Structural Break Detection in Time Series Models

Richard A. Davis
Thomas Lee
Gabriel Rodriguez-Yam

Colorado State University
(http://www.stat.colostate.edu/~rdavis/lectures)

This research supported in part by an IBM faculty award.
How many segments do you see?

\[ \tau_1 = 51 \quad \tau_2 = 151 \quad \tau_3 = 251 \]
Illustrative Example

Auto-PARM=Auto-Piecewise AutoRegressive Modeling

4 pieces, 2.58 seconds.

\[ \tau_1 = 51 \quad \tau_2 = 157 \quad \text{time} \quad \tau_3 = 259 \]
A Second Example

Any breaks in this series?
Introduction

- Examples
  - AR
  - GARCH
  - Stochastic volatility
  - State space models

Model selection using Minimum Description Length (MDL)

- General principles
- Application to AR models with breaks

Optimization using a Genetic Algorithm

- Basics
- Implementation for structural break estimation

Simulation results

Applications

Simulation results for GARCH and SV models
1. Piecewise AR model:

\[ Y_t = \gamma_j + \phi_{j1} Y_{t-1} + \cdots + \phi_{jp_j} Y_{t-p_j} + \sigma_j \varepsilon_t, \quad \text{if } \tau_{j-1} \leq t < \tau_j, \]

where \( \tau_0 = 1 < \tau_1 < \cdots < \tau_{m-1} < \tau_m = n + 1 \), and \( \{\varepsilon_t\} \) is IID(0,1).

**Goal:** Estimate

- \( m = \) number of segments
- \( \tau_j = \) location of \( j^{th} \) break point
- \( \gamma_j = \) level in \( j^{th} \) epoch
- \( p_j = \) order of AR process in \( j^{th} \) epoch
- \( (\phi_{j1}, \ldots, \phi_{jp_j}) = \) AR coefficients in \( j^{th} \) epoch
- \( \sigma_j = \) scale in \( j^{th} \) epoch
2. Segmented GARCH model:

\[ Y_t = \sigma_t \epsilon_t, \]
\[ \sigma_t^2 = \omega_j + \alpha_{j1} Y_{t-1}^2 + \cdots + \alpha_{jp_j} Y_{t-p_j}^2 + \beta_{j1} \sigma_{t-1}^2 + \cdots + \beta_{jq_j} \sigma_{t-q_j}^2, \quad \text{if } \tau_{j-1} \leq t < \tau_j, \]

where \( \tau_0 = 1 < \tau_1 < \ldots < \tau_{m-1} < \tau_m = n + 1 \), and \( \{\epsilon_t\} \) is IID(0,1).

3. Segmented stochastic volatility model:

\[ Y_t = \sigma_t \epsilon_t, \]
\[ \log \sigma_t^2 = \gamma_j + \phi_{j1} \log \sigma_{t-1}^2 + \cdots + \phi_{jp_j} \log \sigma_{t-p_j}^2 + \nu_j \eta_t, \quad \text{if } \tau_{j-1} \leq t < \tau_j. \]

4. Segmented state-space model (SVM a special case):

\[ p(y_t \mid \alpha_t, \ldots, \alpha_1, y_{t-1}, \ldots, y_1) = p(y_t \mid \alpha_t) \text{ is specified} \]
\[ \alpha_t = \gamma_j + \phi_{j1} \alpha_{t-1} + \cdots + \phi_{jp_j} \alpha_{t-p_j} + \sigma_j \eta_t, \quad \text{if } \tau_{j-1} \leq t < \tau_j. \]
Model Selection Using Minimum Description Length

Basics of MDL:
Choose the model which \textit{maximizes the compression} of the data or, equivalently, select the model that \textit{minimizes the code length} of the data (i.e., amount of memory required to encode the data).

\[ M = \text{class of operating models for } y = (y_1, \ldots, y_n) \]

\[ L_F(y) = \text{code length of } y \text{ relative to } F \in M \]

