COMPUTATIONAL STATISTICS & DATA ANALYSIS

Computational Statistics and Data Analysis 52 (2008) 3310–3330

Multiscale spectral analysis for detecting short and long range change points in time series

Lena Ringstad Olsen[a,*], Probal Chaudhuri[b], Fred Godtliebsen[a]

[a] *Department of Statistics, University of Tromsø, Norway*
[b] *Theoretical Statistics and Mathematics Division, Indian Statistical Institute, India*

Received 14 May 2007; received in revised form 30 October 2007; accepted 31 October 2007
Available online 19 November 2007

Abstract

Identifying short and long range change points in an observed time series that consists of stationary segments is a common problem. These change points mark the time boundaries of the segments where the time series leaves one stationary state and enters another. Due to certain technical advantages, analysis is carried out in the frequency domain to identify such change points in the time domain. What is considered as a change may depend on the time scale. The results of the analysis are displayed in the form of graphs that display change points on different time horizons (time scales), which are observed to be statistically significant. The methodology is illustrated using several simulated and real time series data. The method works well to detect change points and illustrates the importance of analysing the time series on different time horizons.

1. Introduction

While monitoring a process, we often measure parameters at certain time intervals. If the process is running properly, the observed time series will usually be stationary. In the analysis of time series data, stationarity is usually an important assumption. A non-stationary process is difficult to analyse, but if we can find stationary segments of the process, these segments can be analysed individually. Hence, we need methods to find if and when the stationarity assumption of a process is violated. A change in statistical properties indicates that something is happening, and we would like to know if and when the process is going out of its stationary state.

There are several practical situations where we want to investigate the stationarity assumption of a process. For a production process, we expect the parameters measured to be stationary when the process is under control and running as scheduled. In such cases we want to know if and at what time point the process becomes non-stationary. The same questions arise in environmental surveys monitoring, for example, for sedimentation rate or pH in a lake: What are the natural fluctuations, has anything been changed significantly, and if so, when? Researchers investigating climate changes measure physical parameters such as temperature, temperature proxies and ice accumulation.

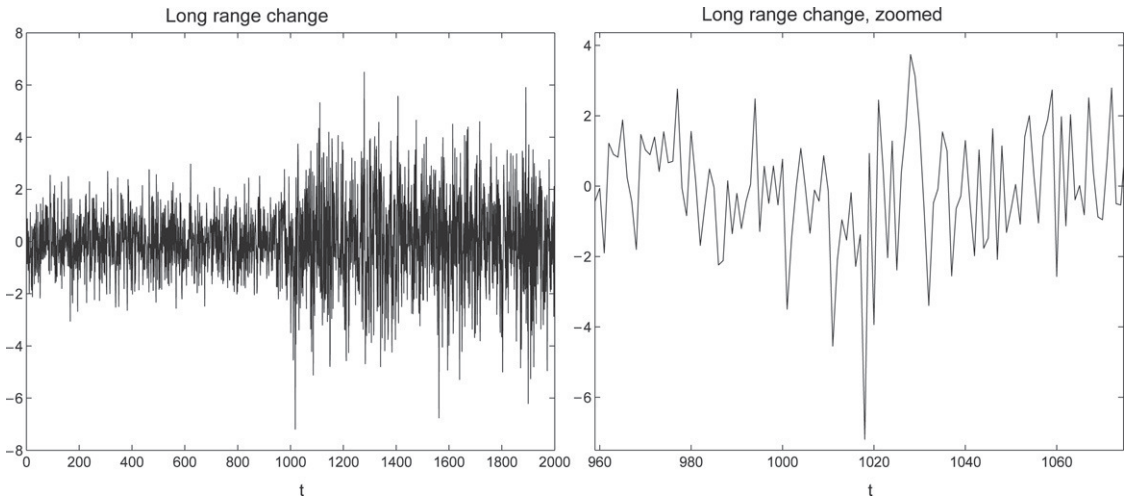

Fig. 1. Long range changes.

The important issues are whether there has been a change in the fluctuations of the temperature or ice accumulation and at which time horizons changes develop. To be able to answer these questions, we need methods that detect significant changes and account for at which scale changes in the statistical behaviour of the measured parameters are detected.

In this paper, change points are interpreted as time points where the stationarity assumption is broken. This means that there are changes in the statistical properties of a process at these time points. We will focus on changes in the time domain that can be found as a change in the power spectrum of a process.

For our analysis of change points, we do not assume any standard model in the time domain like those used in e.g., "intervention analysis" (Box et al., 1994). Further, our objective is not to model the time series for the purpose of making prediction. Balke (1993) detects level shifts and outliers in time series to enable building a model in the time domain for the time series. Polansky (2007) uses Markov chains to model the time series. These examples demonstrate the scope of a large part of the existing time series literature dealing with change points, where the aim is to develop a model for the time series which can be used for further prediction of the process. Other examples are Appel and Brandt (1983), Basseville and Benveniste (1983), Sclove (1983) and Davis et al. (2006) who are using parametric sequential segmentation procedures assuming a piecewise autoregressive process.

Segmentation of a time series by different statistical methods has been considered by several authors. Lai (1995) gives a review of different sequential detection procedures and suggests a class of sequential detection rules for on-line implementation. A common approach for detecting change points, is to use sequential detection procedures based on piecewise modelling of a process. Dahlhaus first addressed the concept of local stationarity; see Dahlhaus (1997), Dahlhaus and Neumann (2001) and Dahlhaus and Sahm (2001). Sato et al. (2007) use wavelet expansions to find locally stationary autoregressive models. In this paper, our objective is not just to identify stationary segments in a time series, but to identify statistically significant change points that exist on different time scales. This provides more insight into the statistical behaviour of time series for different time horizons. Of course, our identification of short and long range change points can be used to identify various stationary segments in a series. Authors like Coates and Diggle (1986), Adak (1998), Mallat et al. (1998) and Donoho et al. (1998) have used various transformations on the time series for identifying stationary segments. Andrews (1993) considers Wald, Lagrange multiplier and likelihood ratio-like tests and gives critical values for the asymptotic distribution. In the present work, we have chosen to use the traditional discrete Fourier transforms and related periodogram estimates that were found to yield good results in our data analysis.

Depending on which time horizon we are looking at, the assessment of whether a process is changing will be different. Fig. 1, left part, shows a time series of 2000 time points where the variation level is changing at $t = 1000$. This change is easily found by visual inspection. Looking more closely at a small neighbourhood around $t = 1000$ (Fig. 1, right part), it is harder to see that there is a change point. This illustrates that for a long time horizon there is clearly a change, but for a short time horizon a change point is harder to locate.

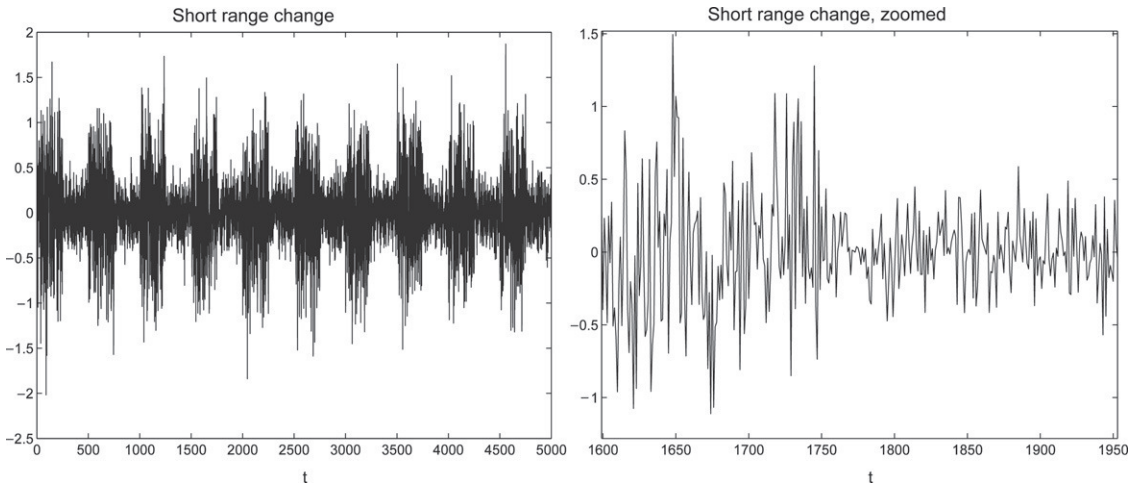

Fig. 2. Short range changes.

Fig. 2 displays a time series of 5000 time points simulated from a $N(0, \sigma^2)$-distribution, where the standard deviation σ changes between 0.5 and 0.2 for segments of 250 time points. Comparing two long time segments (e.g. 1000 time points or more) from this process, the fluctuations between high and low variance will be similar for the two segments, and no change would be claimed. Looking at a shorter time segment, like in the right part of Fig. 2, we interpret the change in variation at $t = 1750$ as a change point.

The examples illustrate that the scale chosen for analysis is of vital importance. If we choose one scale only, we may not be able to detect all changes. In this paper, we will focus on analysis using scale-space which means that we are analysing the data simultaneously for several time horizons. For an introduction to scale-space see Lindeberg (1994) and ter Haar Romeny (2001). In many cases, there is no sudden change, and it is reasonable to allow a structural change a certain amount of time to take its effect. This happens for evolutionary (Priestley, 1965) or quasi-stationary processes (Amin, 1987) where the frequency distribution changes slowly and continuously. Such changes will not be seen on small scales, but will be apparent on longer scales.

In our test procedures, we transform the data to the frequency domain. This is advantageous because it enables us to avoid the need for finding a model in the time domain, for the dependency structure of the time series data, to be able to analyse the process. It also gives us technical advantages like asymptotic independence of the periodogram values for different frequencies, as well as the asymptotic χ^2-distributions of scaled periodogram values. To search for and locate change points, we use two different statistical tests, which are described in Sections 2.1 and 2.2. For both of them we are analysing segments of the time series using their spectra, and the concept of scale is taken care of by using different lengths of the time segments. Some examples to illustrate the method are given in Section 2.4. In all our examples, the scale and the span of a data window is the same since our time steps are one time unit. Technical details about the implementation of the method are given in Section 2.5. Section 3 describes how the method performs in comparison with some other methods and shows some level and power studies. An illustration of the method using real data is given in Section 3.4. Section 4 is a brief discussion and summary of concluding remarks.

