Multiscale breakpoint detection in piecewise stationary processes

Haeran Cho    Piotr Fryzlewicz

University of Bristol, UK

Young Statisticians’ Meeting,
Glasgow, 8 April 2009
Introduction

Stationarity assumption
- Autocovariance structure does not change over time.
- Appealing when analyzing short time series.
- Often unrealistic for longer processes.
- Many naturally occurring phenomena cannot be modeled as stationary processes.
  - Signal processing, quality control, econometrics, etc.

Our aim
- Modeling a nonstationary process by consistently detecting its structural breakpoints.
Introduction

An example of piecewise stationary processes
- can you find any breakpoint in this process?
Retrospective segmentation of piecewise stationary processes

- Find multiple breakpoints occurred in the past.

- Applied in the spectral domain.
  - Any change in the autocovariance structure is detected in wavelet spectra.
  - Does not require any specific model assumption.

- Based on the binary segmentation.
  - Enables simultaneous locating and testing of a breakpoint.
  - Segments a sequence in a recursive manner - multiscale procedure.
Piecewise stationary processes

Definition the locally stationary wavelet (LSW) time series

\[ X_{t,T} = \sum_{i=-\infty}^{-1} \sum_{k=-\infty}^{\infty} W_i(k/T) \psi_i, t - k \xi_i, k \]

- \( i \): scale parameter, \( k \): location parameter.
- \( W_i(z) : [0, 1] \rightarrow \mathbb{R} \): real-valued, scale- and location-dependent piecewise constant function with finite but unknown number of jumps.
- \( \psi_i \): discrete, real-valued, compactly supported, non-decimated wavelet vectors.
- \( \xi_{i,k} \): i.i.d. Gaussian variables.
The wavelet periodogram of $X_{t,T}$ at scale $i$

$$I_{t,T}^{(i)} = \left| \sum_s X_{s,T} \psi_{i,s-t} \right|^2$$

- A sequence of squared wavelet coefficients of $X_{t,T}$.
- Scaled $\chi^2_1$ variables as $I_{t,T}^{(i)} = \mathbb{E}I_{t,T}^{(i)} \cdot Z_{t,T}^2$.
  - $\mathbb{E}I_{t,T}^{(i)}$: “almost” piecewise constant as $W_i(z)$.
  - $\{Z_{t,T}\}_{t=0}^{T-1}$: standard Gaussian variables (autocorrelated).
Wavelet periodograms

First develop a generic segmentation tool applicable to any sequence of the form:

\[ Y^2_{t,T} = \sigma^2_{t,T} \cdot Z^2_{t,T}, \quad t = 1, \ldots, T \]

- \( \sigma^2_{t,T} \) is deterministic and “close” to a piecewise constant function \( \sigma^2(z) \),
- \( \{ Z_{t,T} \}_{t=1}^{T} \) is multivariate Gaussian with mean zero, variance one, and absolutely summable autocorrelation (asymptotically).
- \( I^{(i)}_{t,T} \) and \( E I^{(i)}_{t,T} \) correspond to \( Y^2_{t,T} \) and \( \sigma^2_{t,T} \) respectively.
Binary segmentation

For a sequence of observations \( \{ Y_{t,T}, \ t = 1, \cdots, T \} \),
- Find a potential breakpoint \( b \in (1, T) \).
- Test the hypotheses
  - \( H_0 \): \( \{ Y_{t,T} \}_{t=1}^T \) are from a common model.
  - \( H_1 \): \( \{ Y_{t,T} \}_{t=1}^b \) and \( \{ Y_{t,T} \}_{t=b+1}^T \) are from different models.
- Repeat on each segment until no further breakpoint is detected.
Algorithm

Step 1  Compute $d_k$ for $k \in (1, T)$ where

$$d_k = \frac{\sqrt{T-k}}{\sqrt{T-k}} \sum_{t=1}^{k} Y_{t,T}^2 - \frac{\sqrt{k}}{\sqrt{T-(T-k)}} \sum_{t=k+1}^{T} Y_{t,T}^2,$$

and find $b = \arg \max_k |d_k|$. Denote $d = d_b$ and $m = \sum_{t=1}^{T} Y_{t,T}^2 / T$.

Step 2  Perform hard-thresholding on $|d|/m$ with the threshold $t = \tau T^{\theta} \sqrt{\log T}$; $\hat{d} = d$ if $|d|/m > t$ and $\hat{d} = 0$ otherwise.

Step 3  If $\hat{d} \neq 0$, repeat the same procedure on each segment iteratively.
Consistency of detected breakpoints

- The total number and locations of detected breakpoints are consistent with probability converging to 1.
- Provided true breakpoints \( \eta_l, \ l = 1, \cdots, B \) are not too “close” to each other.

**Theorem**

Detected breakpoints \( \hat{\eta}_l \) for \( l = 1, \cdots, \hat{B} \) satisfy

\[
\Pr \left\{ \hat{B} = B; \ |\hat{\eta}_l - \eta_l| \leq \epsilon_T, \ 1 \leq l \leq B \right\} \rightarrow 1,
\]

where \( \epsilon_T = O(T^{1/2} \log T) \).
Simulation: the example from Introduction

\[ X_t = \begin{cases} 
0.999X_{t-1} + \epsilon_t & \text{for } 1 \leq t \leq 400, \\
0.999X_{t-1} + 1.5\epsilon_t & \text{for } 401 \leq t \leq 750, \\
0.999X_{t-1} + \epsilon_t & \text{for } 751 \leq t \leq 1024 
\end{cases} \]

- dotted red lines: true breakpoints
- dashed blue lines: detected breakpoints
Simulation: overview

- Sim 1: Piecewise stationary (PS) AR(2) process with observable changes
- Sim 2: PS AR(1) process with less observable changes
- Sim 3: PS AR(1) process with a short segment
- Sim 4: PS ARMA(1, 1) process
- Sim 5: Almost unit-root PS AR(1) with changes in variance only
Simulation: examples

Sim 1

Sim 2

Sim 3

Sim 4
**Simulation: outcome**

**Table:** Summary of breakpoint detection from simulations

<table>
<thead>
<tr>
<th></th>
<th>Sim 1 BS</th>
<th>Sim 1 Auto-PARM†</th>
<th>Sim 2 BS</th>
<th>Sim 2 AP</th>
<th>Sim 3 BS</th>
<th>Sim 3 AP</th>
<th>Sim 4 BS</th>
<th>Sim 4 AP</th>
<th>Sim 5 BS</th>
<th>Sim 5 AP</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>95</td>
<td>99</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>33</td>
</tr>
<tr>
<td>2</td>
<td>92</td>
<td>91</td>
<td>96</td>
<td>100</td>
<td>3</td>
<td>0</td>
<td>99</td>
<td>100</td>
<td>95</td>
<td>21</td>
</tr>
<tr>
<td>3</td>
<td>6</td>
<td>9</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>10</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>total</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>

Simulation: almost unit-root like process

Detected breakpoints over 100 realizations.

(a) The breakpoints detected from our method centre on $\eta_1 = 400$, $\eta_2 = 750$,
(b) while those from the Auto-PARM are scattered everywhere.
Dow-Jones Industrial daily average 2007-2009

Daily closing values of Dow-Jones Industrial average.
(08/01/2007-16/01/2009, $T = 512 = 2^9$)
Two estimated breaks around the last week of July 2007 and mid-September 2008. \((t = 136, 425)\)

Soon afterwards the first break, worldwide credit crunch broke out.

The second break coincides with the bankruptcy of Lehman Brothers (14 September 2008), the event which has brought even more volatility to market.
Thanks!