3. Methods

In this chapter, methods used for performing the change point analysis in this thesis are explained. It first starts by providing general details about Markov chains. Later on, the simple Markov switching model feature and more general model specifications namely Markov switching autoregressive model are discussed. Next three sections are devoted to some methods for estimating the values of parameters, predicting a state for a new observation and selecting a suitable model for the datasets. Another change point method in a non-parametric approach is described. Finally, a simulation technique is explained.

3.1. Markov chains

A Markov chain is a random process which has a property that given on the current value, the future is independent of the past. A random process X contains random variables $X_t : t \in T$ indexed by a set of T where $T = \{0, 1, 2, ...\}$ is called a discrete-time process and $T = [0, \infty)$ is called a continuous-time process. Let X_t be a sequence which has values from a state space S. The process begins from one of these states and moves to another state. The move between state is called a step. The process of Markov chains is described here.

Definition 3.1.1. (Grimmett and Stirzaker, 2001, p.214) If a process X satisfies the Markov property, the process X is a first order Markov chain

$$P(X_t = s|X_0 = x_0, X_1 = x_1, ..., X_{t-1} = x_{t-1}) = P(X_t = s|X_{t-1} = x_{t-1})$$

where $t \geq 1$ and $s, x_0, ..., x_{t-1} \in S$

If $X_t = i$ then it is said that the chain is being in state i or the chain is in the ith state at the tth step.

There are transitions between states which describe the distribution for the next state given the current state. This evolution of changing from $X_t = i$ to $X_t = j$ is defined by the transition probability as $P(X_t = j|X_{t-1} = i)$. Markov chains are frequently assumed that these probabilities depend only on i and j and do not depend on t.

Definition 3.1.2. (Grimmett and Stirzaker, 2001, p.214) The chain is time-homogeneous if

$$P(X_{t+1} = j|X_t = i) = P(X_1 = j|X_0 = i)$$

for all t, i, j. The probability of the transition is independent of t. Then, the transition matrix $\mathbf{P}$ is the matrix of transition probabilities

$$p_{ij} = P(X_t = j|X_{t-1} = i)$$

Theorem. *(Grimmett and Stirzaker, 2001, p.215) The transition matrix* $\mathbf{P}$ *is a stochastic matrix that*

- *Each of the entries is a non-negative real number or* $p_{ij} \geq 0$ *for all* i, j
- *The sum of each column equal to one or* $\sum_i p_{ij} = 1$ *for all* j

Definition 3.1.3. (Grimmett and Stirzaker, 2001, p.222) The mean recurrence time of a state i is defined as

$$\mu_i = E(T_i|X = 0 = i) = \sum_n n \cdot f_{ii}(n)$$

State i is positive recurrent or non-null persistent if μ_i is finite. Otherwise, the state i is null persistent.

Definition 3.1.4. (Grimmett and Stirzaker, 2001, p.222) A state i has the period $d(i)$ and is defined as

$$d(i) = gcd\{n : \ p_{ii}(n) > 0\}$$

where gcd is the greatest common divisor. If $d(i) = 1$, then the state is said to be aperiodic. Otherwise, the state is said to be periodic.

Definition 3.1.5. (Grimmett and Stirzaker, 2001, p.222) A state is called ergodic if it is non-null persistent and aperiodic.

Definition 3.1.6. A chain is called irreducible if all states are eventually reached from any state.

Definition 3.1.7. If there is a finite state space, an irreducible Markov chain is the same thing as ergodic Markov chain. It is possible to go from every state to every other state with positive probability.

3.2. Markov switching model

A Markov switching model is used for time series that are evolved over unobserved distinct states or regimes. This is a regime switching model where the shifting back and forth between the regimes is controlled by a latent Markov chain. The model structure consists of two stochastic processes embedded in two levels of hierarchy. One process is an underlying stochastic process that is not observable but it is possible to observe them through another stochastic process which generated the sequence of observation (Rabiner and Juang, 1986). The time of transition to different state and the duration in between is random. In addition, the state assumes to follow the Markov property that the future state depends only on the current state.

The Markov switching model is able to model more complex stochastic processes and describes changes in the dynamic behavior. A general structure of the model can be drawn in graphically as shown in Figure 3.1, where S_t and y_t denote the state sequence and observation sequence in the Markov process, respectively. The arrows from one state to another state in the diagram implied the conditional dependency.