2. Methodology

The spectrum of a vector of observations, $\mathbf{x} = [x_1, x_2, \ldots, x_N]$, can be estimated by the periodogram defined as

$$P(\nu_j) = \frac{1}{N} \cdot \left| \sum_{t=1}^{N} x_t \cdot e^{(-2i\pi\nu_j t)} \right|^2 = \left| X_S\left(\nu_j\right) \right|^2 + \left| X_C\left(\nu_j\right) \right|^2 \quad (1)$$

where the periodogram frequency $\nu_j = j/N$, $j = 0, \ldots, N-1$, N is the number of observations from the time domain, and X_S and X_C are the sine and cosine transforms, respectively. A thorough description of mathematical and

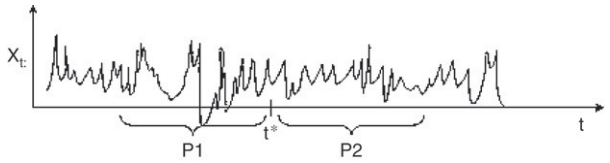

Fig. 3. Illustration of the time segments in the analysis. t^* is the analysed time point, and $P1$ and $P2$ are the periodograms of the observations in the data windows on each side of t^*.

statistical details of periodograms and Fourier transforms can be found in Brockwell and Davis (1991). Assuming that the weak dependency condition, $\theta = \sum_m |m| \cdot |\gamma(m)| < \infty$, is met (where $\gamma(m)$ is the autocovariance of the time series at lag m, and θ is a constant), it is well known that the autocovariance between two different periodogram frequencies approaches zero. The sine and the cosine transforms are asymptotically normally distributed (Shumway and Stoffer, 2000). Hence, zero correlation between periodogram values for different frequencies is equivalent to asymptotic independence between these periodogram values. Tapered versions of the periodogram will usually give a better estimate of the spectrum and reduce the leakage effects. We have chosen to use the periodogram since its values are essentially independent for different frequencies.

The scaled periodogram is asymptotically chi-square-distributed with two degrees of freedom, i.e.

$$\frac{2 \cdot P(\nu_j)}{f(\nu_j)} \sim \chi^2_2,$$

where $f(\nu_j)$ denotes the true spectral density evaluated at the frequency ν_j. When the two segments under comparison are both parts of a stationary process, the components $P2(\nu_j)/P1(\nu_j)$, $j = 0, 1, \ldots, N/2$, are approximately independent and identically $F_{2,2}$-distributed for large N. A sum of periodogram values over L frequencies will be approximately F-distributed with $2L$ and $2L$ degrees of freedom ($F_{2L,2L}$).

To search for and locate change points, we calculate the periodograms (Eq. (1)) for segments of the time series using data windows, and try to determine whether there is a change in the periodogram from one time segment to the next. The number of observations in a time segment is given by N. A change in the spectrum indicates that there is a change in the statistical properties of the time series and that the process is non-stationary. Fig. 3 illustrates two consecutive time segments for which the periodograms are calculated and compared by some test. If two processes are different, the patterns of these two periodograms will be different.

The estimated spectra of the time series data from two consecutive data windows ($P1$ and $P2$), on each side of the time point being checked (t^*), are compared by looking at the ratio of the two periodograms. More precisely, at time point t^*, one gets the ratios

$$R(\nu_1, \ldots, \nu_{\frac{N}{2}}) = \left[\frac{P2(\nu_1)}{P1(\nu_1)}, \ldots, \frac{P2(\nu_{\frac{N}{2}})}{P1(\nu_{\frac{N}{2}})} \right],$$

where N equals the number of observations in each of the data windows, from which $P2$ and $P1$ are calculated. Since the periodogram is symmetric around the Nyquist frequency ($\nu_{N/2}$), we only use the first half of the periodogram. The DC component (ν_0) tends to obscure the periodogram, and this frequency is removed from the analysis. The DC component represents the mean of a time series, and therefore the method is not so good at detecting changes in the mean. For this purpose Mean SiNos for dependent data, Olsen et al. (in press), or SiZer for independent data, Chaudhuri and Marron (1999), will be a better choice.

2.1. Test statistic for comparing segments of time series

At each of the time points we want to test, we calculate the periodograms of a data window just before and just after the time point and find the ratio of these two periodograms. Our default test statistic for testing for non-stationarities is based on the mean of the periodogram ratios (*Mean Ratio Test*). Alternative effective tests could be chosen, but for our purpose this test works well.

The spectra in window 1 and window 2 are given by $f_1(\nu_1), \ldots, f_1(\nu_{N/2})$ and $f_2(\nu_1), \ldots, f_2(\nu_{N/2})$, respectively. We want to test $H_0 : f_1(\nu_j) = f_2(\nu_j)$ or equivalently $H_0 : f_2(\nu_j)/f_1(\nu_j) = 1$, $j = 1, 2, \ldots N/2$.

A natural approach is to use the periodogram values to estimate these ratios and then test H_0. The periodogram is not a consistent estimator of the spectral density; hence we look at a coarser level of the problem and test whether the spectra are equal over a frequency band, $B_k = \{\nu : \nu_k \pm (L-1)/(2N)\}$, where L is the number of periodogram values summed in the frequency band, and the bin centres are given by $\nu_k = ((2k-1)L+1)/(2N)$. This means that we test: $H_0 : f_2(\nu_k)/f_1(\nu_k) = 1, k = 1, 2, \ldots, K$, where $K = \lfloor N/(2L) \rfloor$. When H_0 is true, the mean ratio (MR) equals 1: $\mathrm{MR} = 1/K \sum_{k=1}^{K} f_2(\nu_k)/f_1(\nu_k) = 1$.

To test whether there is a change in the statistical properties, we therefore test $H_0 : \mathrm{MR} = 1$ against $H_1 : \mathrm{MR} \neq 1$. A natural choice of the test statistic is then

$$\widehat{\mathrm{MR}} = \frac{1}{K} \sum_{k=1}^{K} \frac{P2_L(\nu_k)}{P1_L(\nu_k)},$$

where $Pi_L = 1/L \sum_{j=(k-1)L+1}^{kL} Pi(\nu_j)$, $i = 1, 2$. This means that $P1_L$ and $P2_L$ are the periodograms, averaged over L frequencies, of the observations in the time segments illustrated in Fig. 3.

Note that $Cov\,(P1\,(\nu_k)\,, P2(\nu_l)) \rightarrow 0$ for all Fourier frequencies. This also holds when $k = l$. Hence the conditions of the F-distribution are asymptotically fulfilled for the ratio of $P2_L$ and $P1_L$. The variables, $F_k = P2_L(\nu_k)/P1_L(\nu_k), k = 1, 2, \ldots, K$, are asymptotically (as $N \rightarrow \infty$) independent and identically $F_{2L,2L}$-distributed. This means that $E\,[F_k] = 2L/(2L-2)$ and hence $E\left[\widehat{\mathrm{MR}}\right] = 2L/(2L-2)$. We reject H_0 when $\widehat{\mathrm{MR}}$ is significantly different from $2L/(2L-2)$. In the Mean Ratio Test, we test whether the expectation of the ratio of $P2_L$ and $P1_L$ equals the expectation from an $F_{2L,2L}$-distribution.

The number of defined moments for the F-distribution depends on the second degree of freedom in the F-distribution. We will avoid smoothing too much so that we lose features of the periodogram. Hence, the choice of L is a trade-off between getting a well "behaved" F-distribution and possibly losing features of the periodogram. The variance of an F-distribution is defined when the second degrees of freedom is more than or equal to 5. In the Mean Ratio Test, $L = 3$ is chosen to get finite second-order moments. To get a two-sided test, we include the inverse ratios in the test which means that we test both $\widehat{\mathrm{MR}} = (1/K) \sum_{k=1}^{K} P2_L(\nu_k)/P1_L(\nu_k)$ and $\widehat{\mathrm{MR}^I} = (1/K) \sum_{k=1}^{K} P1_L(\nu_k)/P2_L(\nu_k)$.

In practice, it is not necessary to test for changes at every single time point. If we change the tested time point only by one time step from one test to the next, two consecutive tests will be very similar since only one observation in each data window is new. A better strategy is to skip some time points between two tested time points. We do not, however, want to miss any change points. Hence, there is a trade-off between testing too many and testing enough. Not testing all time points could mean that we do not detect "the one" interesting change point. This may also happen by not performing tests using all possible window widths. But testing all time points and possible scales would be far too time-consuming. Our method is an exploratory method, and we are not worried about changes detectable for only one time point since such changes are really hard to separate from spurious detections.

We also want the tests to be relatively independent. We have chosen to use a time shift of 20% of the window width for all our subsequent data analysis. This means that when the window width is 50, every 10th time point is checked, while when the window width is 200, every 40th time point is checked. In this way, the number of tests will be considerably reduced. Simulation studies of the correlation between consecutive tests indicate that using a time shift of 20% of N gives a correlation less than 0.2 for window widths bigger than 50. A more detailed description of time shifts is given in Section 2.5.1.

To get a more stable test and increase the power, we are simultaneously testing the neighbouring windows in each test. A detailed description of the neighbouring testing and applied details of the methodology are given in Section 2.5.

2.2. An alternative test for detecting changes in AR processes

For some AR processes, the Mean Ratio Test has low power, and we wanted to use a test more sensitive to changes in such processes. The Distribution Test finds changes in the spectrum by comparing the distribution of the periodogram ratios for different groups of frequencies. The ratios are split into two groups: one group with the periodogram ratios for the low frequencies ($\nu_k \leq 0.25$) and one group with the high frequencies ($\nu_k > 0.25$). This

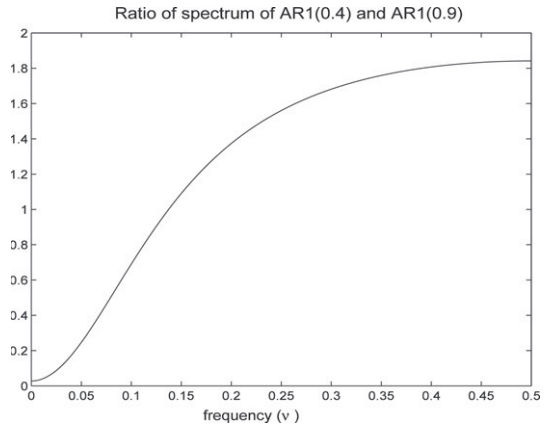

Fig. 4. Ratio of the spectrum of an AR1(0.4) process and that of an AR1(0.9) process.

means that we compare the "two samples"

$$R(\nu_1, \ldots, \nu_{\frac{N}{4}}) = \left[\frac{P2(\nu_1)}{P1(\nu_1)}, \ldots, \frac{P2(\nu_{\frac{N}{4}})}{P1(\nu_{\frac{N}{4}})} \right]$$

and

$$R(\nu_{\frac{N}{4}+1}, \ldots, \nu_{\frac{N}{2}}) = \left[\frac{P2(\nu_{\frac{N}{4}+1})}{P1(\nu_{\frac{N}{4}+1})}, \ldots, \frac{P2(\nu_{\frac{N}{2}})}{P1(\nu_{\frac{N}{2}})} \right].$$

Our test statistic is the Kolmogorov–Smirnov (K–S) distance between the distributions of the two groups of ratios. The K–S distance is denoted by $KS(R(\nu_1, \ldots, \nu_{N/4}), R(\nu_{N/4+1}, \ldots, \nu_{N/2}))$. If the statistical properties of the process are the same for the two segments, the distribution of the ratio of high and low frequencies will be the same. In the case of a scale change of the spectrum (change only in the variance of the process), the two groups of ratios used in our Distribution Test will have the same distribution, and no change will be detected by this test. Such a change will be detected by the Mean Ratio Test.

