Multiscale breakpoint detection in piecewise stationary AR models

Haeran Cho Piotr Fryzlewicz [1]

[1] University of Bristol, UK. E-mail: {haeran.cho, p.z.fryzlewicz}@bris.ac.uk

Keywords: piecewise stationarity, AR, breakpoint detection, multiscale

1 Introduction

In this paper, we are interested in the problem of detecting breakpoints in piecewise stationary autoregressive (AR) processes. For short time series, stationarity assumption is common. However, for longer time series, this assumption is often unrealistic. Besides, many naturally occurring phenomena cannot be modelled as stationary processes. Therefore modelling stochastic time series under nonstationary assumption is often appealing, and finds its application in many areas, such as speech processing, biomedical signal processing, and seismology to name a few.

Much literature exists on various models for nonstationary time series, such as Priestley (1965), Dahlhaus (1997) (locally stationary processes), Adak (1998) (piecewise locally stationary processes), Nason et al. (2000) (locally stationary wavelet processes), and Ombao et al. (2002) (the SLEX model).

In the analysis of nonstationary time series, it is common practice to assume certain conditions on the time-varying structure of time series, so that meaningful estimation is attainable. The simplest approach is to assume that the time series is, or can be well approximated by, a piecewise stationary process (Adak (1998)). Several methods have been suggested for segmenting such nonstationary time series by adapting existing on-line or a posteriori segmentation procedures. In this paper, we are interested in a posteriori segmentation of given time series with all the data available from the past which enables the prediction of future based on the last stationary piece.

One way of segmenting nonstationary time series is to develop test statistics based on the spectral representation of given time series. In Ishii et al. (1979), the statistics measuring the difference between two spectral densities were proposed, such as Kullback information, Bhattacharyya distance, and Chernoff distance, so that it is possible to detect any change in the amplitude or frequency for the observed time series. Ombao et al. (2001) introduced the Auto-SLEX method which uses the aforementioned SLEX transform to obtain the SLEX periodogram and chooses the "best" segmentation minimising the cost of estimated periodogram. Adak (1998) proposed the Tree-based Adaptive Segmented Spectrogram algorithm (TASS), where a complete tree is constructed by recursively halving each segment of the data into very short segments, and the distance between the estimated spectra from the left and right halves of the segment is computed.

On the other hand, by assuming a parametric model, some papers focused on detecting the change in parameters. In the context of multiple regression problem, Bai (1997) studied the least squares estimation of the change in regressors. In Gombay (2008), testing statistics were developed which can detect the change in any one of $(p+2)$ parameters of an AR(p) time series, including the mean, the variance of white noise, and p AR parameters. Instead of using the statistics process of quadratic form created from the estimated white noise sequence, they suggested the test statistics based on the efficient score vector. Under such a parametric model assumption, Lavielle (1998) addressed the off-line nonstationary process segmentation problem by introducing a new random process that takes the value 1 at the change instants and zero between two changes. Without any additional prior information, this change process is defined as a sequence of Bernoulli variables and estimated as the maximum a posteriori (MAP) estimate based on a penalized contrast function.

Davis et al. (2006) proposed the Auto-PARM procedure which adopted the minimum description length (MDL) principle. Under the assumption that the time series is piecewise autoregressive, the Auto-PARM finds the "best" combination of the number of breakpoints, their locations and AR orders of all stationary segments by minimising the code length which is computed according to general coding rules. Also in the paper, it was recalled that since locally piecewise stationary processes are well approximated by piecewise stationary processes (Adak (1998)), locally stationary time series can also be approximated by piecewise stationary processes. Besides, any zero-mean nondeterministic stationary process can be expressed as a