Typically, this term can be decomposed into two pieces (two-part code),

\[ L_F(y) = L(\hat{F}|y) + L(\hat{e} | \hat{F}), \]

where

\[ L(\hat{F}|y) = \text{code length of the fitted model for } F \]

\[ L(\hat{e}|\hat{F}) = \text{code length of the residuals based on the fitted model} \]
Model Selection Using Minimum Description Length (cont)

Applied to the segmented AR model:

\[ Y_t = \gamma_j + \phi_{j1}Y_{t-1} + \cdots + \phi_{j\,p_j}Y_{t-p_j} + \sigma_j \varepsilon_t, \text{ if } \tau_{j-1} \leq t < \tau_j, \]

First term \( L(\hat{F} | y) \):

\[
L(\hat{F} | y) = L(m) + L(\tau_1, \ldots, \tau_m) + L(p_1, \ldots, p_m) + L(\hat{\psi}_1 | y) + \cdots + L(\hat{\psi}_m | y)
= \log_2 m + m \log_2 n + \sum_{j=1}^{m} \log_2 p_j + \sum_{j=1}^{m} \frac{p_j + 2}{2} \log_2 n_j
\]

Second term \( L(\hat{e} | \hat{F}) \):

\[
L(\hat{e} | \hat{F}) \approx -\sum_{j=1}^{m} \log_2 L(\hat{\psi}_j | y)
\]

\[
MDL(m, (\tau_1, p_1), \ldots, (\tau_m, p_m))
= \log_2 m + m \log_2 n + \sum_{j=1}^{m} \log_2 p_j + \sum_{j=1}^{m} \frac{p_j + 2}{2} \log_2 n_j + \sum_{j=1}^{m} (\log_2 (2\pi \hat{\sigma}_j^2) + n_j)
\]
Optimization Using Genetic Algorithm

Basics of GA:
Class of optimization algorithms that mimic natural evolution.

- Start with an initial set of *chromosomes*, or population, of possible solutions to the optimization problem.
- Parent chromosomes are randomly selected (proportional to the rank of their objective function values), and produce offspring using *crossover* or *mutation* operations.
- After a sufficient number of offspring are produced to form a second generation, the process then *restarts to produce a third generation*.
- Based on Darwin’s *theory of natural selection*, the process should produce future generations that give a *smaller (or larger)* objective function.
Optimization Using Genetic Algorithm

Genetic Algorithm: Chromosome consists of $n$ genes, each taking the value of $-1$ (no break) or $p$ (order of AR process). Use natural selection to find a *near* optimal solution. Map the break points with a chromosome $c$ via

$$(m, (\tau_1, p_1), \ldots, (\tau_m, p_m)) \leftrightarrow c = (\delta_1, \ldots, \delta_n),$$

where

$$\delta_t = \begin{cases} -1, & \text{if no break point at } t, \\ p_j, & \text{if break point at time } t = \tau_{j-1} \text{ and AR order is } p_j. \end{cases}$$

For example,

$$c = (2, -1, -1, -1, -1, 0, -1, -1, -1, 0, -1, -1, -1, 3, -1, -1, -1, -1, -1)$$

$t$: 1 6 11 15

would correspond to a process as follows:

$\text{AR}(2), t=1:5; \text{AR}(0), t=6:10; \text{AR}(0), t=11:14; \text{AR}(3), t=15:20$
Generation 0: Start with \( L \) (200) randomly generated chromosomes, \( c_1, \ldots, c_L \) with associated MDL values, \( MDL(c_1), \ldots, MDL(c_L) \).

Generation 1: A new child in the next generation is formed from the chromosomes \( c_1, \ldots, c_L \) of the previous generation as follows:

- with probability \( \pi_c \), crossover occurs.
  - two parent chromosomes \( c_i \) and \( c_j \) are selected at random with probabilities proportional to the ranks of \( MDL(c_i) \).
  - \( k^{th} \) gene of child is \( \delta_k = \delta_{i,k} \) w.p. \( \frac{1}{2} \) and \( \delta_{j,k} \) w.p. \( \frac{1}{2} \)

- with probability \( 1 - \pi_c \), mutation occurs.
  - a parent chromosome \( c_i \) is selected
  - \( k^{th} \) gene of child is \( \delta_k = \delta_{i,k} \) w.p. \( \pi_1 \) ; \(-1 \) w.p. \( \pi_2 \); and \( p \) w.p. \( 1 - \pi_1 - \pi_2 \).
Implementation of Genetic Algorithm—(cont)

Execution of GA: Run GA until convergence or until a maximum number of generations has been reached.