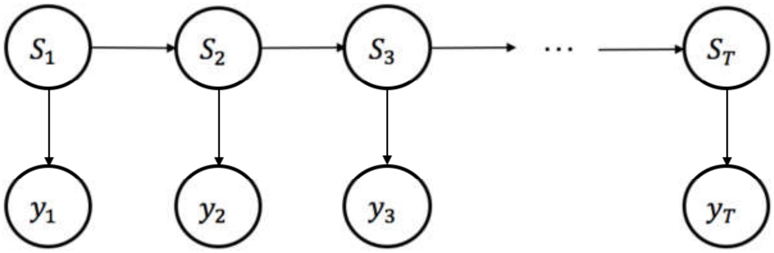

Figure 3.1.: Model structure

The process is given by (Hamilton, 1989)

$$y_t = X_t' \beta_{S_t} + \varepsilon_t \tag{3.1}$$

where y_t is the observed value of the time series at time t

X_t are the predictor variables of the time series at time t

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

ε_t follows a Normal distribution with mean zero and variance given by $\sigma^2_{S_t}$

The Equation 3.1 is the simplest form for the switching model where there are $k-1$ structural breaks in the model parameters. To aid understanding, the baseline model assuming two states ($k = 2$) is discussed. S_t is a random variable which is assumed that the value $S_t = 1$ for $t = 1, 2, ..., t_0$ and $S_t = 2$ for $t = t_0 + 1, t_0 + 2, ..., T$ where t_0 is a known change point. The transition matrix **P** is an 2x2 matrix where row

j column i element is the transition probability p_{ij}. Since the whole process S_t is unobserved, the initial state where $t = 0$ of each state also needs to be specified. The probability which describes the starting distribution over states is denoted by

$$\pi_i = P(S_0 = i)$$

There are several options for computing the probability of the initial state. One procedure is to simply use a naive guess i.e., setting $P(S_0 = i) = 0.5$. Alternatively, the unconditional probability of S_t

$$\pi_1 = P(S_0 = 1) = \frac{1 - p_{jj}}{2 - p_{ii} - p_{jj}}$$

can be used by presuming an ergodic Markov chain (Hamilton, 2005).

3.3. Autoregressive (AR) model

An autoregressive model is one type of time series model that uses for describing the time-varying process. The model is flexible in handling various kinds of time series patterns. The name autoregressive comes from how the model performs a regression of the variable against its own previous outputs (Cryer and Kellet, 1986). The number of autoregressive lags is denoted by p.

Definition 3.3.1. An autoregressive model of order p or AR(p) model can be written as

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + ... + \phi_p y_{t-p} + \varepsilon_t$$

or

$$y_t = c + \sum_{i=1}^{p} \phi_i y_{t-i} + \varepsilon_t$$

where c is a constant, ϕ_i are the coefficients in the autoregression and ε_t is Gaussian white noise with zero mean and variance σ^2.

If p is equal to one, the model AR(1) is called the first order autoregression process.

3.4. Markov switching autoregressive model

This model is an extension of the basic Markov switching model where observations are drawn from autoregression process. Markov switching autoregressive model relaxes the conditional independent assumption by allowing an observation to also depends on a past observation and a current state (Shannon and Byrne, 2009).

Definition 3.4.1. The first order Markov switching autoregressive model is

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

where ϕ_{1,S_t} is the autoregression coefficient of the observed value at time $t-1$ in state S_t and ε_t follows a Normal distribution with mean zero and variance given by $\sigma^2_{S_t}$.

The structure of the model is shown in Figure 3.2. It can be clearly seen that there is a dependency at the observation level and the observation is not independent from one another.

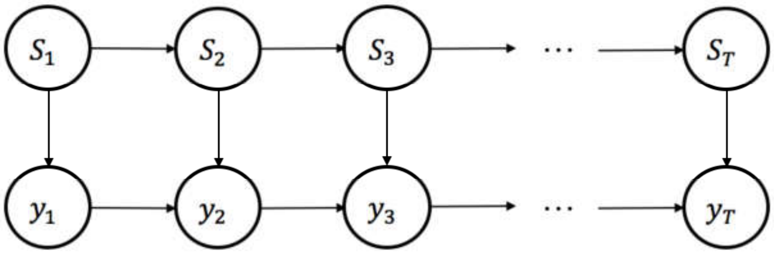

Figure 3.2.: Model structure of Markov switching AR(1)

Assuming two states $S_t = 1$ or 2, the set of parameters for the Markov switching autoregressive model that are necessary for describing the law of probability governs y_t are $\theta = \{\beta_1, \beta_2, \phi_{1,1}, \phi_{1,2}, \sigma_1^2, \sigma_2^2, \pi_1, \pi_2, p_{11}, p_{22}\}$.