sum of two uncorrelated processes, an infinite order moving-average process and a deterministic process (Brockwell and Davis (1996)), which implies that AR processes can be used to approximate the weakly stationary processes. Thus segmentation of piecewise stationary AR processes can be extended to the approximation of locally stationary time series.
The methodology presented in this paper is developed primarily for a posteriori breakpoint detection in piecewise stationary AR processes and therefore can be used in approximating locally stationary time series. It proceeds in two stages; Firstly the given process X_t is transformed into $(p+1)$ sequences which are "pre-estimates" for the autocovariance function at lags $0, 1, \cdots, p$. Then based on each pre-estimate sequence, we look for a piecewise constant function which has the minimum total variation, among the piecewise constant functions with sensibly behaving estimated residuals. The breakpoints in the estimated functions indicate the breakpoints in given time series process. It is shown later that our estimated breakpoints are likely to be within a certain distance of true breakpoints, which converges to zero with probability in rescaled time. Once we locate the breakpoints in sequences, and the distance between two breakpoints is sufficiently large, standard parameter estimation methods, such as maximum likelihood estimation or least squares estimation, can be applied to each segment.
This paper is organised as follows. In Section 2, we describe the basic setting for models of our interest. Transformation and estimation procedures are developed and presented in Section 3 and Section 4, with the description of their theoretical properties and computational implementation. We compare our methodology with others through simulations of varying settings in Section 5.

2 Piecewise stationary AR models

For a piecewise stationary AR(p) process X_t, let m be the number of breakpoints, η_j, $j = 1, \cdots, m$ be the jth breakpoints, $\eta_0 = 1$, and $\eta_{m+1} = n$. Then the process can be written as

$$X_t = \begin{cases} X_t^{(1)}, & \text{for } \eta_0 = 1 \le t \le \eta_1 \\ X_t^{(2)}, & \text{for } \eta_1 + 1 \le t \le \eta_2 \\ \vdots & \\ X_t^{(m+1)}, & \text{for } \eta_m + 1 \le t \le \eta_{m+1} = n \end{cases} \tag{2.1}$$

where each $X_t^{(j)}$is an AR process of order p, i.e.,

$$X_t^{(j)} = \beta_{t,1} X_{t-1}^{(j)} + \beta_{t,2} X_{t-2}^{(j)} + \cdots + \beta_{t,p} X_{t-p}^{(j)} + \epsilon_t. \tag{2.2}$$

$\beta_{t,i}$'s are piecewise constant and at each η_j, there exists at least one parameter function $\beta_{t,i}$ which has a "jump" in its value. The orders of segments do not have to be the same throughout the time but by assuming that p is the maximum order, we can exploit the notation above. By defining that $p = \max_{1 \le j \le m+1} p_j$, there is no problem in applying standard parameter estimation procedures such as the maximum likelihood estimation or least squares estimation within each stationary segment (Brockwell and Davis (1996)). Note that it is reasonable to assume that the breakpoints in $\beta_{t,i}$ are sufficiently sparse over time and the length of each stationary segment goes to infinity as the number of observations grows for the estimation to be attainable; see, for example, Ombao et al. (2001) for further discussion. Condition on this distance will be given later in Assumption 4.1.
Finally, the noise $\{\epsilon_t,\ t = 1, 2, \cdots\}$ is assumed to be a zero-mean random variable with variance σ_ϵ^2. More detailed assumption on the distribution of ϵ_t is discussed in Proposition 3.1.

3 Process transformation

Denote the time-varying autocovariance function of X_t as $\gamma_t(\cdot)$, i.e.,

$$\begin{aligned} \gamma_t(r) &= \mathbb{E}\left(X_t^{(j)} X_{t-r}^{(j)}\right), \\ \text{for } t &= \eta_{j-1} + r + 1, \cdots, \eta_j;\ r = 0, \cdots, p;\ j = 1, \cdots, m+1 \end{aligned}$$

so that $\gamma_t(\cdot)$ changes in time in piecewise constant manner. Since the AR parameters $\beta_{t,1}, \cdots, \beta_{t,p}$ satisfy

$$\begin{pmatrix} \gamma_t(0) & \gamma_t(1) & \cdots & \gamma_t(p-1) \\ \gamma_t(1) & \gamma_t(0) & \cdots & \gamma_t(p-2) \\ \vdots & & & \vdots \\ \gamma_t(p-1) & \gamma_t(p-2) & \cdots & \gamma_t(0) \end{pmatrix} \begin{pmatrix} \beta_{t,1} \\ \beta_{t,2} \\ \vdots \\ \beta_{t,p} \end{pmatrix} = \begin{pmatrix} \gamma_t(1) \\ \gamma_t(2) \\ \vdots \\ \gamma_t(p) \end{pmatrix}, \tag{3.3}$$

