates approach. Whereas they used a CUSUM procedure based on recursive residuals we will introduce monitoring processes based on the computationally more convenient OLS residuals. The OLS residual- and estimates-based types of tests are equivalent in the case where there is only a constant regressor, a common situation in statistical quality control. The idea is as intuitive as for the estimates-based processes: the regression coefficients are just estimated once for the history period and based on these estimates the residuals of the observations in the monitoring period are computed. If there is a structural change in the monitoring period the residuals are expected to deviate systematically from their zero mean. Thus, we introduce monitoring processes based on the OLS residuals

$$\hat{u}_i^{(n)} \quad = \quad y_i - x_i^\top \hat{\beta}^{(n)}. \tag{7}$$

The OLS-based CUSUM process for monitoring is then defined as:

$$B_n^0(t) \quad = \quad \frac{1}{\hat{\sigma}\sqrt{n}} \sum_{i=1}^{\lfloor nt \rfloor} \hat{u}_i^{(n)} \qquad (t \geq 0). \tag{8}$$

The following functional central limit (FCLT) holds for $B_n^0(t)$:

$$B_n^0(t) \quad \Rightarrow \quad W^0(t) \ = \ W(t) - t \cdot W(1), \tag{9}$$

where $W$ and $W^0$ are the (1-dimensional) Brownian motion and Brownian bridge, respectively. The proof for (9) is essentially the same as in Ploberger and Krämer (1992) for the ordinary OLS-based CUSUM test except that $t$ is from the compact interval $[0, T]$, with $T > 1$, rather than from $[0, 1]$: Rewrite (8) as

$$\hat{\sigma} B_n^0(t) \quad = \quad \frac{1}{\sqrt{n}} \sum_{i=1}^{\lfloor nt \rfloor} u_i - \frac{1}{\sqrt{n}} \sum_{i=1}^{\lfloor nt \rfloor} x_i^\top \left( \hat{\beta}^{(n)} - \beta \right). \tag{10}$$

As in Ploberger and Krämer (1992) the following relation holds uniformly in $t$ on $[0, T]$:

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{\lfloor nt \rfloor} x_i^\top \left( \hat{\beta}^{(n)} - \beta \right) \quad = \quad \frac{t}{\sqrt{n}} \sum_{i=1}^{n} u_i + o_{\mathrm{p}}(1). \tag{11}$$

Hence (9) follows from the well-known fact that

$$\frac{1}{\sqrt{n}} \left( \sum_{i=1}^{\lfloor nt \rfloor} u_i - t \sum_{i=1}^{n} u_i \right) \quad \Rightarrow \quad \sigma \left( W(t) - tW(1) \right) \ = \ \sigma W^0(t). \tag{12}$$

The OLS-based MOSUM process for monitoring is defined analogously as:

$$M_n^0(t|h) \quad = \quad \frac{1}{\hat{\sigma}\sqrt{n}} \left( \sum_{i=\lfloor \eta t \rfloor - \lfloor nh \rfloor + 1}^{\lfloor \eta t \rfloor} \hat{u}_i \right) \qquad (t \geq h) \tag{13}$$

$$\quad = \quad B_n^0\left( \frac{\lfloor \eta t \rfloor}{n} \right) - B_n^0\left( \frac{\lfloor \eta t \rfloor - \lfloor nh \rfloor}{n} \right), \tag{14}$$

where $\eta = (n - \lfloor nh \rfloor)/(1-h)$. From (14) together with (9) it follows directly that $M_n^0(t|h)$ satisfies the following FCLT:

$$M_n^0(t|h) \quad \Rightarrow \quad W^0(t) - W^0(t-h), \tag{15}$$

i.e., the OLS-based MOSUM process converges towards the process of the increments of the Brownian bridge. Therefore the limiting process for the OLS-based CUSUM and MOSUM process is the 1-dimensional special case of the $k$-dimensional recursive and moving estimates process. The respective empirical processes are in fact equivalent if $x_i = 1$ for all $i$. Thus, the boundaries given in the previous section can be used as well for the OLS-based processes.

The advantage of estimates-based processes is that there is a process for each regression coefficient, hence it can be determined which coefficient(s) is (are) responsible for the rejection of the null hypothesis. The OLS-based processes on the other hand are much easier to compute because a linear model has to be fit only once for the whole process (and not in every single step) and then just residuals have to be computed.

 

### 3.3 Rescaling of estimates-based processes

The estimates-based processes from (3) and (5) scale the estimates of the regression coefficients with the estimate $Q_{(n)}$ of their asymptotic covariance matrix $Q$ that is based on the observations in the history period. Kuan and Chen (1994) showed by simulation of empirical sizes that the tests can be seriously distorted in dynamic models and suggested to rescale the processes in order to repair this defect. Instead of estimating $Q$ always on the basis of the full history period, each estimate $\hat{\beta}^{(i,j)}$ is scaled with the corresponding estimate of the covariance matrix $Q_{(i,j)}$, i.e., the modified processes can be written as

$$Y_n^*(t) = \frac{i}{\hat{\sigma}\sqrt{n}} \cdot Q_{(i)}{}^{\frac{1}{2}} \left( \hat{\beta}^{(i)} - \hat{\beta}^{(n)} \right), \tag{16}$$

$$Z_n^*(t|\, h) = \frac{\lfloor nh \rfloor}{\hat{\sigma}\sqrt{n}} \cdot Q_{(\lfloor nt \rfloor - \lfloor nh \rfloor, \lfloor nh \rfloor)}{}^{\frac{1}{2}} \left( \hat{\beta}^{(\lfloor nt \rfloor - \lfloor nh \rfloor, \lfloor nh \rfloor)} - \hat{\beta}^{(n)} \right). \tag{17}$$

The respective limiting processes remain of course the same, because both estimates of the covariance matrices also converge to $Q$ as $n \to \infty$, but Kuan and Chen (1994) show that they converge faster in dynamic models. Using the same idea it is quite intuitive that the rescaling of the estimates-based processes might also provide benefits for monitoring in dynamic models.

Following Kuan and Chen (1994) we consider three data generating processes (DGPs):

$$y_i = \varrho \cdot y_{i-1} + u_i, \qquad y_0 = 0, \tag{18}$$

$$y_i = 2 + \varrho \cdot y_{i-1} + u_i, \qquad y_0 = 0, \tag{19}$$

$$y_i = 2 + \varrho \cdot x_i + u_i, \qquad x_i = \varrho \cdot x_{i-1} + \varepsilon_i, \tag{20}$$

with $u_i$ and $\varepsilon_i$ n.i.d.(0,1), and simulate the size of the corresponding tests for a range of sample sizes and of $\varrho$ with $\alpha = 0.1$ and $h = 0.5$. We use four different values for the sample size ($n = 10, 25, 50, 100$) and two different values for the monitoring period ($T = 2, 10$) and compute the empirical size based on 1000 replications. Our simulations show that the problem is the same in the monitoring case, especially for short history and long monitoring periods: For large values of $\varrho$ the empirical size of the ME test is seriously distorted, see Table I.

Kuan and Chen (1994) illustrate this phenomenon with the following equation for the DGP (18):

$$\mathrm{E}\left( \frac{Q_{(n)}}{Q_{(\lfloor nt \rfloor)}} \right) = 1 - \frac{2(1+\varrho^2)(n - \lfloor nt \rfloor)}{n \lfloor nt \rfloor (1 - \varrho^2)} + O(n^{-2}). \tag{21}$$

The second term on the right hand side is a bias term tending to 0 for $n \to \infty$ for fixed $\varrho$, but for a fixed sample size $n$ it approaches infinity for $\varrho \to \pm 1$. This bias is reduced for $t < 1$ and as the term is monotone in $t$ it is even enhanced if $t > 1$, so that rescaling will even increase the distortion of the empirical size of this test. Therefore rescaling makes no sense in the case of the recursive estimates test for monitoring, but it does for the moving estimates test, because the parameter that determines the window size is not $t$ but $h$, and $h \leq 1$. This is confirmed by our simulations: Table II shows that the bias is much smaller for the rescaled processes, especially when the history size $n$ is reasonably large.

