

Apply machine learning to Performance trend analysis

Araya Eamrurksiri

March 28, 2017

- 1 Recall: Thesis objectives
- 2 Markov switching model
 - Markov switching autoregressive model
 - Model estimation
- 3 What has been done?
- 4 Next step

Objectives

- Detect the state of the CPU utilization (degrading, improving or steady state)
- Detect whether there is any change in the test environment that effects the CPU utilization

Markov switching model, [Hamilton, 1989]

- A technique uses for describing the evolution of the process at different period of time
- Model involves multiple structures that can characterize the time series behaviors in different states
- The switching mechanism between the states is assumed to be an unobserved Markov chain - a stochastic process which contains the probability of transition from one state to any other state

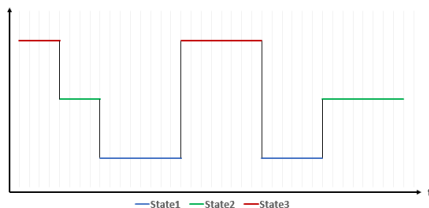


Figure: regime shift between states

Markov switching model, [Hamilton, 1989]

Assuming that S_t denote an unobservable state variable

$$y_t = X_t' \beta_{S_t} + \varepsilon_t, \quad \varepsilon_t \sim N(0, \sigma_{S_t}^2)$$

y_t is the observed value of time series at time t

X_t are the predictor variables of time series at time t

β_{S_t} are the coefficients in state S_t , where $S_t = 1, 2, \dots, k$

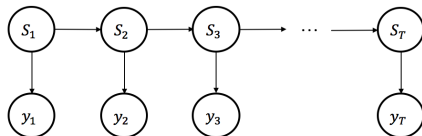


Figure: Model structure

Markov switching model

Given dataset,

$$y_t = X_t' \beta_{S_t} + \varepsilon_t, \quad \varepsilon_t \sim N(0, \sigma_{S_t}^2)$$

- y_t is the CPU utilization
- X_t are components which have an impact on the CPU utilization
- Assume there are three states ($k = 3$): normal, good, bad

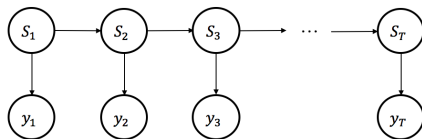


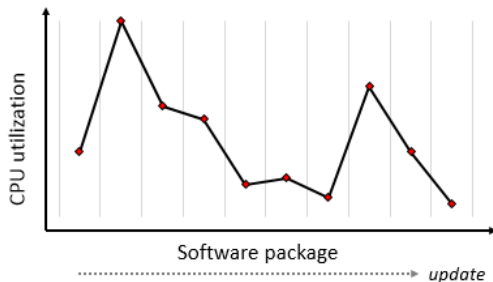
Figure: Model structure

Markov switching autoregressive model

Autoregressive model

$$y_t = c + \sum_{i=1}^p \phi_p y_{t-i} + \varepsilon_t, \quad \varepsilon_t \sim N(0, \sigma^2)$$

where c is constant and ϕ_p are parameters



Markov switching autoregressive model

The observation are drawn from the first order autoregressive model, AR(1).

$$y_t = X_t' \beta_{S_t} + \phi_{1,S_t} y_{t-1} + \varepsilon_t$$

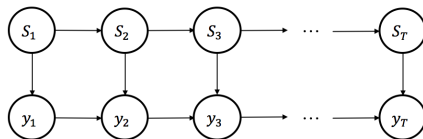


Figure: Model with additional dependencies at observation level

Model Likelihood

$$L(\theta; y_t) = f(y_t|\theta) = \sum_{t=1}^T \sum_{j=1}^k f(y_t|S_t = j; \theta)P(S_t = j)$$

and S_t is non-observable variable

A weighted average of the likelihood function in each state where weights are given by state's probabilities.

What has been done?

One software product \Rightarrow many software packages

One software package \Rightarrow many different types of test cases

Data preprocessing

- Select a test case which has a minimum value of CPU utilization for each software package
- Multiple values separated by a tab character are stored together in column \Rightarrow split a tab-separated values to columns
- Remove incomplete test cases which are not executed properly

ID	Variable1	Variable2	ID	Variable1	A	B	C
1	X	A=2 B=1 C=5	1	X	2	1	5
2	Y	A=4 B=2 C=8	2	Y	4	2	8
3	Z	A=1 C=6	3	Z	1	0	6
.
.
.

Figure: Data example

What has been done?

Study and review source code in the R package in detail

- MSwM: An univariate autoregressive Markov switching model for linear and generalized model by using the E-M algorithm [Sanchez-Espigares, 2014]

What has been done?

Implement and modify code in the package

- Small typo in the code when computing residual variance
- Solve non-invertible Hessian using generalized inverse procedure [Gill, 2004]
- Extension for categorical predictor variables
- Deal with NAs coefficients

What has been done?

Implement and modify code in the package

- Deal with NAs coefficients

A function first initial coefficients with random subsets

- Continuous variable: contains same value in all observations

Solution: Reshuffle data

- Categorical variable: not contain all levels of variable

conditional means for each state: $\hat{y} = X\hat{\beta}$

\Rightarrow NAs

Solution: Remove variables with NAs coefficient before performing matrix multiplication

What has been done?

Next step

References



James D Hamilton (1989)

A new approach to the economic analysis of nonstationary time series and the business cycle

Econometrica: Journal of the Econometric Society, pages 357-384.



Jeff Gill and Gary King (2004)

What to do when your hessian is not invertible: Alternatives to model respecification in nonlinear estimation

Sociological methods & research, 33(1):54-87.



Josep A. Sanchez-Espigares and Alberto Lopez-Moreno (2014)

MSwM: Fitting Markov Switching Models

CRAN R.