Structural Break Estimation for Non-Stationary Time Series Signals

Richard A. Davis, Thomas C.M. Lee & Gabriel A. Rodriguez-Yam, Colorado State University

http://www.stat.colostate.edu/~rdavis

1. Introduction

We consider the problem of modeling a non-stationary time series by segmenting the series into blocks of different autoregressive (AR) processes. The number of break points, denoted by *m*, as well as their location, and the order of the respective AR models are assumed to be unknown. We propose an automatic procedure for obtaining such an optimal partition called ***Auto-PARM*** for ***A***utomatic ***P***iecewise ***A***uto***R***egressive ***M***odeling.

1.1 Piecewise Autoregressive processes

Setup: there exist m and $\tau_0 = 1 < \tau_1 < \ldots < \tau_m < \tau_{m+1} = n + 1$ (n = sample size) such that

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

where $\{\varepsilon_t\}$ is IID(0,1).

Goal: Estimate

m = number of segments

τ_j = location of j^{th} break point

γ_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

σ_j = scale in j^{th} epoch

1.2 Motivation for using piecewise AR models

Piecewise AR is a special case of a *piecewise stationary process* (see Adak 1998),

$$\tilde{Y}_{t,n} = \sum_{j=1}^{m} Y_t^j I_{[\tau_{j-1},\tau_j)}(t/n),$$

where $\{Y_t^j\}$, $j = 1, \ldots, m$ is a sequence of stationary processes. It is argued in Ombao et al. (2001) that if $\{Y_{t,n}\}$ is a locally stationary process (in the sense of Dahlhaus), then there exist a piecewise stationary process $\{\tilde{Y}_{t,n}\}$ and a sequence m_n

$$m_n \to \infty \ \text{ and } \ m_n/n \to 0, \ \text{as } n \to \infty,$$

that approximates $\{Y_{t,n}\}$ (in average mean square).

Roughly speaking: $\{Y_{t,n}\}$ is a locally stationary process if it has a time-varying spectrum that is approximately $|A(t/n,w)|^2$, where $A(u,w)$ is a continuous function in u.

2.1 Model selection using Minimum Description Length (MDL)

The idea behind MDL is to choose the model which *maximizes the compression* of the data or, equivalently, select the model that *minimizes the code length* of the data (i.e., amount of memory required to encode the data).

2.2 The MDL applied to piecewise AR models

$\boldsymbol{M}$ = class of piecewise AR models for $y = (y_1, \ldots, y_n)$

$L_{\boldsymbol{F}}(y)$ = code length of y relative to $\boldsymbol{F} \in \boldsymbol{M}$

Best fitting MDL model is minimizer of

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

where n_j is the length of the j-th segment and $\hat{\sigma}_j^2$ is the Yule-Walker estimate of σ^2 in the j-th segment.

2.3 Consistency

Assume there exist true values m and $0 < \lambda_1 < \ldots < \lambda_m < 1$ with

$$\tau_i = [\lambda_i n], \ i = 1,2, \ldots, m.$$

Theorem. For the piecewise AR model, if the number of breakpoints m is known, then

$$\hat{\lambda}_j \to \lambda_j \ \text{ a.s. } j = 1, 2, \ldots, m.$$

3.1 Basics of the Genetic Algorithm (GA)

The GA is an optimization algorithms that mimics 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.

3.2 Implementation of GA

A 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. An element $\boldsymbol{F} \in \boldsymbol{M}$ is mapped with a chromosome c by

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

For example,

```
c = (2, -1, -1, -1, -1, 0, -1,  -1, -1, -1, 0, -1, -1, -1, 3, -1, -1, -1, -1,-1)
  t: 1                  6                   11              15
```

corresponds to

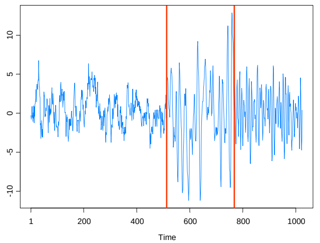
AR(2), t=1:5; AR(0), t=6:10; AR(0), t=11:14; AR(3), t=15:20

4.1 Piecewise stationary with dyadic structure

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

where $\{\varepsilon_t\}$ ~ IID N(0,1).

Sample realization.