Thus, in practical applications of moving estimates processes the rescaled version should always be used.

### 3.4 Boundaries

The shape of the boundaries for empirical fluctuation processes does not make a big difference under the null hypothesis, because they are always chosen to be crossed with the (asymptotic) probability $\alpha$. However, under the alternative they can affect very much the chance to detect certain patterns of structural changes. For example, the CUSUM tests (in historical samples) perform poorly if a change occurs late in the sample period. Zeileis (2004) suggests alternative

 

Table I: Empirical size of the ME test at the 10% level

| DGP | $T$ | $n$ | Autocorrelation coefficient $\varrho$ | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | 0.0 | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 |
| (18) | 2 | 10 | 12.7 | 14.2 | 17.1 | 18.6 | 21.5 | 22.0 | 26.5 | 32.6 | 39.7 | 51.4 |
| | | 25 | 10.6 | 8.3 | 12.6 | 13.2 | 15.1 | 17.0 | 21.3 | 26.1 | 34.1 | 48.5 |
| | | 50 | 14.1 | 12.6 | 12.2 | 11.5 | 14.6 | 14.1 | 16.7 | 20.1 | 24.2 | 36.4 |
| | | 100 | 10.5 | 8.8 | 11.2 | 10.3 | 12.3 | 12.8 | 13.3 | 17.1 | 19.9 | 28.3 |
| | 10 | 10 | 44.8 | 54.1 | 62.6 | 68.9 | 74.0 | 85.5 | 90.7 | 94.2 | 98.6 | 99.9 |
| | | 25 | 21.0 | 23.7 | 30.2 | 39.2 | 49.9 | 63.4 | 75.6 | 89.5 | 98.8 | 100.0 |
| | | 50 | 15.8 | 17.4 | 21.4 | 29.3 | 35.0 | 46.1 | 61.7 | 77.7 | 93.9 | 99.7 |
| | | 100 | 10.7 | 12.2 | 11.9 | 14.4 | 14.4 | 18.8 | 22.2 | 26.8 | 39.2 | 67.1 |
| (19) | 2 | 10 | 16.5 | 24.1 | 24.4 | 32.1 | 39.2 | 49.9 | 62.5 | 81.6 | 93.5 | 93.8 |
| | | 25 | 13.7 | 13.0 | 15.6 | 17.1 | 24.9 | 30.5 | 42.8 | 56.2 | 78.7 | 98.8 |
| | | 50 | 18.3 | 10.5 | 11.1 | 14.7 | 16.8 | 18.3 | 22.7 | 36.8 | 62.9 | 93.7 |
| | | 100 | 10.1 | 11.4 | 11.5 | 13.6 | 14.0 | 12.9 | 17.1 | 21.9 | 40.0 | 78.6 |
| | 10 | 10 | 61.5 | 69.4 | 81.2 | 85.4 | 92.1 | 97.0 | 99.7 | 100.0 | 100.0 | 100.0 |
| | | 25 | 23.5 | 33.6 | 42.0 | 50.2 | 66.0 | 79.3 | 93.3 | 99.2 | 99.9 | 100.0 |
| | | 50 | 18.6 | 24.5 | 30.5 | 36.0 | 48.6 | 62.5 | 79.1 | 90.7 | 99.8 | 100.0 |
| | | 100 | 11.3 | 11.7 | 14.0 | 14.4 | 19.0 | 21.3 | 30.8 | 38.1 | 65.1 | 99.2 |
| (20) | 2 | 10 | 18.5 | 17.1 | 18.2 | 20.6 | 23.1 | 22.4 | 25.8 | 28.0 | 28.7 | 37.3 |
| | | 25 | 12.1 | 12.1 | 12.0 | 13.5 | 11.5 | 15.7 | 15.3 | 17.9 | 23.1 | 26.4 |
| | | 50 | 13.2 | 11.9 | 9.2 | 12.3 | 10.1 | 13.2 | 11.7 | 15.0 | 14.8 | 22.5 |
| | | 100 | 11.2 | 9.6 | 10.1 | 10.0 | 8.4 | 10.9 | 11.0 | 11.2 | 12.4 | 17.2 |
| | 10 | 10 | 61.3 | 67.8 | 80.0 | 84.3 | 92.6 | 98.0 | 99.9 | 100.0 | 100.0 | 100.0 |
| | | 25 | 24.8 | 33.8 | 44.6 | 55.8 | 66.6 | 81.7 | 93.0 | 99.8 | 100.0 | 100.0 |
| | | 50 | 20.2 | 22.9 | 29.0 | 37.5 | 49.8 | 63.3 | 78.3 | 92.7 | 99.7 | 100.0 |
| | | 100 | 10.8 | 10.3 | 11.9 | 10.6 | 11.0 | 12.0 | 13.0 | 13.8 | 17.3 | 30.1 |

 

Table II: Empirical size of the rescaled ME test at the 10% level

| DGP | $T$ | $n$ | Autocorrelation coefficient $\varrho$ | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | 0.0 | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 |
| (18) | 2 | 10 | 12.0 | 15.3 | 16.2 | 19.3 | 18.4 | 23.8 | 24.1 | 27.9 | 32.5 | 39.4 |
| | | 25 | 9.9 | 12.0 | 11.7 | 14.0 | 15.7 | 12.9 | 17.1 | 21.2 | 27.3 | 35.3 |
| | | 50 | 11.9 | 8.7 | 10.1 | 10.5 | 12.3 | 11.9 | 12.4 | 16.0 | 20.3 | 28.4 |
| | | 100 | 7.7 | 10.1 | 10.3 | 10.3 | 9.6 | 10.4 | 10.8 | 11.6 | 15.7 | 22.7 |
| | 10 | 10 | 20.0 | 18.2 | 23.8 | 23.5 | 27.0 | 29.9 | 37.2 | 41.1 | 51.0 | 60.7 |
| | | 25 | 13.6 | 12.5 | 13.0 | 15.1 | 15.7 | 17.4 | 22.4 | 26.9 | 33.6 | 50.5 |
| | | 50 | 10.2 | 11.2 | 13.3 | 13.1 | 13.4 | 17.5 | 18.3 | 19.8 | 31.1 | 40.7 |
| | | 100 | 10.9 | 12.4 | 10.5 | 9.9 | 11.4 | 13.9 | 13.9 | 16.5 | 19.7 | 31.1 |
| (19) | 2 | 10 | 13.2 | 12.0 | 15.1 | 13.9 | 15.8 | 15.8 | 18.9 | 18.6 | 22.8 | 33.5 |
| | | 25 | 9.3 | 9.2 | 10.2 | 11.3 | 12.3 | 11.6 | 10.5 | 12.4 | 15.2 | 21.0 |
| | | 50 | 11.8 | 8.4 | 8.0 | 10.3 | 12.3 | 11.8 | 13.1 | 13.1 | 10.9 | 15.4 |
| | | 100 | 9.6 | 9.8 | 9.2 | 8.5 | 9.9 | 9.4 | 10.7 | 10.4 | 10.5 | 13.6 |
| | 10 | 10 | 17.4 | 16.3 | 19.1 | 22.0 | 20.7 | 24.0 | 27.9 | 27.7 | 31.5 | 43.4 |
| | | 25 | 11.0 | 10.6 | 11.2 | 10.7 | 14.4 | 15.0 | 17.3 | 17.8 | 18.1 | 23.4 |
| | | 50 | 10.0 | 10.4 | 10.6 | 11.0 | 11.6 | 13.8 | 14.5 | 17.4 | 18.1 | 17.9 |
| | | 100 | 10.8 | 10.0 | 9.2 | 10.9 | 11.4 | 10.8 | 12.3 | 12.5 | 13.3 | 11.4 |
| (20) | 2 | 10 | 15.2 | 17.0 | 17.0 | 14.6 | 15.6 | 18.2 | 18.5 | 21.1 | 21.8 | 25.4 |
| | | 25 | 9.4 | 10.4 | 10.6 | 10.3 | 10.3 | 11.7 | 11.5 | 13.2 | 13.3 | 16.8 |
| | | 50 | 10.5 | 9.8 | 11.0 | 7.2 | 10.6 | 10.5 | 11.1 | 11.7 | 12.8 | 16.0 |
| | | 100 | 8.4 | 11.6 | 9.4 | 10.5 | 11.2 | 9.2 | 8.0 | 8.3 | 11.3 | 13.2 |
| | 10 | 10 | 16.8 | 16.2 | 19.4 | 20.3 | 22.1 | 22.8 | 25.1 | 27.1 | 32.7 | 45.5 |
| | | 25 | 8.4 | 9.2 | 9.9 | 13.7 | 12.9 | 16.1 | 16.9 | 21.1 | 19.7 | 21.3 |
| | | 50 | 9.8 | 10.4 | 10.7 | 10.8 | 11.8 | 14.5 | 17.6 | 13.8 | 17.4 | 18.9 |
| | | 100 | 10.1 | 9.7 | 9.5 | 11.4 | 10.1 | 11.1 | 7.6 | 11.4 | 12.1 | 14.5 |