within each stationary segment of X_t, any break in $\beta_{t,i}$ for $i = 1, \cdots, p$ is reflected as a break in at least one of $\gamma_t(r)$ for $r = 0, \cdots, p$. Therefore our pre-estimate sequences are the localised estimates of $\gamma_t(r)$ at each t, i.e.,

$$\tilde{\gamma}_{t,r} = X_t X_{t-r} = \mathbb{E}\left(X_t X_{t-r}\right) + \left\{X_t X_{t-r} - \mathbb{E}\left(X_t X_{t-r}\right)\right\}, \ r = 0, \cdots, p. \tag{3.4}$$

Note that $\gamma_t(r)$ and $\mathbb{E}\left(X_t X_{t-r}\right)$ are not always equivalent since for $t = \eta_j + 1, \cdots, \eta_j + r; j = 1, \cdots, m; r = 1, \cdots, p$, $\mathbb{E}\left(X_t X_{t-r}\right) = 0$. This "boundary" effect will be addressed later in Section 4.
Let $v_{t,r} := X_t X_{t-r} - \mathbb{E}\left(X_t X_{t-r}\right)$ then within each stationary segment of X_t, $v_{t,r}$ is a zero-mean, strong mixing (α-mixing) sequence with exponentially decaying mixing coefficients $\{\alpha(k), k = 1, 2, \cdots\}$ (Davidson (1994)). We further assume that following condition is satisfied for v_t.

Proposition 3.1. *$v_{t,r}$ satisfies the Cramèr's condition, i.e., there exists a positive constant c satisfying*

$$\mathbb{E}|v_{t,r}|^k \leq c^{k-2} k! \mathbb{E} v_{t,r}^2 \tag{3.5}$$

for all $k = 3, 4, \cdots$.

For instance, above condition is satisfied when ϵ_t follows Gaussian distribution. With this proposition, we can establish the following property.

Lemma 3.1. *Denote $\mathcal{I}$ as a family of subsets of $\{1, \cdots, n\}$ and let σ be the standard deviation of $v_{t,r}$ within a stationary segment. Then there exists a positive constant τ which satisfies*

$$\Pr\left(\max_{I \in \mathcal{I}} \frac{1}{\sqrt{|I|}} \left| \sum_{t \in I} v_{t,r} \right| \leq \sigma \sqrt{\tau \log n} \right) \longrightarrow 0 \tag{3.6}$$

for each $r = 0, 1, \cdots, p$ within each stationary segment of X_t.

Lemma 3.1 implies that all partial sums (including "multiscale", wavelet-like ones) of $v_{t,r}$ can be bounded by a term of order $O(\log n)$, like Gaussian distributed variables. This property is used as a criterion in the estimation procedure to decide whether the estimated residuals are "Gaussian-like" or not. The rate of convergence is of polynomial order.

4 Estimation procedure

In this section, we develop an estimation procedure generating a piecewise constant function for each pre-estimate sequence with breakpoints, which indicate the structural breaks in X_t.
Denote

$$\mathcal{A}_n = \mathcal{A}_n(\tilde{\gamma}_r, \mathcal{I}, \sigma, \tau) = \{g : \max_{I \in \mathcal{I}} |w(\tilde{\gamma}_r, g, I)| \leq \sigma \sqrt{\tau \log n}\}, \tag{4.7}$$

$$w(\tilde{\gamma}_r, g, I) = \frac{1}{\sqrt{|I|}} \sum_{t \in I} (\tilde{\gamma}_{t,r} - g_t), \tag{4.8}$$