Auto-PARM results: 3 pieces at τ_1=513, τ_2=769; AR(1); AR(2); AR(2)

Total run time: 16.31 secs

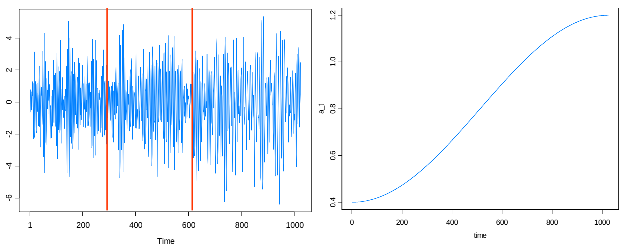
Simulation Results based on 200 reps

# of segments	Auto-SLEX		Auto-PARM		
	%	Change Points	%	mean	std
2	0	1/2	0		
3	60.0	1/4 , 3/4	82.0	.500 .749	.006 .006
4	34.0	1/4, 2/4, 3/4	17.5	.476 .616 .761	.080 .110 .037
5	5.0	2/8, 4/8, 5/8, 6/8, 7/8	0		
≥ 6	1.0		0.5		

4.2 Slowly varying AR(2) model

$$Y_t = a_t Y_{t-1} - .81Y_{t-2} + \varepsilon_t \ \text{ if } 1 \le t \le 1024$$

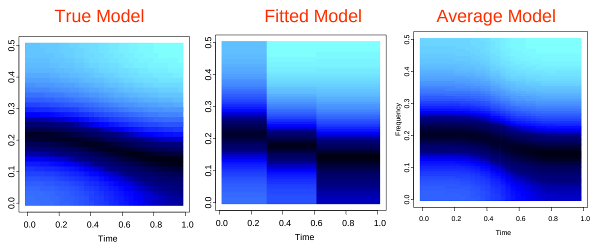
where $a_t = .8[1 - 0.5\cos(\pi t/1024)]$, and $\{\varepsilon_t\}$ ~ IID N(0,1).



GA results: 3 pieces, breaks at τ_1=293, τ_2=615. Total run time 27.45 secs

Fitted model:

	ϕ_1	ϕ_2	σ^2
1- 292:	.365	-0.753	1.149
293- 614:	.821	-0.790	1.176
615-1024:	1.084	-0.760	0.960

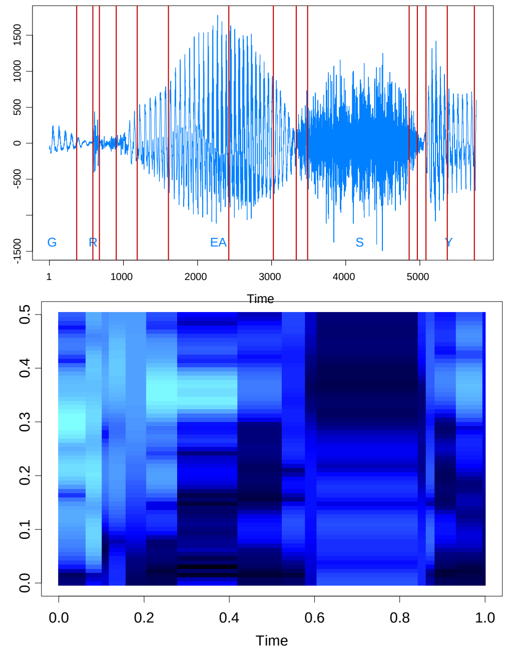


Left: True model log-spectrogram; Center: Auto-PARM log-spectrogram; Right: Average of Auto-PARM log-spectrogram based on 200 reps.

5. Speech signal segmentation

Speech signal: GREASY, n = 5762 observations

Auto-PARM results: m = 15 break points, run time = 18.02s



Bottom: Spectogram based on Auto-PARM model

6. Conclusions

- Introduced ***Auto-PARM*** (an automatic procedure for segmenting a time series signal into piecewise AR models).
- Model selection based on ***MDL*** (minimum description length) principle.
- A ***genetic algorithm*** was used to find a near optimal solution to the model selection problem based on MDL.
- Auto-PARM works well for both detecting segments and for estimating ***time-varying spectra***.

7. References

- Davis, R.A., Lee, T.C.M. and Rodriguez-Yam, G.A. (2005). "Structural Break Estimation for Nonstationary Time Series Models," (To appear in *JASA*.)
- Kittagawa, G. and Akaike, H. (1978). "A Procedure for the Modeling of Non-Stationary Time Series," *Ann of Inst of Stat Math*. **30,** 351-363.
- Ombao, H.C., Raz, J.A., Von Sachs, R., and Malow, B.A. (2001). "Automatic Statistical Analysis of Bivariate Nonstationary Time Series," *JASA* **96,** *543-560.*
- Rissanan, J. (1989). *Stochastic Complexity in Statistical Inquiry*. Singapore: World Scientific.