 

boundaries which are able to increase the detection chances of the OLS-based CUSUM test for early and late changes. Also in the case of monitoring the detection properties for structural changes in the monitoring period strongly depend on the shape of the boundaries—a topic which has not yet been studied in detail. Since structural change tests suffer from a huge alternative, as mentioned in the introduction, alternative boundaries provide a means to derive tests that are suitable for various types of deviations from stability.

Chu et al. (1996) already state that the RE test for monitoring has good chances to detect changes early in the monitoring period, but gets increasingly insensitive to late structural changes. This is due to the fact that most of the size of the test is used at the very beginning of the monitoring period as Figure 1 shows. It depicts the distribution of hitting times (from 10,000 runs) for the Brownian bridge, the asymptotic approximation to the RE process, with the standard boundary (4) using $T = 10$ at level $\alpha = 0.1$ and for the increments of a Brownian bridge (with $h = 0.5$), the asymptotic approximation to the ME process. It is clear that the size of the ME test is spread much more evenly; in fact 25 % of the size of the RE test is used on the interval $[1, 1.09]$. This is caused by the shape of the boundary $b_1(t)$, which can be seen in Figure 3: it starts together with the Brownian bridge in 0 at $t = 1$, hence most random crossings will occur very early. We will introduce boundaries for the RE process that distribute the size more evenly. By controlling size rather than power we do not have to specify one particular alternative but can expect similar results for many types of alternatives as the empirical processes typically start to fluctuate and deviate from their zero mean at the time of the structural change.

In order to obtain a boundary that does not use up the size of the corresponding test at the beginning of the monitoring period it seems natural to choose a boundary with an offset in $t = 1$, but with the correct asymptotic growth rate $t$. The simplest boundary that fulfills these requirements is

$$b_2(t) \quad = \quad \lambda \cdot t. \tag{22}$$

One might want to consider a boundary which is constant at the beginning of the monitoring period like the boundary (6), but this is inappropriate for a process with growing variance such as the Brownian bridge, because simulations show that most of the size will then be used at the point where the boundary changes from being constant to growing. Because there is no (known) closed-form result for the crossing probability of a Brownian bridge for the boundary (22), we simulate the appropriate critical values of $\lambda$ for different values of $T$ as for the ME test. Table III

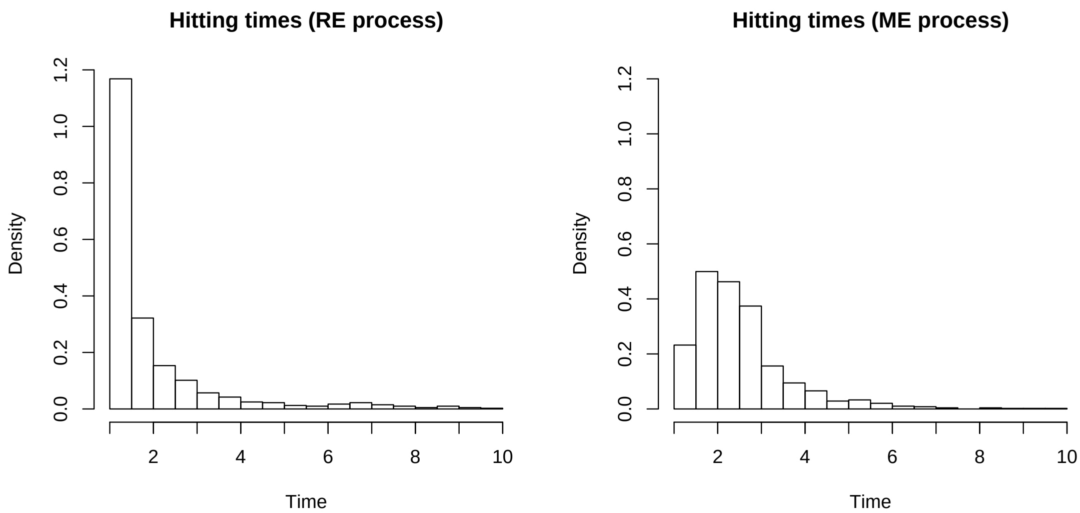


Figure 1: Comparison of hitting times for RE and ME process with standard boundaries

 

Table III: Critical values for the boundary $b_2$

| $\alpha$ (in %) | $T$ = 2 | 3 | 4 | 5 | 6 | 8 | 10 |
|---|---|---|---|---|---|---|---|
| 20.0 | 1.159 | 1.329 | 1.430 | 1.472 | 1.502 | 1.541 | 1.567 |
| 15.0 | 1.253 | 1.445 | 1.544 | 1.589 | 1.619 | 1.668 | 1.688 |
| 10.0 | 1.383 | 1.590 | 1.695 | 1.753 | 1.789 | 1.838 | 1.860 |
| 7.5 | 1.467 | 1.688 | 1.793 | 1.861 | 1.899 | 1.961 | 1.964 |
| 5.0 | 1.568 | 1.814 | 1.939 | 2.006 | 2.046 | 2.090 | 2.128 |
| 4.0 | 1.616 | 1.896 | 2.022 | 2.076 | 2.131 | 2.159 | 2.219 |
| 3.0 | 1.680 | 1.997 | 2.103 | 2.177 | 2.226 | 2.257 | 2.311 |
| 2.0 | 1.801 | 2.114 | 2.217 | 2.301 | 2.397 | 2.380 | 2.454 |
| 1.0 | 1.976 | 2.300 | 2.423 | 2.525 | 2.573 | 2.597 | 2.650 |
| 0.5 | 2.118 | 2.478 | 2.599 | 2.712 | 2.812 | 2.766 | 2.888 |
| 0.1 | 2.435 | 2.789 | 2.973 | 3.288 | 3.226 | 3.230 | 3.401 |

gives the $1 - \alpha$ quantiles for various significance levels $\alpha$ from a simulation of 10,000 Brownian bridges.