Various Strategies:

- include the top ten chromosomes from last generation in next generation.

- use multiple islands, in which populations run independently, and then allow migration after a fixed number of generations. This implementation is amenable to parallel computing.
1. Piecewise stationary with dyadic structure: Consider a time series following the model,

\[ Y_t = \begin{cases} 
  .9Y_{t-1} + \varepsilon_t, & \text{if } 1 \leq t < 513, \\
  1.69Y_{t-1} - .81Y_{t-2} + \varepsilon_t, & \text{if } 513 \leq t < 769, \\
  1.32Y_{t-1} - .81Y_{t-2} + \varepsilon_t, & \text{if } 769 \leq t \leq 1024, 
\end{cases} \]

where \( \{\varepsilon_t\} \sim \text{IID N}(0,1) \).
1. Piecewise stat (cont)

Implementation: Start with $NI = 50$ islands, each with population size $L = 200$.

After every $Mi = 5$ generations, allow migration.

Replace worst 2 in Island 3 with best 2 from Island 2.

Replace worst 2 in Island 4 with best 2 from Island 3.

Replace worst 2 in Island 2 with best 2 from Island 4.

Stopping rule: Stop when the max MDL does not change for 10 consecutive migrations or after 100 migrations.

Span configuration for model selection: Max AR order $K = 10$,

<table>
<thead>
<tr>
<th>$p$</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7-10</th>
<th>11-20</th>
</tr>
</thead>
<tbody>
<tr>
<td>$m_p$</td>
<td>10</td>
<td>10</td>
<td>12</td>
<td>14</td>
<td>16</td>
<td>18</td>
<td>20</td>
<td>25</td>
<td>50</td>
</tr>
<tr>
<td>$\pi_p$</td>
<td>1/21</td>
<td>1/21</td>
<td>1/21</td>
<td>1/21</td>
<td>1/21</td>
<td>1/21</td>
<td>1/21</td>
<td>1/21</td>
<td>1/21</td>
</tr>
</tbody>
</table>
1. Piecewise stat (cont)

GA results: 3 pieces breaks at $\tau_1=513; \tau_2=769$. Total run time 16.31 secs

Fitted model:

<table>
<thead>
<tr>
<th></th>
<th>$\phi_1$</th>
<th>$\phi_2$</th>
<th>$\sigma^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1- 512:</td>
<td>0.857</td>
<td>.9945</td>
<td></td>
</tr>
<tr>
<td>513-768:</td>
<td>1.68</td>
<td>-0.801</td>
<td>1.1134</td>
</tr>
<tr>
<td>769-1024:</td>
<td>1.36</td>
<td>-0.801</td>
<td>1.1300</td>
</tr>
</tbody>
</table>

True Model

Fitted Model
2. Slowly varying AR(2) model:

\[ Y_t = a_t Y_{t-1} - 0.81 Y_{t-2} + \varepsilon_t \quad \text{if} \quad 1 \leq t \leq 1024 \]

where \( a_t = 0.8[1 - 0.5 \cos(\pi t / 1024)] \), and \( \{\varepsilon_t\} \sim \text{IID } N(0,1) \).
2. Slowly varying AR(2) (cont)

GA results: 3 pieces, breaks at $\tau_1=293$, $\tau_2=615$. Total run time 27.45 secs

Fitted model:

<table>
<thead>
<tr>
<th></th>
<th>$\phi_1$</th>
<th>$\phi_2$</th>
<th>$\sigma^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-292</td>
<td>0.365</td>
<td>-0.753</td>
<td>1.149</td>
</tr>
<tr>
<td>293-614</td>
<td>0.821</td>
<td>-0.790</td>
<td>1.176</td>
</tr>
<tr>
<td>615-1024</td>
<td>1.084</td>
<td>-0.760</td>
<td>0.960</td>
</tr>
</tbody>
</table>

True Model

Fitted Model
2. Slowly varying AR(2) (cont)

In the graph below right, we average the spectogram over the GA fitted models generated from each of the 200 simulated realizations.
3. Simulated data from Fearnhead (2005):

True model has 9 changepoints

MAP est of $m=9$ while MAP of $m$ and changepoint locations gives $m=8$ changepoints. Plot is conditional on 9 changepoints.
4. Fearnhead example

**True Model**

**Fitted APARM Model**
Theory

Consistency.