Since $v_{t,r}$ is not stationary throughout the time, neither is the scale σ and its estimation is discussed later in this section.
As $\gamma_t(r)$ is piecewise constant and its breakpoints are sparse, we choose the estimate of $\gamma_t(r)$, say $\hat{\gamma}_{t,r}$, among the piecewise constant sequences which have multiresolution sums of estimated residuals bounded by $\sigma\sqrt{\tau \log n}$, as $v_{t,r}$ in Lemma 3.1. Since what we are looking for is the estimate with sparse breakpoints, we adopt the commonly used practice in linear regression problems with sparse solutions/near-solutions. Donoho (2006) shows that if there exists any sufficiently sparse (near-)solution for a given linear regression

problem, the solution with minimum l_1-norm is a good approximation in the l_2 sense. Besides, while the procedure of looking for a solution with minimum l_0-norm is intractable, minimising l_1-norm changes the problem into a computationally feasible convex problem. Therefore, instead of minimising l_0-norm of $\hat{\gamma}_{t+1,r} - \hat{\gamma}_{t,r}$ directly, our estimation procedure minimises the total variation of $\hat{\gamma}_{t,r}$, which is the same as minimising the l_1-norm.
In short, we find a function $\hat{\gamma}_{t,r} \in \mathcal{A}_n$ which has the minimum total variation

$$\sum_{t=1}^{n-1} |\hat{\gamma}_{t+1,r} - \hat{\gamma}_{t,r}| \,. \tag{4.9}$$

Since $v_{t,r}$ is piecewise stationary throughout $t = 1, \cdots, n$, the variance of $v_{t,r}$ is not constant but changes over time in piecewise constant manner as $\gamma_t(r)$ does. One way of estimating $\sigma_t := \text{var}(v_{t,r})$ is to modify the method presented in Davies et al. (2008).
For sufficiently large constant $C > 1$, let $N = \lfloor n/C \log n \rfloor$, $\bigcup_{j=1}^{N} J_j = \{1, \cdots, n\}$ where J_j's are disjoint and of about equal size, for example, $J_1 = \{1, \cdots, [C \log n]\}$ and so on. Then

$$\hat{\sigma}_{t,r}^{(1)} = 1.4826 \cdot \text{median}\{|\tilde{\gamma}_{j_2,r} - \tilde{\gamma}_{j_1,r}|, \cdots, |\tilde{\gamma}_{j_{N_k},r} - \tilde{\gamma}_{j_{N_k-1},r}|\}/\sqrt{2} \quad \text{for } t \in J_k = \{j_1, \cdots, j_{N_k}\}. \tag{4.10}$$

One the other hand, we can estimate the time-varying $\sigma_t^2 = \text{var}(v_{t,r}) = \text{var}(\tilde{\gamma}_{t,r})$ as a straightforward sample variance over a fixed window. Let b be a fixed integer indicating the window size. Then,

$$\hat{\sigma}_{t,r}^{(2)} = \left\{ \frac{1}{2b} \sum_{j=-b}^{b} \left(\tilde{\gamma}_{t+j,r} - \overline{\tilde{\gamma}_{t,r}}\right)^2 \right\}^{1/2}, \text{ where } \overline{\tilde{\gamma}_{t,r}} = \frac{1}{2b+1} \sum_{j=-b}^{b} \tilde{\gamma}_{t+j,r}. \tag{4.11}$$

In Section 5, the results were obtained using the latter estimate (4.11) for σ_t, as it seemed to give better performance than (4.10).

4.1 Theoretical performance of the methodology

We now show that the estimated function $\hat{\gamma}_{t,r}$ from above method can detect the breakpoints in $\gamma_t(r)$ with high probability. First, we need an assumption on the distance between two breakpoints.

Assumption 4.1. *The length of each segment between two breakpoints, say $d := d_n$, satisfies*

$$\log n / d_n \longrightarrow 0.$$

With this assumption, Theorem 4.1 is derived below.

Theorem 4.1. *Let $\gamma_t(r)$ has a breakpoint at time $t = z$. Then estimated piecewise constant function $\hat{\gamma}_{t,r}$ has a breakpoint at $\hat{z}$ which satisfies*

$$\Pr\left(|\hat{z} - z| \leq O(\log n)\right) \longrightarrow 1.$$

Besides, let z_1 and z_2 be two adjacent breakpoints in $\gamma_t(r)$. Then

$$\Pr\left(TV_{(z_1 + c_1 \log n, z_2 - c_2 \log n)}(\hat{\gamma}_{t,r}) \leq o(1)\right) \longrightarrow 1$$

for some positive constants c_1 and c_2. The rate of convergence is the same as in Lemma 3.1.