The size of the corresponding test is also distributed more evenly (see Figure 2). Figure 3 shows the resulting alternative boundary $b_2(t)$ (at level 0.1 for $T = 2$) in comparison to the standard boundary. It can be seen that the boundaries cross at about $t = 1.3$ which means that the detection chances decrease only for very early changes, but increase for all other changes. This means that if the history period comprises observations from one year and the next year is monitored, the chances to detect a change are decreased only for the first three months (given that a change is immediately discovered with the next observation which is of course not necessarily the case). This is emphasized by simulations under a single shift alternative with a setup like in Chu et al. (1996) and Leisch et al. (2000): the data generating process is n.i.d.(2,1) with a history size of $n = 100$ and a monitoring period $T = 10$. Under the alternative the mean switches from 2 to 2.8 in the monitoring period either at $t = 1, 1.1, 2, 3$. Table IV reports the mean (and standard deviation) of the detection delay from 100,000 replications for the RE test with both boundaries and the

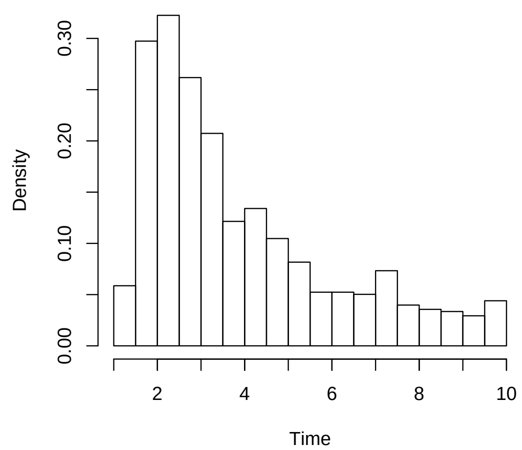


Figure 2: Hitting times for RE process with alternative boundaries

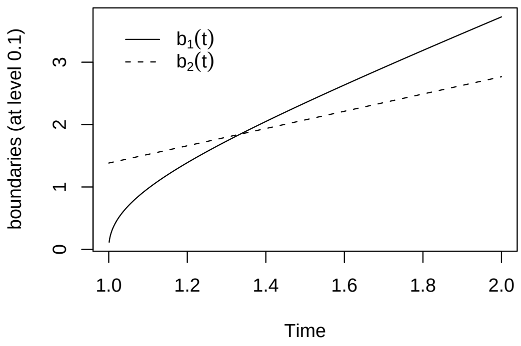


Figure 3: Boundaries for the Brownian bridge

ME test at the 5% level together with the associated type I and II errors. It can be seen that in terms of detection delay the RE test with standard boundaries performs best for early changes: structural changes occurring immediately at the end of the history period or a little bit later at observation 110 are detected on average at observation 120 and 138 respectively. But for changes later in the monitoring period the performance deteriorates dramatically. The ME test on the other hand has an almost constant detection delay of about 50 observations. The RE test with the new boundaries offers a compromise between the two approaches as it performs better than the ME test for early changes and better than the RE test with standard boundaries for late changes. However, it can be outperformed by both tests for either early or late changes. Furthermore, in terms of the standard deviation of the detection delay the results are improved compared to the standard boundaries, and both type I and type II errors are reduced. Note that even for a change as early as $t = 1.1$ the RE test with standard boundaries detects some false changes, i.e., alerts before the actual structural change, and that it fails to pick up certain late changes at all.

It is desirable, of course, to have a more flexible and less heuristic instrument to select the boundaries for fluctuation tests: one might want to choose the boundaries according to a specified prior distribution for the timing of the shift under the alternative. This issue is currently under investigation.

Table IV: Mean (and standard deviation) of detection delay with type I and II error (in %)

| | RE (with $b_1$) | | | RE (with $b_2$) | | | ME | | |
|---|---|---|---|---|---|---|---|---|---|
| shift date | delay(sd) | type I | type II | delay(sd) | type I | type II | delay(sd) | type I | type II |
| 100 | 20(15) | 0 | 0 | 39(15) | 0 | 0 | 51(13) | 0 | 0 |
| 110 | 28(18) | 0.66 | 0 | 41(17) | 0 | 0 | 50(14) | 0 | 0 |
| 200 | 88(51) | 2.61 | 0.01 | 78(41) | 0.55 | 0 | 48(18) | 1.00 | 0 |
| 300 | 149(82) | 3.10 | 0.04 | 117(65) | 1.92 | 0 | 54(20) | 3.28 | 0 |

 

# 4. APPLICATIONS

We demonstrate the usefulness of the methods introduced in the previous section in applications to three data sets: German M1 money demand (Lütkepohl et al., 1999), U.S. labor productivity (Hansen, 2001), and S&P 500 stock returns. These examples illustrate three slightly different approaches to monitoring:

1. The German M1 money demand application illustrates how monitoring can be used for the evaluation of the impact of known events (policy interventions) when it is not known if and when such an intervention becomes effective. An error correction model (ECM) modelling the relatively complex relationship of several variables is used.

2. The U.S. labor productivity application illustrates monitoring as a means of diagnostic checking of a model that is updated with incoming observations. A simple AR(1) model for a univariate time series is used.

3. The S&P 500 stock returns application illustrates the usage of monitoring as an exploratory tool for automatic screening of a large number of time series. Several hundred return series are monitored each for a change in the mean.

The analysis of the three data sets is carried out using the package `strucchange` (Zeileis et al., 2002) which implements all suggested procedures in the R system for statistical computing. To facilitate replication of our results, we use only data sets that are freely available from the World Wide Web and are now also included in the `strucchange` package. The R system and the `strucchange` package are both freely available at no cost under the terms of the GNU General Public Licence (GPL) from the Comprehensive R Archive Network at http://CRAN.R-project.org/

## 4.1 German M1 money demand

Lütkepohl et al. (1999) investigated the stability and linearity of a German M1 money demand function based on data from the German central bank using seasonally unadjusted quarterly time series from 1961(1) to 1995(4). The data are available on the World Wide Web in the data archive of the *Journal of Applied Econometrics* (http://qed.econ.queensu.ca/jae/1999-v14.5/lutkepohl-terasvirta-wolters/). Lütkepohl et al. (1999) found a stable relationship for the M1 money demand for the time before the German monetary unification on 1990-06-01 but a clear structural instability for the extended sample period up to 1995(4), which they modelled by smooth transition regression techniques. Specifically, they established a stable and linear regression relationship for the German M1 money demand using an ECM based on data for the logarithm of real M1 per capita $m_i$, the logarithm of a price index $p_i$, the logarithm of the real per capita gross national product $y_i$ and the long-run interest rate $R_i$.

OLS estimation of their model yields the following equation for the phase from 1961(1) to 1990(2) before the German monetary unification:

$$\begin{aligned} \Delta m_i &= -0.30\Delta y_{i-2} - 0.67\Delta R_i - 1.00\Delta R_{i-1} - 0.53\Delta p_i \\ &\quad -0.12 m_{i-1} + 0.13 y_{i-1} - 0.62 R_{i-1} \\ &\quad -0.05 - 0.13Q1 - 0.016Q2 - 0.11Q3 + \hat{u}_i, \end{aligned} \tag{23}$$

where $Q1$–$Q3$ are seasonal dummies and all coefficients (except for the intercept) are highly significant; the fitted model gives an adjusted $R^2 = 0.943$. In a cointegration relationship, the estimators of coefficients on I(1) variables converge at a faster rate than those of the coefficients on I(0) variables, hence they may be treated as known when monitoring the error correction model. We therefore aggregate the cointegrated variables $m_{i-1}$, $y_{i-1}$ and $R_{i-1}$ to a single variable $e_{i-1} = -0.12 m_{i-1} + 0.13 y_{i-1} - 0.62 R_{i-1}$ to assure stationarity of this regressor in the ECM. This way of testing for structural change in ECMs is similar to the procedure suggested by Hansen (1992).