A natural alternative choice of test statistic could be a one-sample Kolmogorov–Smirnov test, testing for deviations from the F-distribution. This approach did not work very well, and the proposed test was chosen after examining a number of other test statistics. Several different test statistics for testing the difference between two estimated spectra have been proposed; see Coates and Diggle (1986) and references therein. Splitting the ratio vector in the middle is motivated by looking at the ratio of the spectrum of an autoregressive process of order 1 (AR1 process) with zero mean. Such an AR1 process is given by $x_t = \phi \cdot x_{t-1} + \varepsilon_t$, where ϕ is the first-order autocorrelation coefficient and ε_t is Gaussian white noise. The power spectrum of an AR1 process with correlation coefficient ϕ and residual variance σ^2 is given by $f(\nu) = \sigma^2/(1 + \phi^2 - 2 \cdot \phi \cdot \cos(2 \cdot \pi \cdot \nu))$.

Fig. 4 shows the ratio of the spectra of two AR1 processes with correlation coefficient 0.4 and 0.9, respectively. The ratio is monotonically increasing, and the plot demonstrates that the differences between groups of ratios are largest when we compare the ratios for very high and very low frequencies. In this situation, the best way to detect a change will be to compare the upper and lower halves of the periodogram ratios.

2.3. Multiple testing

One of the challenging parts in our testing procedure is adjusting for multiple testing. For each window width, we are doing a lot of tests along the time series. The correct adjustment is hard to find when the distribution is not specified. In our case, the tests are overlapping which makes the consecutive tests correlated. There is no standard way to adjust for multiple testing over sliding windows comparing consecutive segments of the same series. The false discovery rate, FDR, (Benjamini and Yekutieli, 2002) and Holm's procedure (Holm, 1979) are two techniques that may be used. Holm's procedure controls the familywise error rate, but is less conservative than the Bonferroni

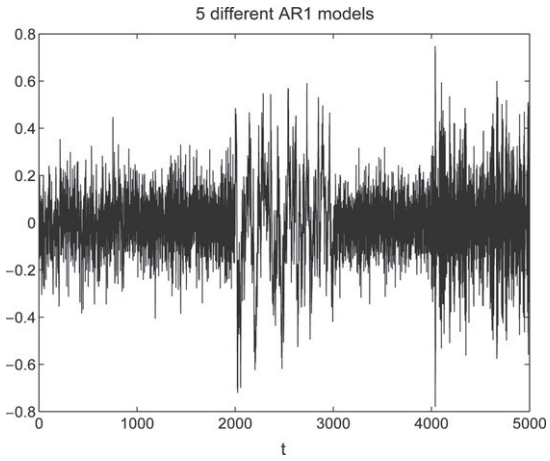

Fig. 5. Data simulated from five different AR1 models, using $\phi = 0.5, -0.5, 0.9, -0.1$ and -0.9, respectively. 1000 data points are simulated from each model.

procedure. The FDR has a different interpretation and controls the number of false rejections of the null hypothesis. The FDR is less conservative than Holm's procedure. The simplest version of the FDR is equal to Simes' procedure (see Simes (1986), or Sarkar and Chang (1997)). Simes' method controls the type I error nominal level and is always more powerful than the Bonferroni procedure for dependent test statistics.

We have looked at three different approaches to the adjustment of simultaneous testing. In one approach, simulated critical values are found by simulating the maximum over the same number of tests as will be performed in the specific time series investigated. In this way the adjustment of the significance level is automatically taken care of. An alternative approach is to use the critical value from the F-distribution for each single test and adjust for simultaneous testing by the FDR or Holm's procedure. The results are given in Sections 3.2 and 3.3.

2.4. Illustrations of the methodology—synthetic data

Technical details on the implementation of the methodology will be given in the next section. First, we are presenting some simple illustrations of the methodology using simulated time series data. In all examples, simulated critical values have been used to adjust for multiple testing. This section is essentially to demonstrate the methodology so that the reader can be familiar with the output of the proposed method. This is why we have chosen very simple examples that illustrate only the basic points about our change point detection methodology. There are more simulations in Sections 3.2 and 3.3, where we learn much more about the performance of the methodology by looking at the observed level and power. At the end of the paper, there are analyses of real data, where we discover features in real time series and try to interpret and understand them.

The significant changes found by a test method are displayed in a significance plot where the horizontal axis shows at which time points and the vertical axis shows for which scales (lengths of the time segments) a change is observed to be statistically significant. If there is a significant difference between the observations in two time segments compared, the pixel in the plot at the actual time point t^* and window width N is coloured white. This means that white areas in the plot show time points at time horizons where there are significant differences between the two time segments compared and hence significant changes in the time series. The significance plot displays dark grey at time points where the test does not find a significant change. Light grey is shown if there are less than N data points on one side of t^*, which occurs near the boundaries.

2.4.1. Five AR1 models

Fig. 5 shows a time series where 1000 time points are simulated from each of five different AR1 processes. The four change points occur at $t = 1000, 2000, 3000$ and 4000. By visual inspection of the plot, the three last change points are easy to identify by eye, but not the first.

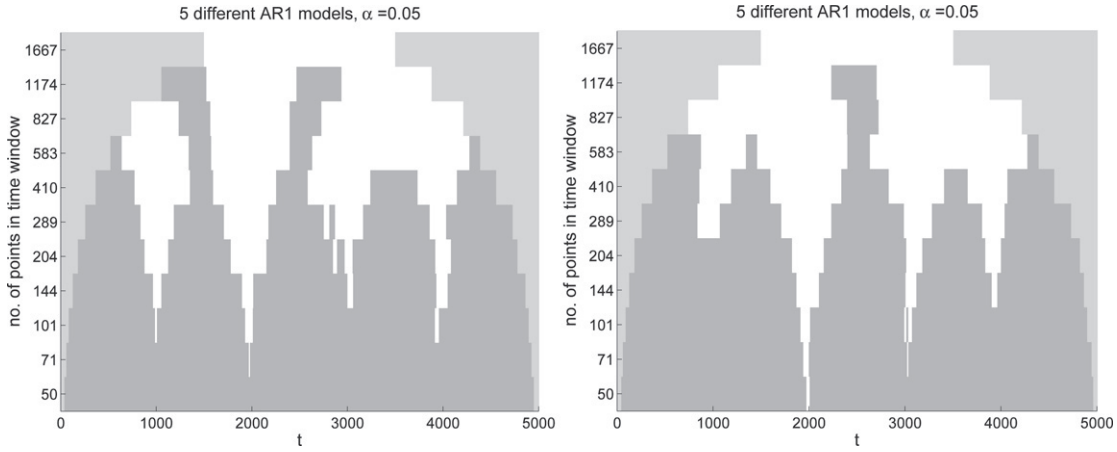

Fig. 6. Distribution Test (left) and Mean Ratio Test (right) analysis of the AR1 data in Fig. 5.

The left part of Fig. 6 shows the significance plot obtained using the Distribution Test for the time series in Fig. 5, when the time point tested is shifted by 20% of the window width. Simulated critical values are used to adjust for simultaneous testing. We see from the plot that at a scale of 50 time points in the time window, no changes are detected. This means that on a time horizon where the observations in two segments of 50 time points are compared, the Distribution Test does not detect any of the changes. For window width 71, there is a white pixel at $t = 2000$ which means that the change occurring at this time point is detected. For window width 101, there are white pixels in the plot around $t = 1000$, 2000 and 4000. These white streaks imply that a change is detected. From window width 144 and coarser, there are white pixels at and around all the known change points ($t = 1000$, 2000, 3000 and 4000). Hence, all the change points are detected for a time horizon of 144 time steps or more.

An analysis using the Mean Ratio Test is shown in the right part of Fig. 6. Also the Mean Ratio Test detects all four change points. The biggest difference between the two methods is for the change at $t = 1000$. Here the Distribution Test detects a change for all window widths of 101 time points or more, while the Mean Ratio Test starts detecting the change at window width 289. For the change points at $t = 2000$, 3000 and 4000 the window width for which the change point starts to be detected differs only by one or two scale units. (Note that the scale unit is logarithmic.) The Distribution Test has a slightly more distinct detection of the change points. For this kind of time series data, the Distribution Test seems to work better than the Mean Ratio Test. This is as expected since the construction of the Distribution Test is motivated by the spectrum of an AR1 process. (It is also in agreement with the power studies in Section 3.3.)

As the window width increases, the number of time points where a change is detected increases. This is expected since when the length of the two time segments compared increases, the number of time points where one of the segments covers two different processes increases. Hence, a difference between two segments will be detected in a longer time span around the change point.

2.4.2. *Three AR02 models*

Suppose the true underlying model is an AR2 model, given by $x_t = \phi_1 \cdot x_{t-1} + \phi_2 \cdot x_{t-2} + \varepsilon_t$. Fig. 7 shows a time series where 1000 time points are simulated from each of three different AR2 models where $\phi_1 = 0$ which means that the first lag autocorrelation $\rho = 0$. (This is what we call an AR02 model.) By visual inspection, we observe that there are two change points. The change at $t = 2000$ is very clear. It is also relatively clear that something is happening around $t = 1000$, but the exact location of the change is not obvious.

Fig. 8, left part, shows a Distribution Test analysis of the time series in Fig. 7. The Distribution Test does not detect any of the change points. This is because a change in ϕ_2 mainly affects the variance of the process and the Distribution Test is not able to detect such changes.

The right part of Fig. 8 displays the Mean Ratio Test analysis of the AR02 data. This test correctly detects a change at $t = 1000$ and at $t = 2000$, and the changes are detected for all time scales.

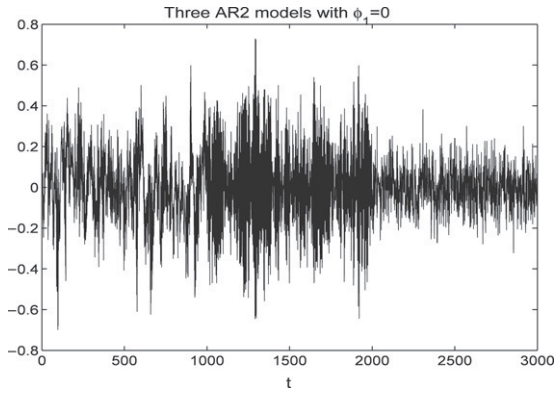

Fig. 7. Data simulated from three different AR2 models. 1000 data points are simulated from each model where $\phi_1 = 0$ and $\phi_2 = 0.8, -0.9$ and 0.2, respectively.