Suppose the number of change points $m$ is known and let

$$\lambda_1 = \tau_1/n, \ldots, \lambda_m = \tau_m/n$$

be the relative (true) changepoints. Then

$$\hat{\lambda}_j \rightarrow \lambda_j \quad \text{a.s.}$$

where $\hat{\lambda}_j = \hat{\tau}_j/n$ and $\hat{\tau}_j$ = Auto-PARM estimate of $\tau_j$.

Consistency of the estimate of $m$?

- For $n$ large, Auto-PARM estimate is $\geq m$.
- Have not proved equality.
Examples

Speech signal: GREASY
Speech signal: GREASY

$n = 5762$ observations

$m = 15$ break points

Run time = 18.02 secs
Examples

Mine explosion seismic trace in Scandinavia: (Shumway and Stoffer 2000, Stoffer et al. 2005)

Two waves: P (primary) compression wave and S (shear) wave
Examples

AR orders: 1 7 17 13 15
Example: EEG Time series

Data: Bivariate EEG time series at channels T3 (left temporal) and P3 (left parietal). Female subject was diagnosed with left temporal lobe epilepsy. Data collected by Dr. Beth Malow and analyzed in Ombao et al (2001). (n=32,768; sampling rate of 100Hz). Seizure started at about 1.85 seconds.

GA bivariate results: 11 pieces with AR orders T3: 17, 2, 6, 15, 2, 3, 5, 9, 5, 4, 1

Data: Bivariant EEG time series at channels T3 (left temporal) and P3 (left parietal). Female subject was diagnosed with left temporal lobe epilepsy. Data collected by Dr. Beth Malow and analyzed in Ombao et al (2001). (n=32,768; sampling rate of 100Hz). Seizure started at about 1.85 seconds.

GA bivariate results: 11 pieces with AR orders T3: 17, 2, 6, 15, 2, 3, 5, 9, 5, 4, 1

T3 Channel

P3 Channel
Remarks:

- the general conclusions of this analysis are similar to those reached in Ombao et al.
- prior to seizure, power concentrated at lower frequencies and then spread to high frequencies.
- power returned to the lower frequencies at conclusion of seizure.
Remarks (cont):  
- T3 and P3 strongly coherent at 9-12 Hz prior to seizure.  
- strong coherence at low frequencies just after onset of seizure.  
- strong coherence shifted to high frequencies during the seizure.
Application to GARCH

**Garch(1,1) model:**  
\[ Y_t = \sigma_t \varepsilon_t, \quad \{\varepsilon_t\} \sim \text{IID}(0,1) \]

\[ \sigma_t^2 = \omega_j + \alpha_j Y_{t-1}^2 + \beta_j \sigma_{t-1}^2, \quad \text{if } \tau_{j-1} \leq t < \tau_j. \]

\[ \sigma_t^2 = \begin{cases} 
.4 + .1 Y_{t-1}^2 + .5 \sigma_{t-1}^2, & \text{if } 1 \leq t < 501, \\
.4 + .1 Y_{t-1}^2 + .6 \sigma_{t-1}^2, & \text{if } 501 \leq t < 1000 
\end{cases} \]

**CP estimate = 506**

AG = Andreou and Ghysels (2002)

<table>
<thead>
<tr>
<th># of CPs</th>
<th>GA %</th>
<th>AG %</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>80.4</td>
<td>72.0</td>
</tr>
<tr>
<td>1</td>
<td>19.2</td>
<td>24.0</td>
</tr>
<tr>
<td>( \geq 2 )</td>
<td>0.4</td>
<td>0.4</td>
</tr>
</tbody>
</table>
Application to GARCH (cont)