 

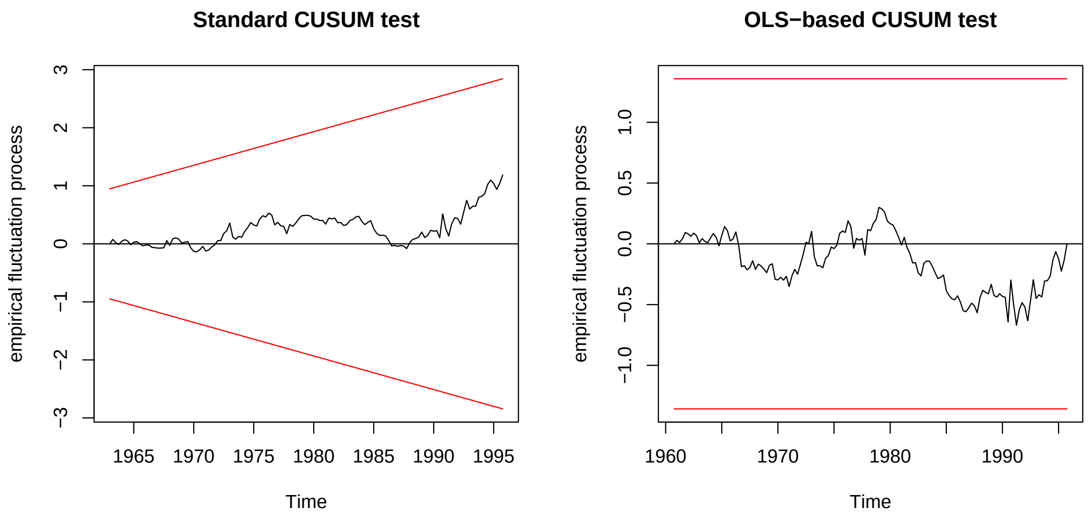


Figure 4: Historical residual-based fluctuation tests

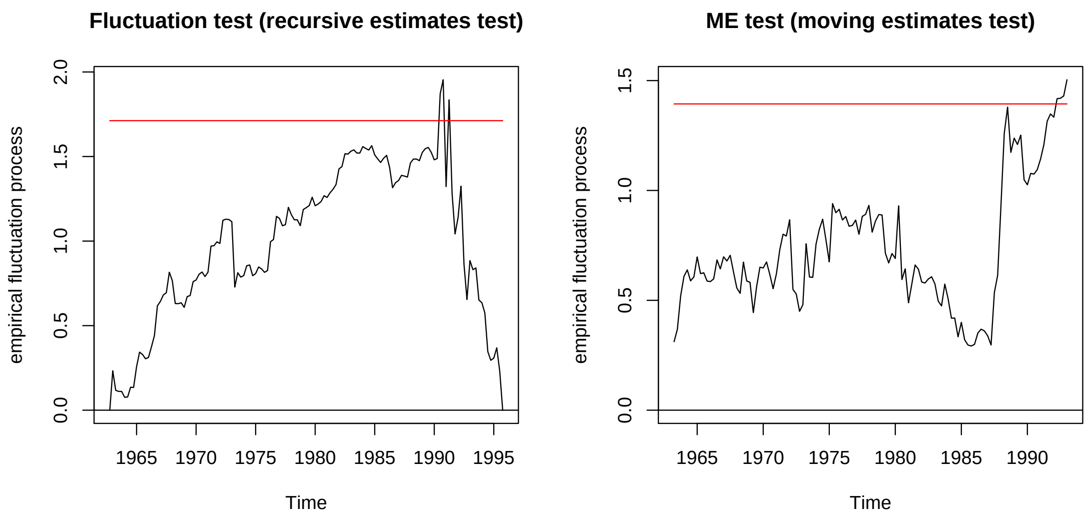


Figure 5: Historical estimates-based fluctuation tests

In the following, we first show that the structural change modelled by Lütkepohl et al. (1999) can be detected using fluctuation tests in this ECM, i.e., we investigate the behaviour under the alternative. Second, we present some Monte Carlo evidence confirming that estimation of the cointegration relationship does not interfere with subsequent monitoring, i.e., we investigate the behaviour of the tests in an ECM under the null hypothesis. In order to detect a change, two strategies can be employed: one can use either historical tests trying to find a structural change *ex post* or the monitoring methods introduced in the previous section trying to detect the structural change *online*. Although we focus on the latter approach in this paper we will first carry out historical tests as well.

**Historical fluctuation tests:** First, two residual-based fluctuation tests—the recursive (or standard) and the OLS-based CUSUM test—are applied to the model (23). Figure 4 shows that both CUSUM processes lack any significant fluctuation and that neither process crosses its 5% level boundary; thus, both tests fail to detect structural change in the data. However, two estimates-based tests—the RE and the ME test with $h = 0.15$—both show a clear peak or shift respectively in the beginning of the 1990s (cf. Figure 5), i.e., both tests detect the structural change after the monetary unification that is also described by Lütkepohl et al. (1999). The reason that the residual-based tests are insensitive to this change while the estimates-based tests are not is the well-known fact that the power of both CUSUM tests depends on the angle between the shift and the mean regressor; in particular they do not have power against shifts orthogonal to the mean regressor (Krämer et al., 1988, Ploberger and Krämer, 1992). Assuming that there is just one structural shift, which appears to be a reasonable hypothesis for the present data, the shift $\Delta\beta$ is estimated by $\hat{\beta}^{(1,118)} - \hat{\beta}^{(119,140)}$, the difference of the estimated coefficients in the history and the monitoring period. The mean regressor $\bar{x}_\infty = \lim_{n\to\infty} 1/n \sum_{i=1}^{n} x_i$ is estimated by the empirical mean. This leads to an estimate for the angle of 90.106°, confirming the suspicion that an orthogonal shift is the cause of the lack of power for the residual-based tests.

**Monitoring:** Now we will confirm the instability of the coefficients in the regression relationship for the money demand function using the tools introduced in Section 3: the OLS-based CUSUM process with the alternative boundary $b_2(t)$ from (22) and the rescaled moving estimates process. We consider the observations from 1961(1)–1990(2) as the history period of the monitoring process and the observations after the monetary unification from 1990(3)–1995(4) as the monitoring period. Thus, we put ourselves in the position of a researcher in 1990 who wants to find out whether the model established for the pre-1990 money demand becomes unstable following the monetary unification. For this a significance level of 5% for a monitoring period of $T = 2$ is used.

Figure 6 shows the OLS-CUSUM process with the history period left of the vertical dashed line and the monitoring period on the right. Whereas the process does not exhibit much fluctuation before 1990(2) it does so after the start of the monitoring period: it crosses the standard boundary after ten observations in 1992(4) and the alternative boundary another seven observations later. As the break occurs immediately after the end of the history period the standard boundaries perform a bit better, but note that there is almost a crossing in 1990(4) after just two observations. In a "real" monitoring situation it would be hard to decide if such a crossing was just a type I error or caused by a structural change. The ME process (with $h = 0.5$) also has a clear shift (see Figure 7), but crosses its boundary a little bit later: in the third quarter of 1994. Hence we can find overwhelming evidence that there has been a structural change in the money demand relationship after the monetary unification.

Although the detection delay of 2.5 years (for the OLS-based CUSUM test with boundary $b_1$) might seem rather long at first sight, this might still be useful in practice. First, this needs to be contrasted with the observation frequency and the dimension of the model. From this point of view, a method detecting a break after 10 observations for a model containing 11 parameters would appear to do rather well. Second, with the monitoring approach we are able to do somewhat better than the common historical tests: For instance, using $F$ statistics (Andrews, 1993; Andrews & Ploberger, 1994) we would need a sufficient number of observations after the shift in order to sensibly estimate model (23) for this second phase—this is certainly not the case with only 10 observations.