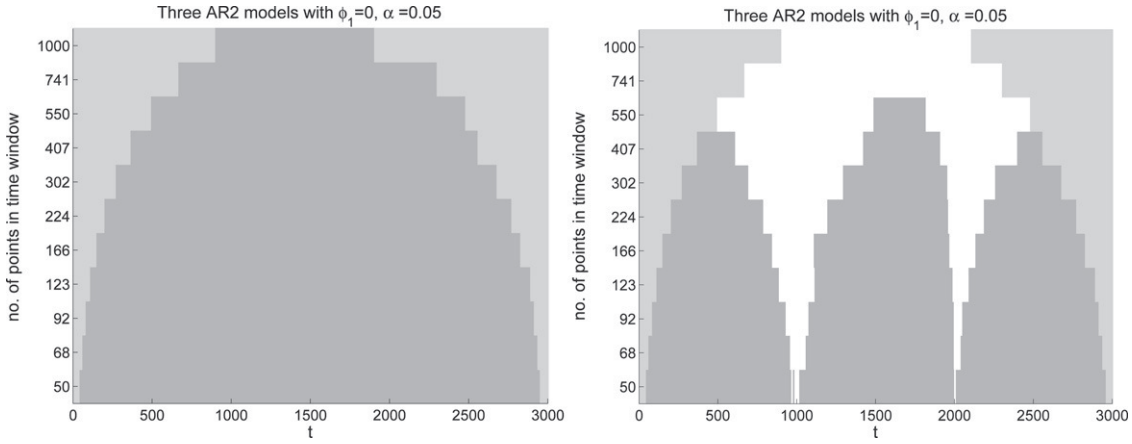

Fig. 8. Distribution Test (left) and Mean Ratio Test (right) analysis of the AR02 data in Fig. 7.

2.5. Implementational details

2.5.1. Time shifts

Recall that $P1$ and $P2$ are the periodograms of the N observations before and after, respectively, the time point we want to test. Fig. 9 shows a plot of the correlation between consecutive ratios calculated from overlapping time segments for several window widths. The shift, expressed as a percentage, is the number of time steps from one checked time point to the next relative to the window width (N). If the time shift is 40%, and the window width is 100, every 40th time point is checked. A shift of 40% also means that 40% of the observations in the periodogram $P1$ are different from the observations in the $P1$ in the next test. In this case, the overlap from one *pair* of windows (i.e. the $2N$ observations from which $P1$ and $P2$ are calculated) to the next is 80%.

As expected, the correlation between the tests decreases when the time shift from one tested time point (t^*) to the next increases. Shifting t^* by 10%, we observe a correlation between 0.1 and 0.4. The correlation is highest for small window widths, and it slightly increases with increasing autocorrelation of the time series process. The correlations displayed in Fig. 9 represent window widths $N = 50$, 100, 200, 500, and 1000, time shifts of 10%, 20%, ..., 200%, and time series from three different time series models. The correlation decreases with increasing window width. Looking at, for instance, the observed correlations using 10% shift, the lowest light triangle shows that the observed correlation for the AR1(0.9) process is 0.1 when $N = 1000$. From Fig. 9 we see that the correlation decreases with increasing time shift and crosses zero when the shift is 60%. For shifts between 70% and 130% the correlation is negative and for longer time shifts the correlation is approximately zero.

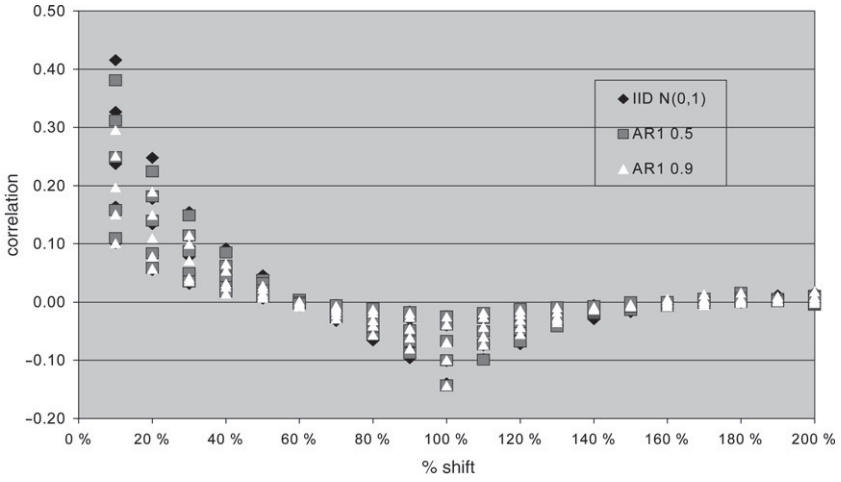

Fig. 9. Correlation between consecutive ratios using different time shifts of t^*.

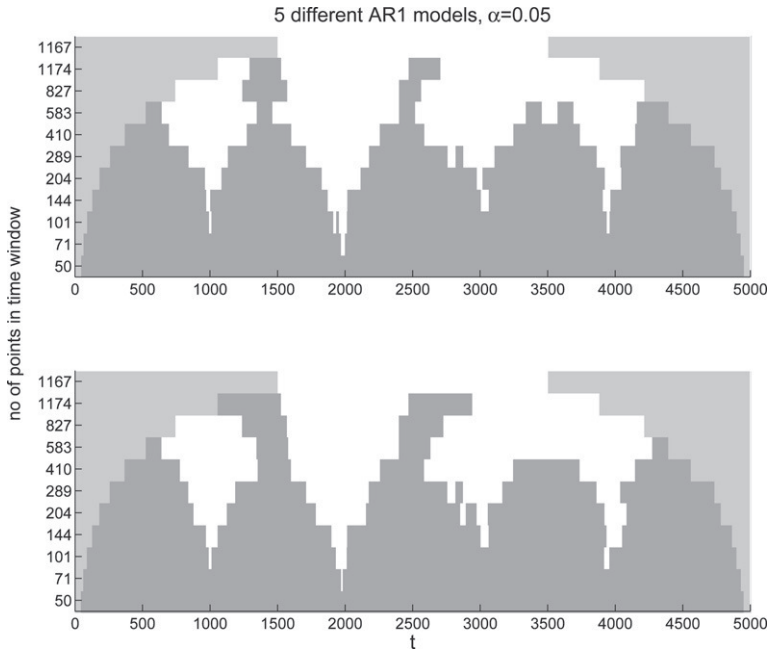

Fig. 10. Upper: Distribution Test analysis of five AR1 processes, shifting t^* by 20% of N (but not including neighbours). Lower: The same analysis but including neighbouring segments in the test.

Analysing every single time point, there would be white areas broken by grey pixels around the change points. This is because close to the change point a part of the change will occur in one of the data windows and the difference will be on the edge of being significant such that some tests will turn out to be significant and some will not. By reducing the number of simultaneous tests, we avoid most of the unstable detections around the change points and speed up the calculations considerably. We have chosen to use a shift of 20% for our data analysis. The simulations shown in Fig. 9 indicate that 20% shift gives a correlation less than 0.2 for window widths bigger than 50.

2.5.2. Including neighbouring segments in the tests

The upper part of Fig. 10 shows the significance plot for the time series of five different AR1 processes displayed in Fig. 5, when the time point tested is shifted by 20% of the window width.

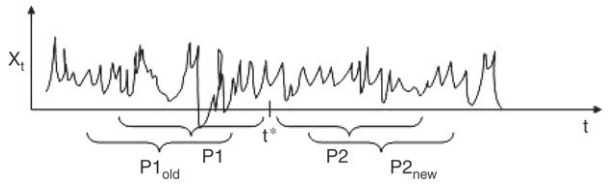

Fig. 11. Illustration of the time segments in the test including neighbours.

To get a clearer significance map with fewer grey streaks in the neighbourhood of the detected change points, we also test simultaneously the neighbouring windows. The idea of testing neighbours is from Ligges et al. (2002) who suggest a solution to the problem of finding locally stationary time segments in a series arising from a vocal sound signal. They use a two-sample Kolmogorov–Smirnov test to test whether there is a difference between two consecutive periodograms. For each tested time point, they also include the neighbouring periodograms from non-overlapping time segments and accept significance if all the three tests show a significant change. Including the neighbours makes the algorithm more robust.

Adjusting our tests based on comparisons of neighbouring segments is carried out as follows: The periodograms $P1$ and $P2$, on each side of the tested time point, are calculated as before. In addition, the periodogram for the observations from the window shifted by 20% of N to the right of $P2$, $P2_{\text{new}}$, and the periodogram shifted to the left of $P1$, $P1_{\text{old}}$, are calculated. See Fig. 11 for an illustration. We now get the ratios $R_{\text{old}} = P2/P1_{\text{old}}$, $R = P2/P1$ and $R_{\text{new}} = P2_{\text{new}}/P1$. This means that the two periodograms used in the numerator and the denominator in a ratio are calculated from time segments which never overlap, but the time segments used for periodograms $P1_{\text{old}}$ and $P1$, and for $P2$ and $P2_{\text{new}}$ overlap. For each of the ratio vectors R_{old}, R and R_{new} we calculate the test statistic for the Mean Ratio or Distribution Test. The intersection of the three tests is tested, and significance is shown in the significance plot only if all three tests show a significant difference. This means that our test statistic is given by the minimum of the three test statistics, i.e. $\min\left(KS\left(R_{\text{old}_1}, R_{\text{old}_2}\right), KS\left(R_1, R_2\right), KS\left(R_{\text{new}_1}, R_{\text{new}_2}\right)\right)$ (where subscripts 1 and 2 means the ratios for $\nu_1 < \nu \leq \nu_{N/4}$ and $\nu_{N/4} < \nu \leq \nu_{N/2}$, respectively) for the Distribution Test and

$$\min\left\{\max\left(\widehat{\text{MR}}_{\text{old}}, \widehat{\text{MR}^I}_{\text{old}}\right), \max\left(\widehat{\text{MR}}, \widehat{\text{MR}^I}\right), \max\left(\widehat{\text{MR}}_{\text{new}}, \widehat{\text{MR}^I}_{\text{new}}\right)\right\}$$

for the Mean Ratio Test.

The lower part of Fig. 10 shows a Distribution Test of the AR1 process from Fig. 5 where the neighbouring segments are included in the test. (The lower part of Fig. 10 is the same as that displayed in the left part of Fig. 6). Compared to the upper part of Fig. 10, it shows that simultaneous testing of neighbours improves the result. For this example there are no spurious detections and only two grey streaks near the detected change point at $t = 2000$. These occur for window width 204 and 289, and the separation of the change points at $t = 1000$ and 2000 is better.