**Garch(1,1) model:** \( Y_t = \sigma_t \varepsilon_t, \quad \{\varepsilon_t\} \sim \text{IID}(0,1) \)

\[
\sigma_t^2 = \omega_j + \alpha_j Y_{t-1}^2 + \beta_j \sigma_{t-1}^2, \quad \text{if } \tau_{j-1} \leq t < \tau_j.
\]

\[
\sigma_t^2 = \begin{cases} 
0.4 + 1Y_{t-1}^2 + 0.5\sigma_{t-1}^2, & \text{if } 1 \leq t < 501 \\
0.4 + 1Y_{t-1}^2 + 0.8\sigma_{t-1}^2, & \text{if } 501 \leq t < 1000
\end{cases}
\]

*CP estimate = 502*

*AG = Andreou and Ghysels (2002)*

<table>
<thead>
<tr>
<th># of CPs</th>
<th>GA %</th>
<th>AG %</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>1</td>
<td>96.4</td>
<td>95.0</td>
</tr>
<tr>
<td>(\geq 2)</td>
<td>3.6</td>
<td>0.5</td>
</tr>
</tbody>
</table>
Application to GARCH (cont)

More simulation results for Garch(1,1): \( Y_t = \sigma_t \varepsilon_t, \quad \{ \varepsilon_t \} \sim \text{IID}(0,1) \)

\[
\sigma_t^2 = \begin{cases} 
0.05 + 0.4Y_{t-1}^2 + 0.3\sigma_{t-1}^2, & \text{if } 1 \leq t < \tau_1, \\
1.00 + 0.3Y_{t-1}^2 + 0.2\sigma_{t-1}^2, & \text{if } \tau_1 \leq t < 1000 
\end{cases}
\]

<table>
<thead>
<tr>
<th>( \tau_1 )</th>
<th>( \text{Mean} )</th>
<th>( \text{SE} )</th>
<th>( \text{Med} )</th>
<th>( \text{Freq} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>GA</td>
<td>52.62</td>
<td>11.70</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>Berkes</td>
<td>71.40</td>
<td>12.40</td>
<td>71</td>
</tr>
<tr>
<td>250</td>
<td>GA</td>
<td>251.18</td>
<td>4.50</td>
<td>250</td>
</tr>
<tr>
<td></td>
<td>Berkes</td>
<td>272.30</td>
<td>18.10</td>
<td>271</td>
</tr>
<tr>
<td>500</td>
<td>GA</td>
<td>501.22</td>
<td>4.76</td>
<td>502</td>
</tr>
<tr>
<td></td>
<td>Berkes</td>
<td>516.40</td>
<td>54.70</td>
<td>538</td>
</tr>
</tbody>
</table>

State Space Model Setup:

**Observation equation:**

\[ p(y_t | \alpha_t) = \exp\{\alpha_t y_t - b(\alpha_t) + c(y_t)\}. \]

**State equation:** \( \{\alpha_t\} \) follows the piecewise AR(1) model given by

\[ \alpha_t = \gamma_k + \phi_k \alpha_{t-1} + \sigma_k \varepsilon_t, \quad \text{if} \quad \tau_{k-1} \leq t < \tau_k, \]

where \( 1 = \tau_0 < \tau_1 < \ldots < \tau_m < n \), and \( \{\varepsilon_t\} \sim \text{IID } N(0,1). \)

**Parameters:**

- \( m \) = number of break points
- \( \tau_k \) = location of break points
- \( \gamma_k \) = level in \( k^{\text{th}} \) epoch
- \( \phi_k \) = AR coefficients \( k^{\text{th}} \) epoch
- \( \sigma_k \) = scale in \( k^{\text{th}} \) epoch
Remark: The exact likelihood is given by the following formula

\[ L(\psi; y_n) = L_a(\psi; y_n) Er_a(\psi), \]

where

\[ Er_a(\psi) = \int \exp\{R(\alpha_n; \alpha^*)\} p_a(\alpha_n | y_n; \psi) d\alpha_n. \]