 

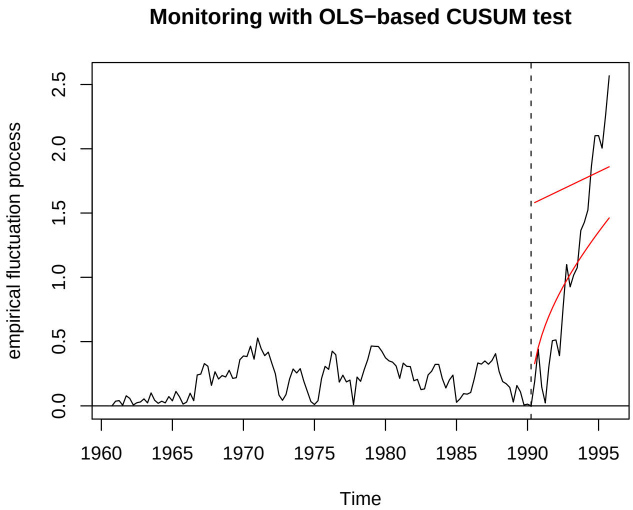


Figure 6: OLS-based CUSUM process

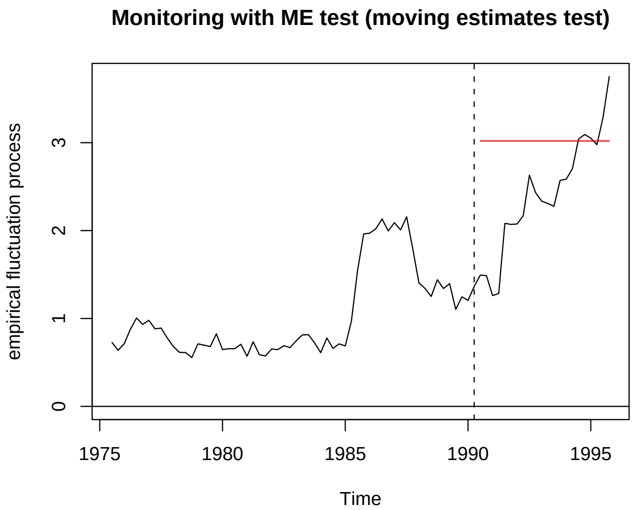


Figure 7: Moving estimates process

 

Given the data up to 1995(4) there are various ways to assure that the observed crossings are not type I errors: First, from visual inspection there seems to be a clear drift in the fluctuation processes, beginning after the end of the history period and continuing after the boundary crossing. Second, the historical estimates-based tests above also found a structural change and their processes also indicated a shift in about 1990. Third, the OLS breakpoint estimate for a shift in the regression coefficients of the ECM is 1990(3), which misses the timing of the monetary unification just by one observation.

**Size simulation:** In order to ensure that aggregation of the three I(1) variables to a single stationary variable as described at the beginning of this section does not lead to size distortions in the monitoring procedures we use a simulation study. Based on the stable regression relationship in the history period (observations $i = 1, \dots, 118$) the ECM with coefficients $\hat{\beta}$ from (23) and the standard deviation $\hat{\sigma} = 0.0127$ (also obtained from the history period) is estimated. To assure stability for the remaining observations $i = 119, \dots, 140$ the following DGP is used for each $i$:

1. $m_i = m_{i-1} + \Delta m_{i-1}$,
2. $\Delta m_i = x_i^\top \hat{\beta} + \xi_i$,

where $x_i$ is the vector of all regressors and the $\xi_i$ are n.i.d.$(0, \hat{\sigma}^2)$. Subsequently, the I(1) variables are again aggregated to a single stationary variable $e_i$ and the monitoring procedures are applied to this partly artificial data set (without a break) with the same settings as for the original data. As the simulation of the artificial data starts with $i = 119$ the estimation of the ECM differs only for the estimates-based monitoring processes whereas the OLS-based procedures estimate the ECM only once on the history period, which is identical in the original and the semi-artificial data. To analyze whether the estimation of the ECM affects the OLS-based procedures the artificial DGP could also be started at $i = 2$ using time $i = 1$ as starting values; however, this leads to almost identical results.

The empirical rejection probabilities (type I errors) based on 10,000 replications of this procedure are reported in Table V. It can be seen that only the RE test with standard boundaries exceeds the nominal 5% significance level. The reason for this is not the presence of I(1) terms in the model equation but the size distortions for the RE test in dynamic models. As described in Section 3.3, these cannot be remedied by rescaling. The reason that the RE test with the new boundaries does not suffer from the same problem is that the standard boundaries are more prone to random crossing in such a short monitoring period (as argued in Section 3.4). The low rejection probabilities for the other tests are caused by the fact that not the full monitoring period up to $T = 2$ is used, rendering the tests rather conservative. In particular, the monitoring period is too short for the MOSUM and ME type processes to produce erroneous boundary crossings. In summary, the size simulation confirms the results from the previous section and justifies the structural change analysis of ECMs as carried out above.

Table V: Size simulation for ECM on artificial M1 data (entries are percentages)

| OLS-CUSUM ($b_1$) | OLS-CUSUM ($b_2$) | RE ($b_1$) | RE ($b_2$) | OLS-MOSUM | ME |
|---|---|---|---|---|---|
| 0.61 | 0.01 | 37.62 | 0.21 | 0.00 | 0.00 |

## 4.2 U.S. labor productivity

In his recent overview of "The new econometrics of structural change," Hansen (2001) examines U.S. labor productivity in the manufacturing/durables sector, a monthly time series with observations from 1947(2) through 2001(4) which is available from Bruce Hansen's homepage (`http://www.ssc.wisc.edu/~bhansen/`). He uses a first order autoregressive model for the U.S. labor productivity in the manufacturing/durables sector which is measured by $x_i$, the growth rate of the Industrial Production Index to average weekly labor hours.

 

Hansen (2001) finds a clear structural change in about 1994 and two weaker changes in 1963 and 1982. For illustration, we choose the time from 1964(1) until 1979(12) as the history period, because we are interested in monitoring the two later changes. We exclude the time before 1964 because there must not be a break in the history period. OLS estimation of the coefficients in the historical AR(1) model yields

$$x_i \quad = \quad 0.0025 - 0.186 x_{i-1} + \hat{u}_i, \tag{24}$$

with both coefficients being highly significant. As for the money demand data we monitor the data using the OLS-based CUSUM process with boundaries $b_2(t)$ from (22) and the rescaled moving estimates process with $h = 0.5$ using a 5% level and $T = 2.5$.

This approach is slightly different from the one in our first example: there is no known event that might cause an instability in the model considered. We rather assume that we are in a position where we have established a model equation we want to work with, and we want to learn whether we have to update it or not.

It could also be argued that in practice one wants to update the model anyway with every new observation instead of waiting until a change occurs. In this case, monitoring can be used as a supplementary diagnostic check which comes "for free" as in the RE and ME test the regression coefficients are updated for every new observation (either for a recursively growing or a moving data window). Here, "blind" updating of the coefficients could lead to uninterpretable results if the time period used for estimation contains a structural change, however, simultanous monitoring would discover this and the model could be adapted to this situation.

The OLS-based CUSUM process for the labor productivity data in Figure 8 provides information at several levels: First, both versions of the OLS-based CUSUM test find a significant structural change at the 5% level as the process crosses both the standard and the alternative boundary. Second, Figure 8 conveys the impression that there are two structural changes: the first in about 1983, where the path starts to depart from zero, and the second in the early 1990s. Third, neither the standard nor the alternative boundaries detect the first shift at the 5% level, but the process crosses both boundaries after the second break: the new boundary already in 1998(8) and the standard one in 2000(5).

Figure 9 provides rather similar results for the ME test. However, the moving estimates process