2.5.3. Critical values

As explained in Section 2.5.2, a change is accepted as significant if the tests of all the three ratios show significance. Hence, we can compare the minimum of the test statistic for the three ratios with the appropriate critical value for this test. The crucial question is how this critical value can be found. Our approach is based upon the fact that, for stationary series where the weak dependency condition is met, the periodogram ratios $P2/P1_{\text{old}}$, $P2/P1$ and $P2_{\text{new}}/P1$ are asymptotically distribution free in nature. By this, we mean that the components within any of the three above-mentioned ratios are essentially independent and $F_{2,2}$-distributed for a wide range of distributional models, including the standard models such as IID normals and AR processes.

Note that this result only holds asymptotically, i.e. when the window width, N, is large. Simulations indicate, however, that for a variety of stationary time series (e.g. IID or AR processes), the above result holds even for small values of N. In other words, the periodogram ratios tend to have distributions that appear to be approximately independent of the statistical properties in the time series. Hence, for a given window width N and time series length T, the critical values in the test can be found by simulating time series from e.g. an IID $N(0, 1)$-distribution or simulating ratios from an F-distribution. In our approach, we simulate realizations of length T from an $N(0, 1)$-distribution. Then we calculate the test statistics for all chosen time points (for a given N). For each simulated time series of length T, we choose the maximum value of these test statistics. In this way we also automatically adjust for

simultaneous testing. This procedure is repeated 10 000 times. Finally, for a test with target level α, the α-quantile, $q_{N,T}$, for this (N, T) pair is chosen as the value where α per cent of the 10 000 simulated test statistics values are greater than $q_{N,T}$. Simulation results indicate that this approach is plausible as long as the dependence structure in the original time series is not too strong. In e.g. the case of an AR1 process, the parameter ϕ should be less than 0.9 for the above arguments to hold.

3. Results

3.1. Comparison to other methods

Scale-space approaches using wavelets represent an alternative to our approach. The concept of scale in our approach is the same as in wavelet analysis in the sense that scale can easily be related to physical scale in the time domain for both methods. For more details about the wavelet scale-space technique, see Percival and Walden (2000). The paper by von Sachs and Neumann (2000) describes a wavelet procedure that gives an objective answer to the question of whether and where non-stationarities occur in a time series. To this end, they use several scales to detect changes in the covariance structure. Their aim is different from ours since we focus on the fact that the detected non-stationarities on one scale can be very different from detected non-stationarities on another (very different) scale. In contrast to our approach, von Sachs and Neumann (2000) limit their study to dyadic scales. Moreover, their correction for multiple testing is very different from our procedure for handling this problem.

Ombao et al. (2001) describe a procedure, SLEX (smooth localized complex transform), which is related to ours. The SLEX model uses the SLEX vectors which are orthogonal and localized Fourier vectors in the Cramér representation of the spectrum. It does, however, differ from ours in at least two important ways. First, it is not an issue in Ombao et al. (2001) that the non-stationarities found depend greatly on what scale the analysis is performed on. Second, we detect non-stationarities in a more objective way than Ombao et al. (2001) since we are using statistical quantiles, while Ombao et al. (2001) use subjectively chosen cost functions and limits, to decide when data corresponding to two different time intervals come from two different models. For more information about SLEX, see Ombao et al. (2002) and Huang et al. (2004).

Bai (1997) uses a test which is based on a parametric model and tests whether there is a change in one of the parameters. If change points are found at a level, the new intervals have to be tested over again to see whether there are significant change points using other samples from the time series. The examples given are based on a regression model with two regressors. In our case, we just have a time series and no regressors. We want to avoid the need for estimating parameters. Bai's test splits a time series in pieces due to where there may be a change point according to the Wald test. This means that Bai (1997) do not keep the sense of scale, just chop the time series into pieces.

Davis et al. (2006) consider the problem of modelling a class of non-stationary time series using piecewise AR processes. The primary objective is to estimate structural breaks for a time series. The number and locations of the piecewise AR segments, as well as the orders of the respective AR processes, are assumed unknown. The goal is to find the "best" combination of the number of segments, the lengths of the segments, and the orders of the piecewise AR processes. Assuming that the true underlying model is a segmented autoregression, this procedure is shown to be consistent for estimating the location of the breaks. We want to avoid any model estimation, and our objective is to focus on the importance of scale. This is not an issue in Davis et al. (2006). An approach similar to ours, including the AR modelling of Davis et al. (2006), could be to find the best AR model for each window and compare the windows. In this way the scale-space could be added to the analysis.

Our focus is different from those of other methods for change point detection. We consider change points that are statistically significant at different time scales. SiZer (Chaudhuri and Marron, 1999) and Dependent SiZer (Park et al., 2004) are similar methods. SiZer assumes a regression model where the residuals are independent, and Dependent SiZer assumes that the autocorrelation function is known. Both methods only detect changes in the mean. A Bayesian multiscale approach is described in Øigård et al. (2006). Park et al. (2007) describe a scale-space method combining wavelet analysis with the visualization tools SiZer (Chaudhuri and Marron, 1999) and SiNos (Olsen et al., in press). A method that is fairly close and comparable to what we are doing is that of the algorithm described in Adak (1998). Adak (1998) also uses a nonparametric procedure for change point detection, but Adak (1998) does not consider a multiscale approach like ours. Adak (1998) segments the time series, running through several levels of a dyadic tree which at each level divides the segments of the time series into two new segments of dyadic length. We have chosen to make a comparison to the TASS procedure in Adak (1998).

3.1.1. Comparison with the tree-based segmented spectrogram algorithm (TASS)

To illustrate the tree-based adaptive segmented spectrogram algorithm (TASS), Adak (1998) uses a stationary AR2 process to study the observed level. The same process, where one of the coefficients is changing, is used to study the power. The terms observed level and power are not explicitly discussed in the paper. In Adak (1998) there are two tables showing the distribution of how many nodes are detected, and the performance of the algorithm in finding the stationary segmentation is discussed.

The first stationary AR2 process is given by $x_t = 1.69x_{t-1} - 0.81x_{t-2} + \varepsilon_t$ and the second is given by $x_t = 1.38x_{t-1} - 0.81x_{t-2} + \varepsilon_t$. In both cases the ε_ts are IID $N(0, 1)$-distributed. Adak (1998) uses the first AR2 model to study the observed level. For power studies, a process where the first AR2 model is changing to the second AR2 model in the middle of the time series is used. The simulation studies are performed using 5000 simulations.

Using the stationary series, Adak (1998) detects more than one node in 17% to 48% of the simulations, depending on the distance measure used, without using any penalty on the number of segments. Including a penalty parameter, between 6% and 13% of the simulations detect more than one segment. This means that the observed level of the TASS algorithm is between 6% and 13% for this series. A nominal level is not given as there is no use of "statistical significance" in this sense in the algorithm.

The overall observed levels that we get for the two AR2 processes using our methods are quite different depending on which of the two processes we choose. Adak (1998) only shows the results using the first AR2 model and not the second. For the first AR2 process the DT and the MRT have observed levels of 25% and 41%, respectively, averaged over 11 window widths using a nominal level of 5%. The observed level is strongly decreasing with increasing window width for the MRT, and the largest window width gives an observed level of 11%. This indicates that for this process, the convergence to the χ^2-distribution for the squared periodogram values is slow. For the DT, the observed level is low (6% and 11%) for the two smallest window widths, high for the mid-range window widths (up to 49%) and decreases to 11% for the largest window width. Be aware that the apparently very bad observed level is the *overall* level for each window width. For example, if the level is 50% this means that in 50% of the simulations there is at least *one* time point for which a change is detected at the given scale. That is, a level of 50% does *not* mean that a change is detected in 50% of the time points at a scale. If we look at a significance plot, it is relatively clear that these detections are spurious detections. Hence, it is not straightforward to compare the numbers reported by Adak for the TASS algorithm and our overall observed levels, directly. If we compare using the largest window width from our procedure, the observed levels are fine and clearly compete with TASS.

If we look at the second AR2 process, the observed level averaged over the scale is 7.1% and 7.6% for the DT and MRT, respectively. For the DT the level varies from 1% to 15%, converging to 5% for the largest window width. For the MRT the level is 14% for the smallest window width, converging to 5% for window widths larger than 350. This indicates that for this AR2 process the convergence to the χ^2-distribution for the squared periodogram values is much faster than that for the first AR2 process.

The power of detecting the change from the first AR2 model to the second is 79% for the largest window width in the DT. For the MRT the power is over 90% for window width 140 and 100% for window widths of 230 or more. The power of TASS is between 68% and 87%. This means that the power is better for our method, but again it is hard to compare directly since TASS does not take the scale into account.

Another way to compare the performance of the methods is to run an analysis on the "same" process and compare the results. Adak (1998) shows an example using the changing AR2 process described above, where the change occurs after one third of the total time. The resulting segmentation tree is given in the left part of Fig. 12. TASS only makes dyadic segmentations of the data, i.e. the splits in the data occur at dyadic points. The price that one must pay for the fast algorithm is that if a change in the spectrum occurs at a non-dyadic point, then the algorithm will try to estimate the dyadic segmentation that best approximates the true partitioning of the data. For a non-trained eye, it is hard to interpret how many segments TASS detects in this case.

The right part of Fig. 12 displays an MRT analysis of the same series. We see that there are two spurious detections, one occurring at window width 50 and one at window width 307. The true change is detected for all window widths from 141 and coarser.

We cannot say that any of the methods outperforms another since the methods are rather different and have different qualities. TASS has better observed levels, but the spurious detections of our method are usually easy to separate from the real detections. TASS does not take into account the issue of scale. If a change occurs on a large scale (slowly developing), TASS would possibly detect that there are two segments, but not pinpoint for which time span or at

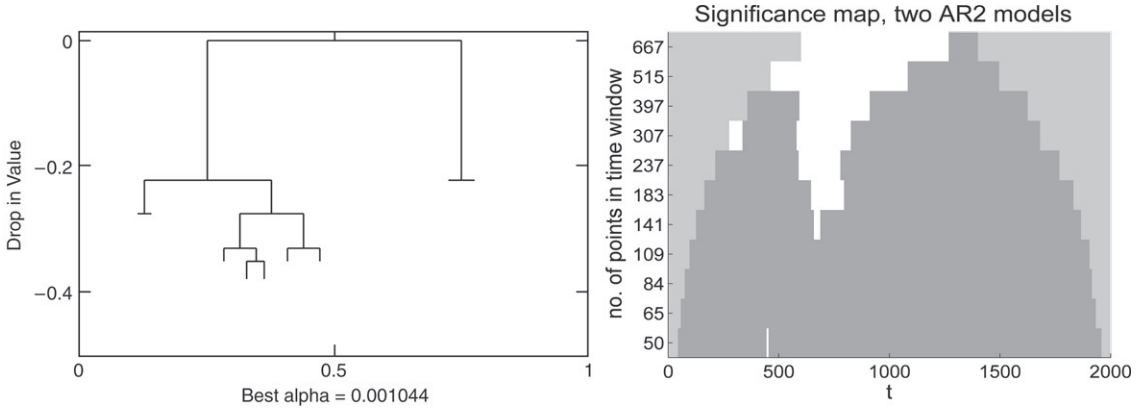

Fig. 12. TASS (left) and MRT analysis of the changing AR2 process where the changes occur at $t = 1/3$ and $t = 667$, respectively.