It turns out that \( \log(Er_a(\psi)) \) is nearly linear and can be approximated by a linear function via importance sampling,

\[ e(\psi) \sim e(\hat{\psi}_{AL}) + \dot{e}(\hat{\psi}_{AL})(\psi - \hat{\psi}_{AL}) \]
Count Data Example

Model: $Y_t \mid \alpha_t \sim Pois(\exp\{\beta + \alpha_t\}), \ \alpha_t = \phi \alpha_{t-1} + \varepsilon_t, \ \{\varepsilon_t\} \sim \text{IID } N(0, \sigma^2)$

True model:
- $Y_t \mid \alpha_t \sim Pois(\exp\{.7 + \alpha_t\}), \ \alpha_t = .5\alpha_{t-1} + \varepsilon_t, \ \{\varepsilon_t\} \sim \text{IID } N(0, .3), \ t < 250$
- $Y_t \mid \alpha_t \sim Pois(\exp\{.7 + \alpha_t\}), \ \alpha_t = -.5\alpha_{t-1} + \varepsilon_t, \ \{\varepsilon_t\} \sim \text{IID } N(0, .3), \ t > 250.$
- GA estimate 251, time 267secs
SV Process Example

Model: \( Y_t | \alpha_t \sim N(0, \exp\{\alpha_t\}), \quad \alpha_t = \gamma + \phi \alpha_{t-1} + \varepsilon_t, \quad \{\varepsilon_t\} \sim \text{IID } N(0, \sigma^2) \)

- \( Y_t | \alpha_t \sim N(0, \exp\{\alpha_t\}), \quad \alpha_t = -.05 + .975 \alpha_{t-1} + \varepsilon_t, \quad \{\varepsilon_t\} \sim \text{IID } N(0, .05), \quad t \leq 750 \)
- \( Y_t | \alpha_t \sim N(0, \exp\{\alpha_t\}), \quad \alpha_t = -.25 + .900 \alpha_{t-1} + \varepsilon_t, \quad \{\varepsilon_t\} \sim \text{IID } N(0, .25), \quad t > 750 \)
- GA estimate 754, time 1053 secs
SV Process Example

Model: \( Y_t | \alpha_t \sim N(0, \exp{\alpha_t}) \), \( \alpha_t = \gamma + \phi \alpha_{t-1} + \epsilon_t \), \( \{\epsilon_t\} \sim \text{IID } N(0, \sigma^2) \)

True model:
- \( Y_t | \alpha_t \sim N(0, \exp{\alpha_t}) \), \( \alpha_t = -.175 + .977\alpha_{t-1} + \epsilon_t \), \( \{\epsilon_t\} \sim \text{IID } N(0, .1810) \), \( t \leq 250 \)
- \( Y_t | \alpha_t \sim N(0, \exp{\alpha_t}) \), \( \alpha_t = -.010 + .996\alpha_{t-1} + \epsilon_t \), \( \{\epsilon_t\} \sim \text{IID } N(0, .0089) \), \( t > 250 \).
- GA estimate 251, time 269s
True model:
- \( Y_t \mid \alpha_t \sim N(0, \exp\{\alpha_t\}) \), \( \alpha_t = -.175 + .977\alpha_{t-1} + \varepsilon_t \), \( \{\varepsilon_t\} \sim \text{IID } N(0, .1810), \ t \leq 250 \)
- \( Y_t \mid \alpha_t \sim N(0, \exp\{\alpha_t\}) \), \( \alpha_t = -.010 + .996\alpha_{t-1} + \varepsilon_t \), \( \{\varepsilon_t\} \sim \text{IID } N(0, .0089), \ t > 250 \).

Fitted model based on no structural break:
- \( Y_t \mid \alpha_t \sim N(0, \exp\{\alpha_t\}) \), \( \alpha_t = -.0645 + .9889\alpha_{t-1} + \varepsilon_t \), \( \{\varepsilon_t\} \sim \text{IID } N(0, .0935) \)
Fitted model based on no structural break:

- \( Y_t \mid \alpha_t \sim N(0, \exp\{\alpha_t\}) \), \( \alpha_t = -0.0645 + 0.9889\alpha_{t-1} + \varepsilon_t \), \( \{\varepsilon_t\} \sim \text{IID } N(0, 0.0935) \)