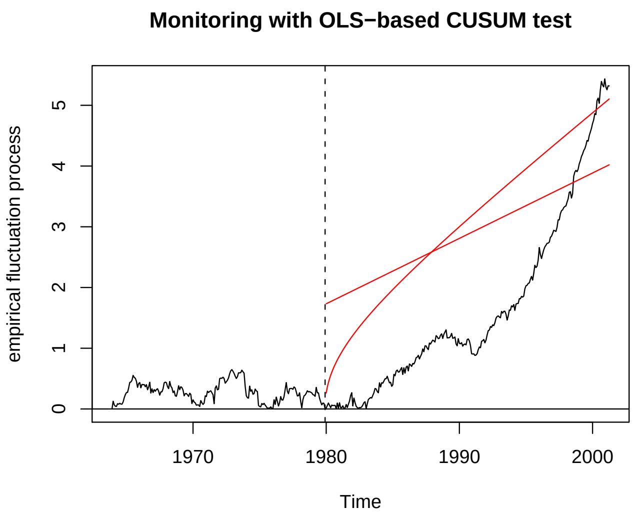


Figure 8: OLS-based CUSUM process

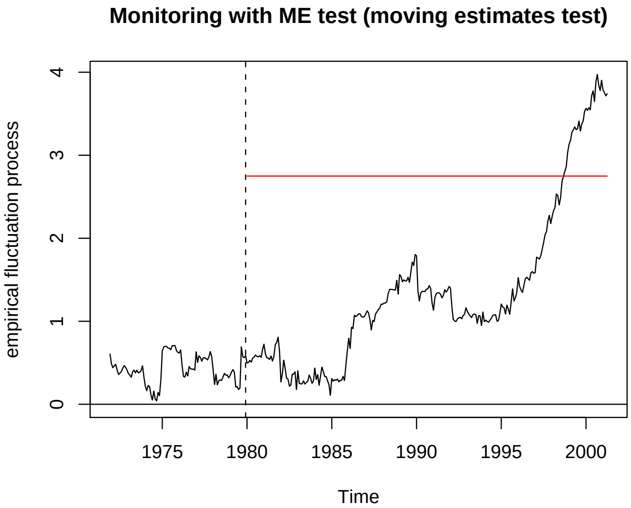


Figure 9: Moving estimates process

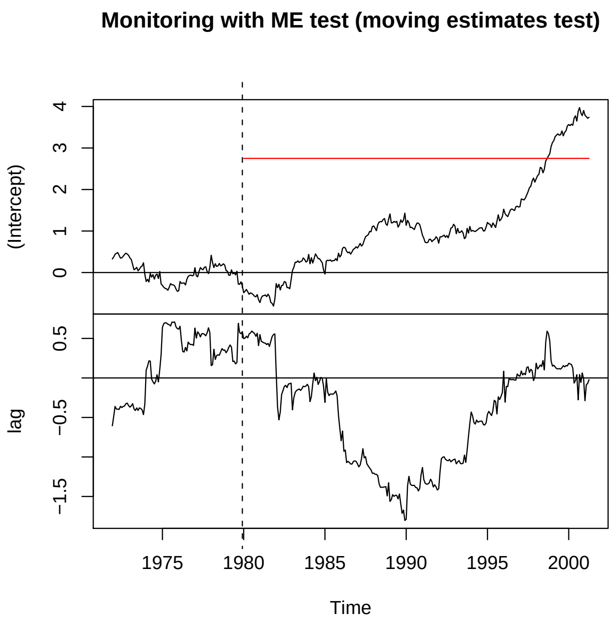


Figure 10: 2-dimensional moving estimates process

 

drifts off somewhat later compared to the CUSUM process. It also does not detect the first change at the 5% level but crosses its boundary after the second change in 1998(10).

The reason that the OLS-based CUSUM test performs better than the ME test on this particular data set is the usage of the new boundaries. An RE test with standard boundaries would even fail to detect a significant change at the 5% level.

However, an additional benefit of the moving estimates is that they shed light on the nature of the break. The 2-dimensional process for the estimates of the intercept and the coefficient on $x_{i-1}$ can be plotted separately as in Figure 10. This shows that the break in the 1980s affects both parameters, because both processes have a shift, but not significantly so. The second break in the 1990s just affects the intercept but not the AR coefficient, thus we are able to conclude that the type of the detected structural shift is instability in the intercept term.

The results of the monitoring procedures confirm the analysis of Hansen (2001) as they are able to convey visually that two changes occured in the early 1980s and the mid-1990s, respectively, matching the breakpoints (estimated by OLS using the full sample) of about 1982 and 1994 rather well. It is also not surprising that monitoring finds evidence only after the second break as the first shift in 1982 was fairly weak compared to the clear shift in 1994.

## 4.3 S&P 500 stock returns

In many situations, e.g., when analyzing financial data, one has to deal with a multitude of high-frequency time series and not every time series can be analyzed in great detail. Therefore, some automatic screening mechanism is needed which selects a small number of "interesting" time series that can be examined more closely in a second stage. If interesting events are associated with a structural change then monitoring can be used as an exploratory tool for screening. This approach is illustrated in the following using a simple intercept model for the individual means of all S&P 500 stock return series.

The data are the daily closing prices $p_i$ for the stocks that are included in the S&P 500. They are available from Yahoo! Finance at http://finance.yahoo.com/, an online portal quoting data provided by Reuters.[1] The sampling period used in this application is the time from 2001-08-01 until the end of the year 2001—using the time before 2001-09-11 as the history period and monitoring the remaining observations. The question is for which stocks the terroristic attacks of September 11 had an impact on the mean return after Wall Street was closed for a week. Employing the screening approach outlined above, each of 484 price series (16 time series had to be excluded from the analysis due to missing data) is monitored for structural changes using the OLS-based CUSUM test (with boundary $b_2$) and the OLS-based MOSUM test. (We consider only OLS-based procedures because they are equivalent to the corresponding estimates-based procedures, both rescaled or not rescaled, in a model with only an intercept.) The monitoring period is $T = 4$ and the significance level of all procedures is again 5%. This setup controls only the individual and not the overall significance level (which is much higher) but this would only have to be adjusted if the task were to find out whether a significant change in any series occured which is not the aim of this application. We rather use an exploratory screening tool for detecting a set of time series that require further analysis.

The OLS-based CUSUM test finds changes in 167 out of 486 series in the full sampling period up to the end of 2001, and the OLS-based MOSUM test finds changes in 235 series. Focusing on the events immediately after the re-opening of Wall Street on 2001-09-17 and 18, the OLS-based CUSUM and MOSUM tests find changes in 20 and 25 of the series, respectively, 17 of which are detected by both procedures. These time series selected by the automatic screening could then be analysed manually.

To illustrate the typical outcome of this approach, we discuss the results of the monitoring procedures in more detail for two quotes: Delta Air Lines (DAL) and Lucent Technologies (LU). We

[1] Yahoo! Finance can also easily be queried from within R.

chose the return series of an airline company where a shock can be expected after September 11 and of a communication company[2] where a shock is less likely.

Estimation of the mean of both return series on the history period from 2001-08-01 until 2001-09-10 yields the following results

$$\begin{aligned} \text{LU:} \quad \Delta \log(p_i) &= -0.013 + \hat{u}_i, \\ \text{DAL:} \quad \Delta \log(p_i) &= -0.007 + \hat{u}_i, \end{aligned}$$

with both intercepts not being significantly different from zero at the 5% level. These two models are subsequently monitored for structural changes following the re-opening of Wall Street after a week on 2001-09-17.

The empirical fluctuation processes for the stock returns of Lucent Technologies (see Figure 11) exhibit only very moderate fluctuation during both the history and the monitoring period; in particular, they do not cross their boundaries. Thus, both the CUSUM and MOSUM procedure agree with the expectation that September 11 had no significant impact on the level of the LU stock returns.

In contrast, the empirical fluctuation processes for Delta Air Lines (see Figure 12) exhibit excessive fluctuations after the end of the history period and both CUSUM and MOSUM test reject the null hypothesis of structural stability after the first observation in the monitoring period on 2001-09-17. Although a type I error can be excluded due to the clear peak in both processes which would have been detected at virtually any significance level (i.e., cannot be explained under the null hypothesis of constancy of the mean) it is unclear whether this single observation is just an outlier or corresponds to a (temporary) change in the mean of the returns. From visual inspection of the fluctuation processes it seems that the slump is followed by an upward trend which causes the MOSUM process to even cross its upper boundary. This suggests that the shock of one extremely large negative return is followed by a sequence of unusually large positive returns. This impression is confirmed by a sensitivity analysis of the procedures for the influence of the 2001-09-17 observation which shows that even if the 2001-09-17 observation would have been within the range of the following returns a significant (upward) shift in the returns would have been detected by both procedures, albeit not after only one observation. In both the original and the sensitivity setup the CUSUM and MOSUM processes return to values around zero at the end of the monitoring period with only moderate fluctuations within the boundaries.

This application shows that monitoring methods are able to provide signals rapidly in a high-frequency context.