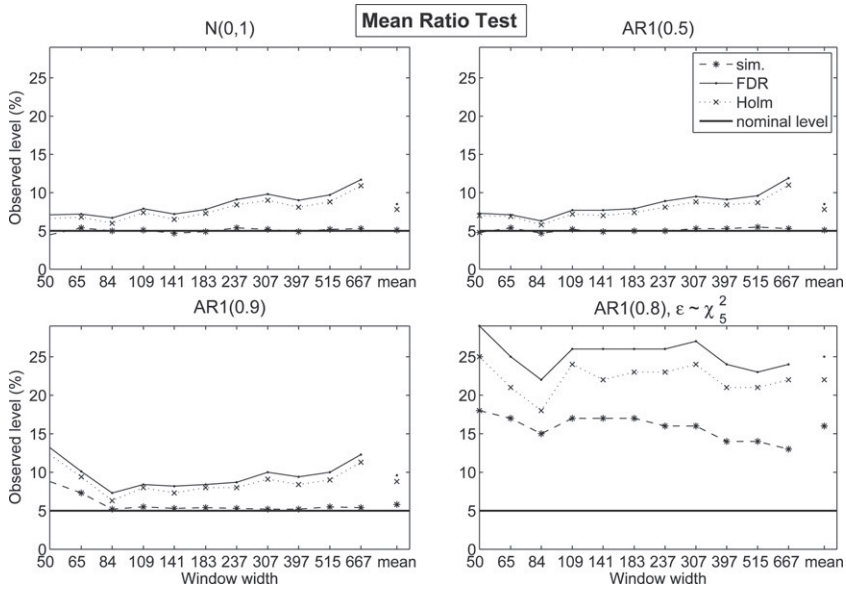

Fig. 13. Observed levels for the Mean Ratio Test.

which scale the change occurs. According to the example, we think that our method gives a clearer picture of when the change occur and in addition you get information about at which scale the change happens.

3.2. Observed significance levels

Figs. 13–16 display the observed levels for the Mean Ratio Test and the Distribution Test, respectively, for different window widths in time series simulated from eight different models. The models are given by:

$N(0, 1)$: Independent Gaussian random noise, $x_t = \varepsilon_t \sim N(0, 1)$.
AR1(0.5): $\varepsilon_t \sim N(0, 1)$, $x_t = 0.5x_{t-1} + \varepsilon_t$.
AR1(0.9): $\varepsilon_t \sim N(0, 1)$, $x_t = 0.9x_{t-1} + \varepsilon_t$.
AR1(0.8), $\varepsilon_t \sim \chi^2_5$: $x_t = 0.8x_{t-1} + \varepsilon_t$.
AR1(0.8), $\varepsilon_t \sim \chi^2_{10}$: $x_t = 0.8x_{t-1} + \varepsilon_t$.
AR1(0.8), $\varepsilon_t \sim Po(5)$: $x_t = 0.8x_{t-1} + \varepsilon_t$.
Markov 1: Markov process, one state, 100 outcomes; $p = 1/100$. That is, x_t discrete uniform $x_t \sim U[1, 100]$.
Markov 2: Markov process, one state, 100 outcomes 1–35 and 66–100: $p = 1/200$, outcomes 36–65: $p = 13/600$.

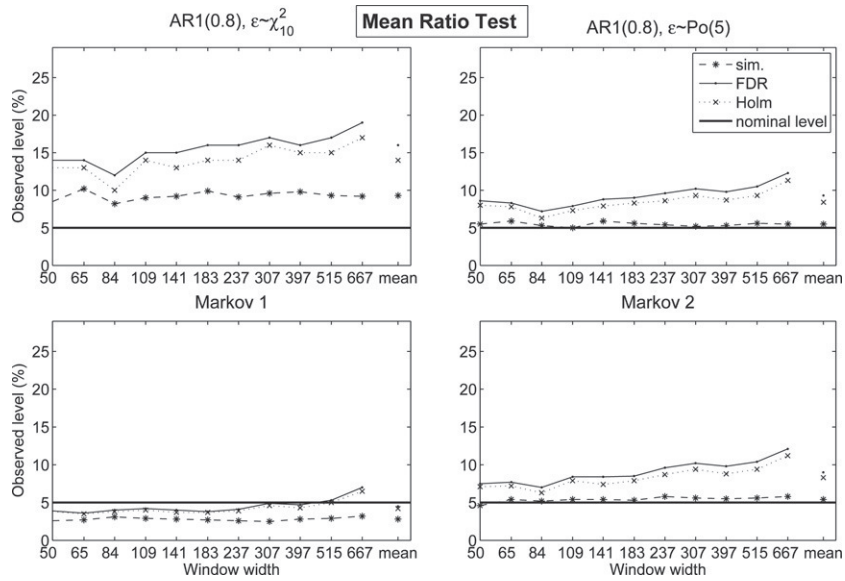

Fig. 14. Observed levels for the Mean Ratio Test.

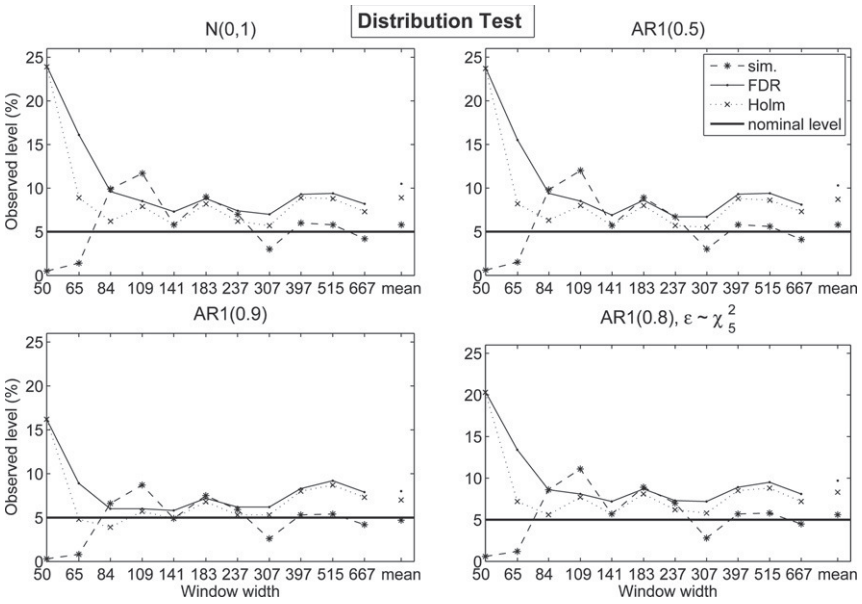

Fig. 15. Observed levels for the Distribution Test.

For all processes 10 000 realizations of the given model have been used. All figures display a comparison of the observed level using three different ways to adjust for simultaneous testing; quantiles simulated from the actual number of tests, FDR and Holm's method. From Figs. 13 and 14 we see that the observed levels for the Mean Ratio Test are lowest using simulated quantiles, and using FDR gives the highest observed levels. For all processes investigated except Markov 1, the observed levels are higher than the nominal 5% level. This means that using simulated quantiles to adjust for simultaneous testing gives an observed level closest to the nominal level. If the process is very skewed, like an AR1(0.8) with χ^2_5-distributed noise, the observed level is too high and about 16% using simulated quantiles. Also, for the same process using χ^2_{10}-distributed noise, the level is too high, but now the average level is 9% using simulated quantiles. For the Markov 1 process, the observed levels are too low. For this process, using Holm's method to adjust for simultaneous testing gives an observed level closest to the nominal level

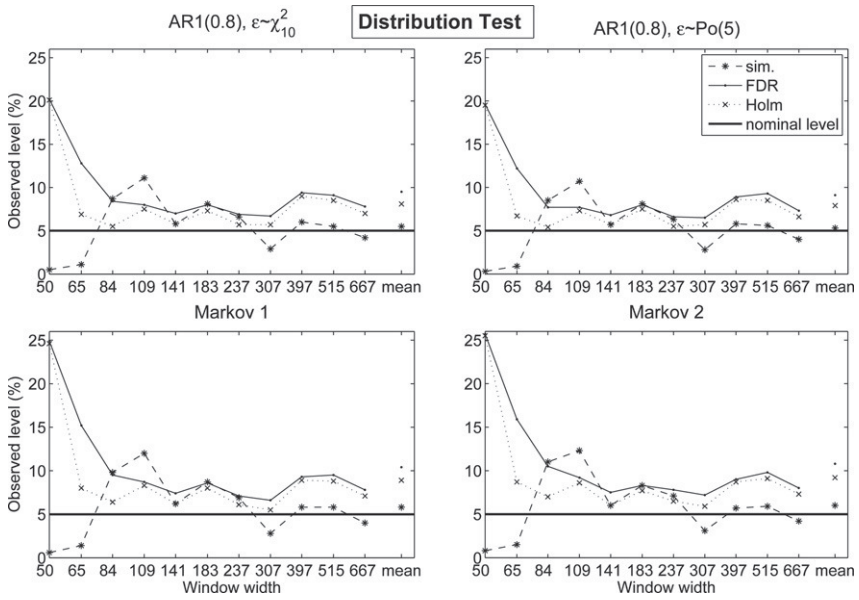

Fig. 16. Observed levels for the Distribution Test.

and using simulated quantiles gives an average observed level of 3%. We observe that using Holm or FDR, the observed levels are slightly increasing with increasing window width when the process is a Markov 1.

For the Distribution Test (Figs. 15 and 16), the observed levels vary quite a lot for different window widths, particularly for small windows. Simulated quantiles give an observed level closer to the nominal 5% level than FDR and Holm, as an average, for all distributions. This is true even if we exclude the two smallest and most unstable window widths. It should be mentioned that the levels can be very bad for data more correlated than data from an AR1(0.9) model. We also note that the curves of the observed levels look very similar for all the processes. This means that the test is rather independent of the underlying process.

For almost all processes, using simulated quantiles is the adjustment method giving a mean observed level (averaging over all given window widths) closest to the nominal level, and the mean observed level is close to the nominal 5% level for both tests. Two exceptions occur for the Mean Ratio Test: For the Markov 1 process ($U[1, 100]$) the level is too low and lower using simulated quantiles than FDR or Holm. For an AR1(0.8), $\varepsilon \sim \chi^2_5$, the mean observed level is 16%, but still closest to the nominal level using simulated quantiles. From the simulation studies we conclude that using simulated quantiles gives the most correct observed level for both the Mean Ratio Test and the Distribution Test.