Remark 4.1. *In rescaled time interval $[0, 1]$, as in other nonparametric estimation literature, let $s = t/n$, $u = z/n$, and $\hat{u} = \hat{z}/n$. Then above result can be rewritten as*

$$\Pr\left(|\hat{u} - u| \leq O\left(\frac{\log n}{n}\right)\right) \longrightarrow 1,$$

and

$$\Pr\left(TV_{(u_1 + c_1 \frac{\log n}{n}, u_2 - c_2 \frac{\log n}{n})}(\hat{\gamma}_{s,r}) \leq o(1)\right) \longrightarrow 1.$$

As mentioned earlier, $\gamma_t(r) \neq \mathbb{E}\tilde{\gamma}_{t,r}$ for $t = \eta_j + 1, \cdots, \eta_j + r; j = 1, \cdots, m; r = 1, \cdots, p$. However, it can be shown that this boundary effect does not have influence on the results in Theorem 4.1 due to Assumption 4.1 and therefore it is theoretically ignorable.

4.2 Implementation of the estimation algorithm

Throughout this section, we drop r from the notation when there is no confusion. We can rewrite (4.9) as

$$||\hat{\gamma}_t||_1 = \sum_{t=1}^{n} |\hat{\gamma}_t - \hat{\gamma}_{t-1}| = \sum_{t=1}^{n} x_t^+ + x_t^-$$

where $\hat{\gamma}_0 = 0$, $x_t = \hat{\gamma}_t - \hat{\gamma}_{t-1} = x_t^+ - x_t^-$ and $x_t^+ = \max\{0, x_t\}$, $x_t^- = \max\{0, -x_t\}$. Then

$$\begin{pmatrix} \hat{\gamma}_1 \\ \hat{\gamma}_2 \\ \vdots \\ \hat{\gamma}_n \end{pmatrix} = \begin{pmatrix} 1 & 0 & \cdots & 0 & -1 & 0 & \cdots & 0 \\ 1 & 1 & \cdots & 0 & -1 & -1 & \cdots & 0 \\ \vdots & & & \vdots & \vdots & & & \vdots \\ 1 & 1 & \cdots & 1 & -1 & -1 & \cdots & -1 \end{pmatrix} \begin{pmatrix} x_1^+ \\ x_2^+ \\ \vdots \\ x_n^+ \\ x_1^- \\ x_2^- \\ \vdots \\ x_n^- \end{pmatrix},$$

$$\text{say, } \hat{\gamma} = (\mathbf{K}, -\mathbf{K})\mathbf{x}.$$

Let $\mathbf{L} := (\mathbf{K}, -\mathbf{K})$ and $\mathbf{K}'$ be column-wise normalised $\mathbf{K}$, i.e.,

$$\mathbf{K}' := \begin{pmatrix} 1/\sqrt{n} & 0 & \cdots & 0 \\ 1/\sqrt{n} & 1/\sqrt{n-1} & \cdots & 0 \\ \vdots & & & \vdots \\ 1/\sqrt{n} & 1/\sqrt{n-1} & \cdots & 1 \end{pmatrix}.$$

The condition of $\hat{\gamma}_t$ in (4.7) and (4.8) can be rewritten as

$$\left|\mathbf{K}'^T(\tilde{\gamma} - \mathbf{L}\mathbf{x})\right| \leq \sigma\sqrt{\tau \log n}, \tag{4.12}$$

where $\tilde{\gamma} = (\hat{\gamma}_1, \cdots, \hat{\gamma}_n)^T$ and the inequality applies coordinate-wise. Then the problem becomes minimising l_1-norm of $\mathbf{x}$ subject to the inequality constraints (4.12) and can be solved by means of standard linear programming techniques, see for example, Boyd and Vandenberghe (2004). This algorithm does not cover multiresolution sums over all subsets of $\mathcal{I}$, but normalised columns of $\mathbf{K}'$ naturally compute the sums of residuals over selected supports $I \in \{\{1, \cdots, n\}, \{2, \cdots, n\}, \cdots, \{n\}\} \subset \mathcal{I}$. We obtain the desired solution $\hat{\gamma}$ by computing $\mathbf{L}\mathbf{x}$.