## 5. CONCLUSIONS

Online monitoring of regression relationships that are known to be stable for a history period is often more natural and more practical than the commonly employed retrospective tests. In this paper, we have presented a unified approach to the online monitoring of econometric models which includes three new extensions to tests based on regression estimates: processes based on OLS residuals, rescaled processes, and alternative boundaries. These offer advantages concerning ease of computation, finite sample properties in dynamic models, and power against certain alternatives.

These methods might also prove useful in the classical non-sequential context, for instance when a structural change is suspected at the end of the sample following a certain event (see e.g., Fair, 2003). For this situation, Andrews (2003) recently suggested a class of (non-sequential) end-of-sample instability tests.

The determination of optimal asymptotic boundaries, in the sense of minimal detection delay, deserves further study.

[2]Lucent Technologies was chosen because the language S which is implemented in the R system was developed at Bell Labs/Lucent Technologies.

 

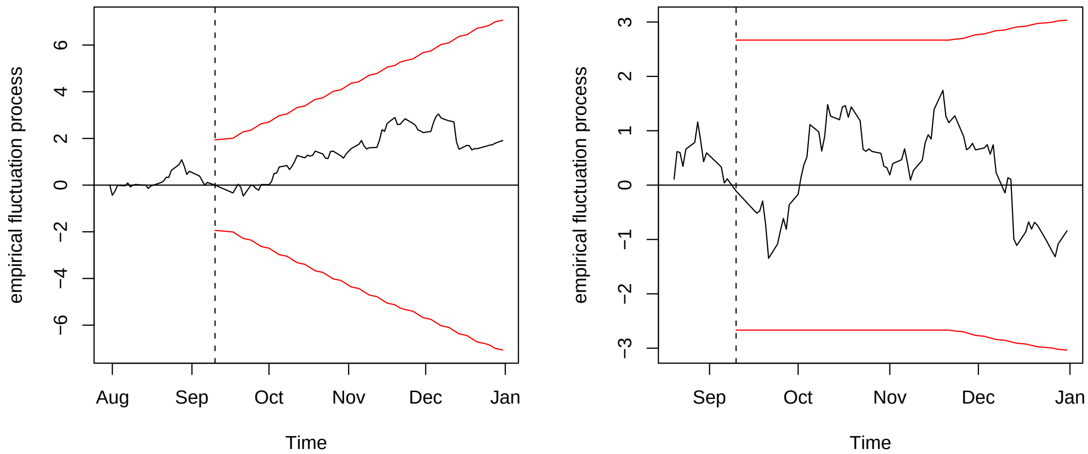


Figure 11: OLS-based CUSUM and MOSUM process for Lucent Technologies

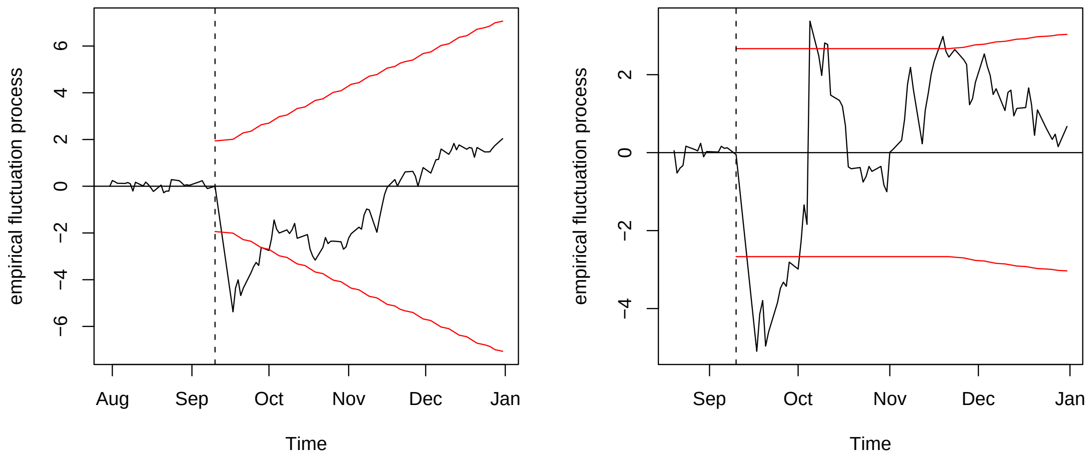


Figure 12: OLS-based CUSUM and MOSUM process for Delta Air Lines

## Acknowledgements

The research of Achim Zeileis, Friedrich Leisch and Kurt Hornik was supported by the Austrian Science Foundation (FWF) under grant SFB#010 ('Adaptive Information Systems and Modeling in Economics and Management Science').

The work of Christian Kleiber was supported by the Deutsche Forschungsgemeinschaft, Sonderforschungsbereich 475.

We are grateful to co-editor Steven N. Durlauf and two anonymous referees for valuable comments on a previous version.

## REFERENCES

Andrews DWK. 1993. Tests for parameter instability and structural change with unknown change point. *Econometrica* **61**: 821–856.

Andrews DWK. 2003. End-of-sample instability tests. *Econometrica* **71**: 1661–1694.

Andrews DWK, Ploberger W. 1994. Optimal tests when a nuisance parameter is present only under the alternative. *Econometrica* **62**: 1383–1414.

Bai J. 1997. Estimation of a change point in multiple regression models. *Review of Economics and Statistics* **79**: 551–563.

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

Carsoule F, Franses PH. 2003. A note on monitoring time-varying parameters in an autoregression. *Metrika* **57**: 51–62.

Chu CSJ, Hornik K, Kuan CM. 1995. The moving-estimates test for parameter stability. *Econometric Theory* **11**: 669–720.

Chu CSJ, Stinchcombe M, White H. 1996. Monitoring structural change. *Econometrica* **64**: 1045–1065.

Cleveland WS. 1993. *Visualizing Data*. Summit, New Jersey: Hobart Press.

Cribari-Neto F, Zarkos SG. 1999. R: Yet another econometric programming environment. *Journal of Applied Econometrics* **14**: 319–329.

Fair RC. 2003. Testing for a new economy in the 1990s. *Business Economics* Forthcoming.

Hansen BE. 1992. Testing for parameter instability in linear models. *Journal of Policy Modeling* **14**: 517–533.

Hansen BE. 2001. The new econometrics of structural change: Dating breaks in U.S. labor productivity. *Journal of Economic Perspectives* **15**: 117–128.

Krämer W, Ploberger W, Alt R. 1988. Testing for structural change in dynamic models. *Econometrica* **56**: 1355–1369.

Kuan CM, Chen MY. 1994. Implementing the fluctuation and moving-estimates tests in dynamic econometric models. *Economics Letters* **44**: 235–239.

Kuan CM, Hornik K. 1995. The generalized fluctuation test: A unifying view. *Econometric Reviews* **14**: 135–161.

Leisch F, Hornik K, Kuan CM. 2000. Monitoring structural changes with the generalized fluctuation test. *Econometric Theory* **16**: 835–854.

 

Lütkepohl H, Teräsvirta T, Wolters J. 1999. Investigating stability and linearity of a German M1 money demand function. *Journal of Applied Econometrics* **14**: 511–525.

Ploberger W, Krämer W. 1992. The CUSUM test with OLS residuals. *Econometrica* **60**: 271–285.

Ploberger W, Krämer W, Kontrus K. 1989. A new test for structural stability in the linear regression model. *Journal of Econometrics* **40**: 307–318.

Racine J, Hyndman R. 2002. Using R to teach econometrics. *Journal of Applied Econometrics* **17**: 175–189.

R Development Core Team. 2003. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-00-3.
URL http://www.R-project.org/

Stock JH, Watson MW. 1996. Evidence on structural instability in macroeconomic time series relations. *Journal of Business & Economic Statistics* **14**: 11–30.

Zeileis A. 2004. Alternative boundaries for CUSUM tests. *Statistical Papers* Forthcoming.

Zeileis A, Leisch F, Hornik K, Kleiber C. 2002. `strucchange`: An R package for testing for structural change in linear regression models. *Journal of Statistical Software* **7**: 1–38.
URL http://www.jstatsoft.org/v07/i02/

 