3.3. Observed power

The observed power measures how often we are detecting a known change point. Figs. 17 and 18 summarize the observed power estimated from 10 000 realizations of time series where there is a change from one process to another at some point in the time series process. In each case there is exactly one change point. Four different processes are used for the simulation studies:

Two noise levels : $IID\ N(0, \sigma^2)$ where σ changes from 1 to 2.
Two AR1 processes : $x_t = \phi x_{t-1} + \varepsilon_t$, $\varepsilon_t \sim N(0, 1)$, ϕ changes from 0.1 to 0.6.
Markov 1, two states : Two-state Markov process, state 1: $p = 1/100$ for all outcomes 1–100, state 2: $p = 1/200$ for outcomes 1–35 and 66–100.
Markov 2, two states : Two-state Markov process, state 1: $p = 1/100$ for all outcomes 1–100, state 2: $p = 1/200$ for outcomes 1–25 and 76–100.

The observed power of detecting changes by the Mean Ratio Test for the distributional changes described above is shown in Fig. 17. One should be careful interpreting these results because some of the changes detected might be

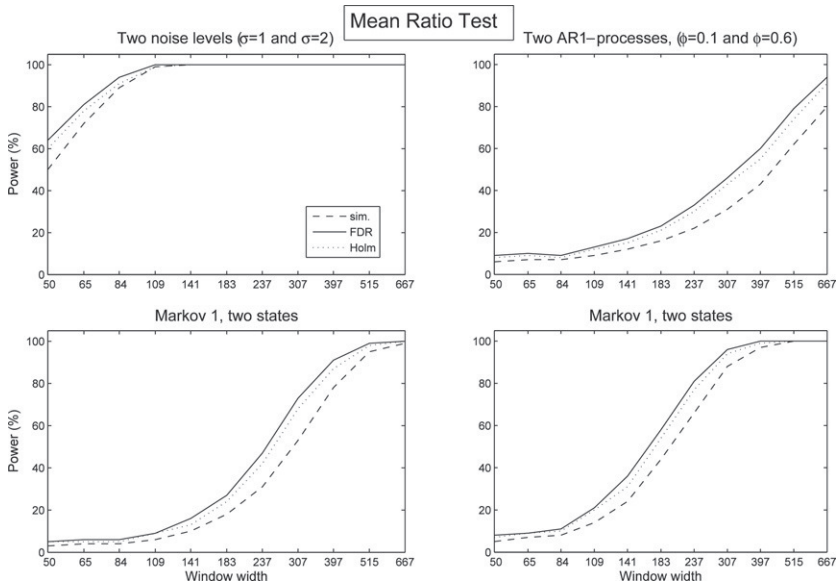

Fig. 17. Observed power for the Mean Ratio Test.

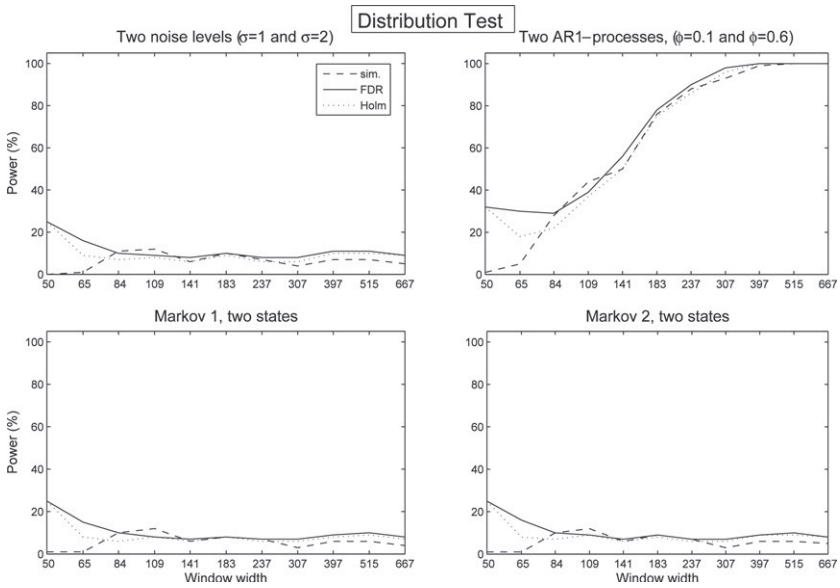

Fig. 18. Observed power for the Distribution Test.

spurious as we cannot pinpoint the exact location of the change point. Some of the pixels detected as change points might or might not be actual change points. This means that there is always a possibility of overestimating the power. The observed power is highest using FDR to adjust for simultaneous testing. This is as expected since the observed levels were highest using this adjustment. For an AR1 process where the correlation coefficient changes, the power of the test is good only for very large window widths (>500).

Fig. 18 displays the observed power for using the Distribution Test to detect change points. The Distribution Test is especially designed for AR1 processes. We see that the only process where the Distribution Test is able to detect a change is for the AR1 process. For this process, the power is clearly better using the Distribution Test than the Mean Ratio Test. FDR gives the highest power, though the differences between the three adjustment methods are small.

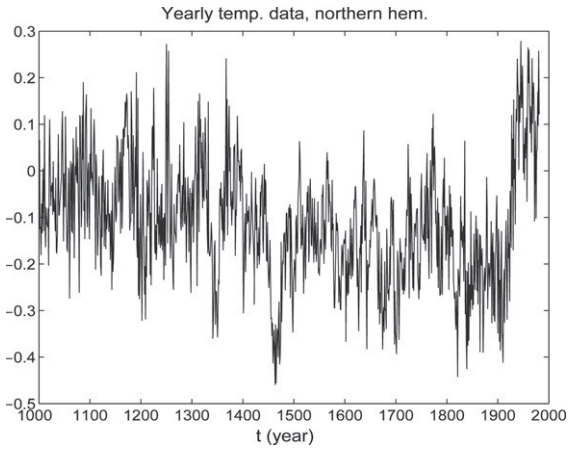

Fig. 19. Annual average temperature anomalies over the northern hemisphere for the last millennium.

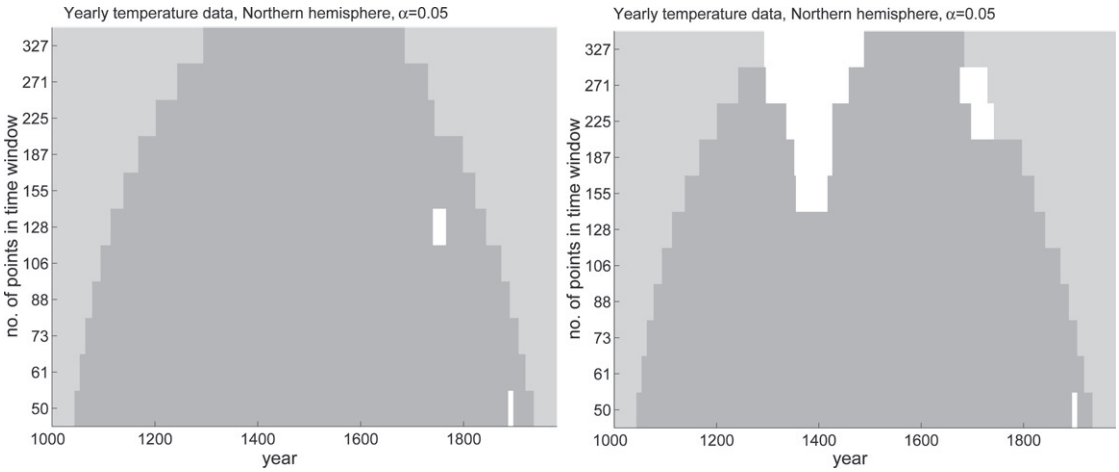

Fig. 20. Distribution (left) and Mean Test (right) analysis of the annual temperature data in Fig. 19.

The three procedures used for multiple testing seem to be comparable in terms of their observed powers, while quantiles simulated over the same number of tests have observed levels closer to the nominal levels than each of the other two procedures. For the rest of the paper, we have restricted ourselves to using the simulated quantiles in order to save computing time as well as space required for reporting the results.

3.4. Real data

3.4.1. Yearly average temperature data, northern hemisphere

Temperatures in the past can be estimated from thickness of the layers in ice cores drilled from glaciers, varves from glaciers and tree rings (Mann et al., 1999). Fig. 19 shows the development of the average annual temperature anomalies (deviations from the average temperature in 1902–1980) over the northern hemisphere during the last millennium (1000–1980). The data are reconstructed by Mann et al. (1999). In Fig. 19 the most evident change seems to be a sudden increase in the temperature around 1930. Since the mean of the time series (DC component) is excluded in our analysis methods, the methods are not suitable for detecting changes in the mean if these changes do not influence any other statistical properties of the time series. A change in the mean may, however, cause a low frequent component in the spectrum. This will often make us able to detect also a change in the mean.

Fig. 20, left and right parts, displays the Distribution and the Mean Ratio Test analyses, respectively, of the temperature data in Fig. 19. Using the smallest time scale, $N = 50$ years, both methods detect a change in 1900. Normally, we are reluctant to interpret a change that is detected only for one window width at one scale as real, but in

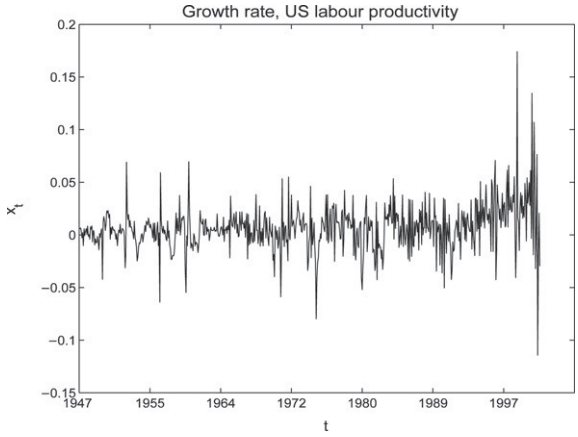

Fig. 21. Growth rate, US labour productivity.

Fig. 22. Distribution (left) and Mean Test (right) analysis of the labour productivity data in Fig. 21.

this case the two methods detect a change at the same time point and scale. Moreover, both methods detect a change in the period 1700–1750. The Mean Ratio Test detects a change in 1350–1450. Looking at Fig. 19, we see that there is a drop in the temperature, lasting from 1450 to 1470. A long period with lower temperatures known as the Little Ice Age *sensu lato* took place in the period 1450–1850 which probably means that we detect the start of this period.

The change point at $t = 1900$ is only detected for the smallest scale. This means that this change is rather abrupt. The two other change points are detected at large scales. This means either that the change is small, such that a certain size of windows is needed to be able to detect the change, or that there is a long range change slowly developing. In this case, we also need a relatively large window width to be able to detect the change. Note that none of the changes are detected for all scales. This means that if we had been analysing for a fixed window width, we would not have found all the changes.