Remark 4.2. *Note that the constraint in (4.12) can be interpreted as a bound over transformation-invariant residuals in a linear regression problem $\tilde{\gamma} \cong \mathbf{L}\mathbf{x}$. Essentially this procedure works similarly to the Dantzig selector (Candès and Tao (2007)), since both are seeking for a sparse solution with a minimum l_1-norm constraint and bounded estimated residuals. However $\mathbf{K}$, the design matrix equivalent in our problem, does not satisfy the "uniform uncertainty principle" conditions, which require that any submatrix with S or fewer columns of $\mathbf{K}$ should behave as if it were almost orthogonal, where S denotes the sparsity. Even though $\mathbf{K}$ does not meet these requirements, nonzero entries in our solution indicate where breakpoints occur and highly correlated columns of $\mathbf{K}$ are adjacent to each other. Furthermore, $\mathbf{K}$ is a n-by-n matrix, which means that the number of covariates and the number of observations are the same, while the Dantzig selector also covers the cases when the number of covariates is larger than the number of observations.*

Remark 4.3. *There remains the choice of τ. The choice of $\tau = 2$ can be also justified by the similarity between our procedure and the Dantzig selector. See also Davies and Kovac (2001), where their choice of τ is 2.5. In Section 5, grid of values were used for τ and with breakpoints obtained from each value, the time series was fitted and the sum of squared residuals (RSS) were computed so that τ with minimum RSS was chosen.*

4.3 Post-processing

It is always of interest to tell which estimated breakpoints are significant and which ones are from either the l_1-norm constraint or irregular fluctuations in pre-estimate sequences/estimated scale σ. In this section, we introduce a post-processing procedure to serve this purpose by adapting the arguments in the proof of Theorem 4.1. r is dropped from the subscript throughout this section.
For an estimated breakpoint, say ν, let

$$\hat{\gamma}_1 := \hat{\gamma}_{\nu-}, \quad \hat{\gamma}_2 := \hat{\gamma}_{\nu+}, \quad \gamma_1 := \gamma_{\nu-}, \quad \gamma_2 := \gamma_{\nu+}, \quad \sigma_1 := \sigma_{\nu-}, \quad \sigma_2 := \sigma_{\nu+}$$

and denote the lengths of two estimated segments before and after ν as d_1 and d_2. Then it follows that

$$\begin{aligned} \hat{\gamma}_1 - \frac{2\sigma_1\sqrt{\tau \log n}}{\sqrt{d_1}} &\leq \gamma_1 \leq \hat{\gamma}_1 + \frac{2\sigma_1\sqrt{\tau \log n}}{\sqrt{d_1}}, \\ \hat{\gamma}_2 - \frac{2\sigma_2\sqrt{\tau \log n}}{\sqrt{d_2}} &\leq \gamma_2 \leq \hat{\gamma}_2 + \frac{2\sigma_2\sqrt{\tau \log n}}{\sqrt{d_2}}, \end{aligned}$$

with probability converging to 1, and therefore

$$|\gamma_1 - \gamma_2| \leq |\hat{\gamma}_1 - \hat{\gamma}_2| + 2\sqrt{\tau \log n}\left(\frac{\sigma_1}{\sqrt{d_1}} + \frac{\sigma_2}{\sqrt{d_2}}\right). \tag{4.13}$$

As pointed out in the proof of Theorem 4.1, the distance between the true breakpoint and the estimated breakpoint is bounded from below in the worst case as

$$k \geq \frac{4(\sigma_1 + \sigma_2)^2 \tau \log n}{(\gamma_2 - \gamma_1)^2} \tag{4.14}$$

From (4.13), the right-hand side of (4.14) can be replaced by

$$k \geq \hat{k} := \frac{4(\sigma_1 + \sigma_2)^2 \tau \log n}{\left\{|\hat{\gamma}_1 - \hat{\gamma}_2| + 2\sqrt{\tau \log n}\left(\sigma_1/\sqrt{d_1} + \sigma_2/\sqrt{d_2}\right)\right\}^2}.$$

Then for each estimated breakpoint ν, we can compute the intervals $[\nu - \hat{k}, \nu + \hat{k}]$. If an estimated breakpoint is included in either of adjacent intervals apart from its own, it is likely that the corresponding estimated breakpoints are actually of the same true breakpoint. Thus our post-processing procedure computes these intervals for all estimated breakpoints and merge such two breakpoints into the one with narrower interval. This procedure is conducted repeatedly until every breakpoint is located outside the intervals other than its own.

5 Simulation study

As the implementation of the methodology is still in progress, we only present simulation results for illustrative purposes.
For the simulation experiment, the piecewise stationary AR(2) process with breakpoints in (5.15) was repeatedly generated 50 times.

$$X_t = \begin{cases} 0.4X_{t-1} - 0.6X_{t-2} + \epsilon_t, & 1 \leq t \leq 150 \\ -0.2X_{t-1} + \epsilon_t, & 151 \leq t \leq 300 \\ 0.5X_{t-1} + \epsilon_t, & 301 \leq t \leq 450 \end{cases} \tag{5.15}$$

Assuming the order of the process was known, we applied the breakpoint detection procedure with $\hat{\sigma}_t$ estimated as in (4.11). The $\hat{\sigma}$ estimated as (4.11) seemed to work better than (4.10) since the former was more likely to adjust to the fluctuation of the pre-estimate sequences than the latter. For τ, as mentioned in the Remark 4.3, sequence $2, 3, \cdots, 6$ was given as candidates and the one with minimum RSS was selected.
A typical realisation of the time series and the estimation result is given in Figure 1 (a). Vertical dotted lines indicate the true breakpoint (at $t = 150,\ 300$) and thicker dashed lines indicate the detected breakpoint with our procedure. In Figure 1 (b) and (c), pre-estimate sequence of lags 1, 2 and the estimated functions after post-processing (thicker) of time series in (a) are shown.

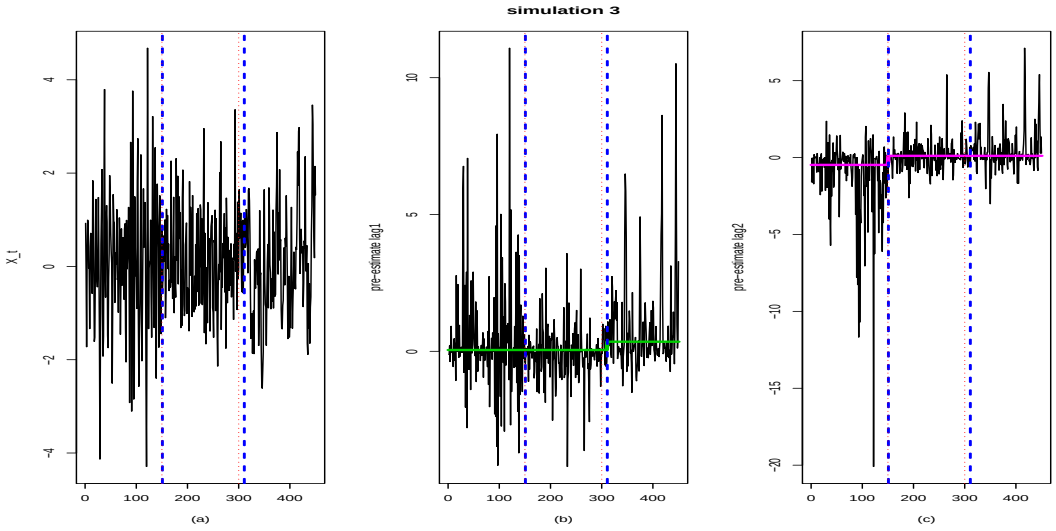

Figure 1: (a) An example of process (5.15) and detected breakpoints; (b) $\tilde{\gamma}_{t,1}$ and $\hat{\gamma}_{t,1}$; (c) $\tilde{\gamma}_{t,2}$ and $\hat{\gamma}_{t,2}$

Table 1: Summary of the estimated breakpoints from the process (5.15)

Number of segments	Breakpoints (%)	Distance from true breakpoint	Breakpoints (%)	
			$t = 150$	$t = 300$
0	0	$[\log(n)] = 6$	56	26
1	0	$[2\log(n)] = 12$	70	42
2	88	$[3\log(n)] = 18$	76	58
3	12	$[4\log(n)] = 24$	78	74
> 3	0	$[5\log(n)] = 30$	86	76
All	100			

The results of the experiment is summarised in Table 1. The number of breakpoints detected and the distance between the detected and true breakpoints are shown in the table.