3.4.2. Labour productivity

Fig. 21 shows the labour productivity for each month in the U.S. manufacturing/durables sector in the period 1950–2000. Looking at the productivity index, there seems to be an increase in the productivity from about 1993. Other potential changes are hard to detect by visual inspection.

Hansen (2001) finds that there are significant changes around 1963, 1982 and 1994 with confidence limits: [1959, 1971], [1977, 1988], [1992, 1996]. The Distribution Test, shown in the left part of Fig. 22, detects a change in 1970, 1981–83 and 1991. We see that the two first detections correspond relatively well to the changes found by Hansen (2001). In Hansen (2001), there is a plot of the Chow Test Sequence as a function of breakdate with the corresponding

χ^2-critical value plotted. Using this as the critical value, a change point in 1970 could be claimed and supports the change found by the Distribution Test. The χ^2-critical value is not used for the Chow Test if the change is not known a priori. For detecting an unknown change, Andrew's critical value should be used as described in Hansen (2001). Using Andrew's critical value, a change in 1970 would not be claimed.

The Mean Ratio Test, shown in the right part of Fig. 22, detects changes in 1953, 1962, 1966–70 and in 1995. The only change the Distribution and Mean Ratio Test agree about is the change around 1970. But the change in 1991 detected by the Distribution Test and in 1995 by the Mean Ratio Test could be early and late detections of the 1994 change from Hansen (2001). The change detected around 1953 is more difficult to explain. Our suggestion is that since the change is detected on three different scales, this is a real change even if it has not been detected by Hansen's methods.

Looking at the five changes we detect, we see that the change point in 1962 is only detected for small scales. This indicates that it is a short range change. The changes detected in 1953, 1966–70 and in 1999 are all detected for intermediate scales and could possibly be short (or intermediate) range changes too. Note that we stop detecting these changes when the data windows start covering more than one change. This means that one should be careful about the interpretation of these change points in terms of short and long range changes. Possibly they are changes not depending on the scale, but still we would have lost detecting some of them if we had chosen one single scale for our analysis.

4. Discussion and concluding remarks

We have demonstrated that the number and location of change points in a time series depend heavily on what scale the analysis is performed. An excellent example of this fact is given by the northern hemisphere data set in Fig. 20, where none of the changes were detected for all time scales, and all changes would not have been found if one "optimal" scale had been chosen for the analysis. Due to some technical advantages, our scale-space approach for detecting change points is performed in the frequency domain. Our default procedure is entitled the 'Mean Ratio Test', and it detects change points on short and long scales for a broad range of random processes. Since the autoregressive process of order 1 occurs frequently in practice, we have designed one method particularly for this situation. This method is entitled the 'Distribution Test', and it outperforms the Mean Ratio Test when the observed time series is of AR1 type. For other processes, our analyses of observed level and power indicate that the Mean Ratio Test is preferable.

Changes detected on small scales are more or less abrupt changes. Changes detected on large scales may be long term changes, but these are difficult to separate from abrupt, but small changes. For small abrupt changes a larger data window is needed to get enough power to detect the change. For the synthetic data examples discussed in Section 2.4, all the change points, when detected, are detected on most of the scales considered, and sometimes they are not detected at all on any scale depending on the method. For these examples, the method does not identify the change points as short and long range change points. This is expected given the way these synthetic time series data are generated. For real data analyses in Section 3.4, the situation is different from what we have seen in the case of synthetic data. In both examples, we indeed see some change points that are visible on some smaller scales, but not visible on larger scales. We also see some change points only visible on larger scales. For these data, we detect both short range and some long range change points.

Our method is based on an objective measure of whether a change is significant or not, and there is no need for any subjective choice of thresholds or limits. We do not estimate any parameters, and modelling is not a part of the objective for the method. The change points are easy to find visually as they are marked with white pixels in a grey–white plot, and there is no need for judgement about whether a change point is there or not. This is an advantage compared to Adak (1998) or SLEX (Ombao et al., 2002) where one needs some experience in interpreting the plots to be able to make judgements about how many change points there are. The main difference between our method and the traditional existing change point detection methods is that we are focusing on the importance of which (time) scale the changes are detected on. The interpretation and/or presence of change points may be very different depending on which scale they occur.

All the programming is done using Matlab. The computation time is very short since the quantiles are simulated in advance. Analysing a time series of $T = 2000$ data points takes 0.4 s, while it takes 1.2 s analysing $T = 10\,000$ data points. This is achieved using Matlab 7.3.0 on a HP Intel(R) Core(TM)2 CPU T7200 @ 2.00 GHz 997 MHz, 2.00 GB RAM.

References

Adak, S., 1998. Time-dependent spectral analysis of nonstationary time series. Journal of the American Statistical Association 93, 1488–1501.

Amin, M.G., 1987. Time and lag window selection in Wigner–Ville distributions. In: IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 1529–1532.

Andrews, D.W.K., 1993. Tests for parameter instability and structural change with unknown change point. Econometrica 61, 821–856.

Appel, U., Brandt, A.V., 1983. Adaptive sequential segmentation of piecewise stationary time series. Information Sciences 29, 27–56.

Bai, J., 1997. Estimation of a change point in multiple regression models. The Review of Economics and Statistics 4, 551–563.

Balke, N.S., 1993. Detecting level shifts in time series. Journal of Business & Economic Statistics 11, 81–92.

Basseville, M., Benveniste, A., 1983. Sequential segmentation of nonstationary digital signals using spectral analysis. Information Sciences 29, 57–73.

Benjamini, Y., Yekutieli, D., 2002. The control of the false discovery rate in multiple testing under dependency. The Annals of Statistics 29, 1165–1188.

Box, G.E.P., Jenkins, G.M., Reinsel, G.C., 1994. Time Series Analysis: Forecasting and Control. Prentice-Hall, New Jersey (Chapter 12).

Brockwell, P.J., Davis, R.A., 1991. Time Series: Theory and Methods. Springer-Verlag, New York.

Chaudhuri, P., Marron, J.S., 1999. SiZer for exploration of structures in curves. Journal of the American Statistical Association 94, 807–823.

Coates, D.S., Diggle, P.J., 1986. Tests for comparing two estimated spectral densities. Journal of Time Series Analysis 7, 7–20.

Dahlhaus, R., 1997. Fitting time series models to nonstationary processes. Annals of Statistics 25, 1–37.

Dahlhaus, R., Neumann, M.H., 2001. Locally adaptive fitting of semiparametric models to non-stationary time series. Stochastic Processes and their Applications 91, 277–308.

Dahlhaus, R., Sahm, M., 2001. Local likelihood methods for nonstationary time series and random fields. Resenhas Journal 4, 457–477.

Davis, R.A., Lee, T.C.M., Rodriguez-Yam, G.A., 2006. Structural break estimation for nonstationary time series models. Journal of the American Statistical Association 101, 223–239.

Donoho, D., Mallat, S., von Sachs R., 1998. Estimating covariances of locally stationary processes: rates of convergence of best basis methods. Technical Report 517. Stanford University, Department of Statistics.

Hansen, B.E., 2001. The new econometrics of structural change: dating breaks in US labor productivity. Journal of Economic Perspectives 15, 117–128.

Holm, S., 1979. A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics 6, 65–70.

Huang, H.-Y., Ombao, H., Stoffer, D.S., 2004. Discrimination and classification of nonstationary time series using the SLEX model. Journal of the American Statistical Association 99, 763–774.

Lai, T.L., 1995. Sequential changepoint detection in quality control and dynamical systems. Journal of the Royal Statistical Society, Series B 57, 613–658.

Ligges, U., Weihs, C., Hasse-Becker, P., 2002. Detection of locally stationary segments in time series. Technical Report 11/2002. Department of Statistics, University of Dortmund, Germany.

Lindeberg, T., 1994. Scale-space theory: A basic tool for analysing structures at different scales. Journal of Applied Statistics 21, 224–270.

Mallat, S., Papanicolaou, G., Zhang, Z., 1998. Adaptive covariance estimation of locally stationary processes. The Annals of Statistics 26, 1–47.

Mann, M.E., Bradley, R.S., Hughes, M.K., 1999. Northern hemisphere temperatures during the past millennium: inferences, uncertainties and limitations. Geophysical Research Letters 26, 759–762.

Øigård, T.A., Rue, H, Godtliebsen, F., 2006. Bayesian multiscale analysis for time series data. Computational Statistics & Data Analysis 51, 1719–1730.

Olsen, L.R., Sørbye, S.H., Godtliebsen, F., 2007. A scale space approach for detecting changes in statistical behaviour of dependent data. Scandinavian Journal of Statistics (in press).

Ombao, B.C., Raz, J.A., von Sachs, R., Guo, W., 2002. The SLEX model of a non-stationary random process. Annals of the Institute of Mathematical Statistics 54, 171–200.

Ombao, B.C., Raz, J.A., von Sachs, R., Malow, B.A., 2001. Automatic statistical analysis of bivariate nonstationary time series. Journal of the American Statistical Association 96, 543–560.

Park, C., Marron, J.S., Rondonotti, V., 2004. Dependent SiZer: goodness-of-fit tests for time series models. Journal of Applied Statistics 31, 999–1017.

Park, C., Godtliebsen, F., Taqqu, M., Stoev, S., Marron, J.S., 2007. Visualization and inference based on wavelet coefficients, SiZer and SiNos. Computational Statistics & Data Analysis 51, 5994–6012.

Percival, D.B., Walden, A.T., 2000. Wavelet Methods for Time Series Analysis. Cambridge University Press, 620 pp.

Polansky, A., 2007. Detecting change-points in Markov chains. Computational Statistics & Data Analysis 51, 6013–6026.

Priestley, M., 1965. Evolutionary spectra and non-stationary processes. Journal of the Royal Statistical Society 27, 204–237.

Sarkar, S.K., Chang, C.-K., 1997. The simes method for multiple hypothesis testing with positively dependent test statistics. Journal of the American Statistical Association 92, 1601–1608.

Sato, J.R., Morettin, P.A.M., Arantes, P.R., Amaro, E., 2007. Wavelet based time-varying vector autoregressive modelling. Computational Statistics & Data Analysis 51, 5847–5866.

Sclove, S.L., 1983. Time series segmentation: a model and a method. Information Sciences 29, 7–25.

Shumway, R.H., Stoffer, D.S., 2000. Time Series Analysis and Its Applications. Springer-Verlag, New York.

Simes, R.J., 1986. An improved Bonferroni procedure for multiple tests of significance. Biometrika 73, 751–754.

ter Haar Romeny, B.M., 2001. Front-end Vision and Multiscale Image Analysis. Kluwer Academic Publishers, Dordrecht, Netherlands.

von Sachs, R., Neumann, M.H., 2000. A wavelet based test for stationarity. Journal of Time Series Analysis 21, 597–613.