In most cases, our procedure accurately estimated the number of breakpoints but sometimes overestimation was made with one more breakpoint. However, as shown in (3.3), any change in $\beta_{t,i}$, $i = 1, \cdots, p$ is reflected as a change in at least one of $\gamma_t(r)$ sequences for $r = 0, \cdots, p$. Therefore our method, if it happens at all, will tend to oversplit rather than undersplit. In some cases, oversplitting may be preferable over undersplitting as the latter can introduce nonstationarity bias (Ombao et al. (2001)).

References

Adak, S. (1998). Time-dependent spectral analysis of nonstationary time series. *J. Am. Stat. Assoc.*, 93(444):1488–1501.

Bai, J. (1997). Estimation of a change point in multiple regression models. *Rev. Econ. Stat.*, 79(4):551–563.

Bosq, D. (1998). *Nonparametric statistics for stochastic process: estimation and prediction*. Springer.

Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press.

Brockwell, P. J. and Davis, R. A. (1996). *Time series: theory and methods*. Springer.

Candès, E. and Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *Ann. Statist.*, 35(6):2313–2351.

Dahlhaus, R. (1997). Fitting time series models to nonstationary processes. *Ann. Statist.*, 25(1):1–37.

Davidson, J. (1994). *Stochastic limit theory: An introduction for econometricians*. Oxford University Press.

Davies, P. L. and Kovac, A. (2001). Local extremes, runs, strings and multiresolution. *Ann. Statist.*, 29(1):1–48.

Davies, P. L., Kovac, A. and Meise, M. (2008). Confidence regions, regularization and non-parametric regression. *Ann. Statist. To appear.*

Davis, R. A., Lee, T. C. M. and Rodriguez-Yam, G. A. (2006). Structural break estimation for non-stationary time series. *J. Am. Stat. Assoc.*, 101(473):223–239.

Donoho, D. L. (2006). For most large underdetermined systems of linear equations, the minimal l1-norm near-solution approximates the sparsest near-solution. *Commun. Pur. Appl. Math.*, 59(7):907–934.

Fryzlewicz, P., Sapatinas, T. and Subba Rao, S. (2006). A Haar-Fisz technique for locally stationary volatility estimation. *Biometrika*, 93(3):687–704.

Gombay, E. (2008). Change detection in autoregressive time series. *J. Multivar. Anal.*, 99(3):451–464.

Ishii, N., Twata, A. and Suzumura, N. (1979). Segmentation of non-stationary time series. *Int. J. Systems Sci.*, 10(8):883–894.

Lavielle, M. (1998). Optimal segmentation of random processes. *IEEE Trans. on Signal Processing*, 46(5):1365–1373.

Nason, G. P., von Sachs, R. and Kroisandt, G. (2000). Wavelet processes and adaptive estimation of the evolutionary wavelet spectrum. *J. Roy. Stat. Soc. B*, 62(2):271–292.

Ombao, H. C., Raz, J. A., von Sachs, R. and Malow, B. A. (2001). Automatic statistical analysis of bivariate nonstationary time series. *J. Am. Stat. Assoc.*, 86(454):543–560.

Ombao, H. C., Raz, J. A., von Sachs, R. and Guo, W. (2002). The SLEX model of a non-stationary random process. *Ann. I. Stat. Math.*, 54(1):171–200.

De la Peña, V. H. (1999). A general class of exponential inequalities for martingales and ratios. *Ann. Prob.*, 27(1):537–564.

Priestley, M. B. (1965). Evolutionary spectra and non-stationary processes. *J. Roy. Stat. Soc. B*, 27(2):204–237.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *J. Roy. Stat. Soc. B*, 58(1):267–288.