3.5. Parameter estimation

There are various ways to estimate parameters of Markov switching model. Methods which have been widely used are as follow: E-M algorithm (Hamilton, 1990; Kim, 1994) used the maximum likelihood criterion, Segmental K-mean (Juang and Rabiner, 1990) used K-means algorithm and maximized the state-optimized likelihood criterion, and Gibbs sampling (Kim et al., 1999) used a Markov chain Monte Carlo simulation method based on the Bayesian inference. In this thesis framework, E-M algorithm is used in estimating parameters and is briefly described a general procedure.

3.5.1. The Expectation-Maximization algorithm

E-M algorithm is originally designed to deal with the incomplete or missing values in data (Dempster et al., 1977). Nevertheless, it can potentially implement in Markov switching model since the unobserved state S_t can be viewed as the missing values. The set of parameters is estimated by iterative two-step procedure. The algorithm starts with an arbitrary initial parameters and finds the expected values of the state process given the observations. Next, the new maximum likelihood from the derived parameters in previous step is calculated. These two steps are repeated until the maximum value of the likelihood function is reached (Janczura and Weron, 2012).

3.5.1.1. E-step

Assume that $\theta^{(n)}$ is the derived set of parameters in M-step from the previous iteration and the available observations of time $t-1$ is denoted as $\Omega_{t-1} = (y_1, y_2, ..., y_{t-1})$. The general idea of this step is to calculate the expectation of S_t under the current estimation of the parameters. The obtained result is called smoothed inferences probability and is denoted by $P(S_t = j|\Omega_T; \theta)$. The E-step consists of filtering and smoothing algorithm and the process is described as follows (Kim, 1994):

Filtering Filtered probability is the probability of the non-observable Markov chain being in a given state j at time t, conditional on information up to time t. The algorithm starts from $t = 1$ to $t = T$. The starting points for the first iteration where $t = 1$ is chosen as arbitrary values. The probabilities of each state given that the available observation is up to time $t - 1$ is calculated.

$$P(S_t = j|\Omega_{t-1}; \theta^{(n)}) = \sum_{i=1}^{k} p_{ij}^{(n)} P(S_{t-1} = i|\Omega_{t-1}; \theta^{(n)}) \tag{3.2}$$

and the conditional densities of y_t given Ω_{t-1} are

$$f(y_t|\Omega_{t-1}; \theta^{(n)}) = \sum_{j=1}^{k} f(y_t|S_t = j, \Omega_{t-1}; \theta^{(n)}) P(S_t = j|\Omega_{t-1}; \theta^{(n)}) \tag{3.3}$$

where $f(y_t|S_t = j, \Omega_{t-1}; \theta) = \frac{1}{\sqrt{2\pi\sigma_{S_t}^2}} exp\left\{-\frac{(y_t-\beta_{S_t})^2}{2\sigma_{S_t}^2}\right\}$ is the likelihood function in each state for time t. This is simply the Gaussian probability density function.

Then, with the new observation at time t, the probabilities of each state are updated by using Bayes' rule

$$P(S_t = j|\Omega_t; \theta^{(n)}) = \frac{f(y_t|S_t = j, \Omega_{t-1}; \theta^{(n)}) P(S_t = j|\Omega_{t-1}; \theta^{(n)})}{f(y_t|\Omega_{t-1}; \theta^{(n)})} \tag{3.4}$$

It is computing iteratively until all the observation is reached i.e., $t = T$.

Smoothing Smoothed probability is the probability of the non-observable Markov chain being in state j at time t, conditional on all available information. The algorithm iterates over $t = T-1, T-2, ..., 1$. The starting values are obtained from the final iteration of the filtered probabilities.

By nothing that

$$\begin{aligned} P(S_t = j|S_{t+1} = i, \Omega_T; \theta^{(n)}) &\approx P(S_t = j|S_{t+1} = i, \Omega_t; \theta^{(n)}) \\ &= \frac{P(S_t = j, S_{t+1} = i|\Omega_t; \theta^{(n)})}{P(S_{t+1} = i|\Omega_t; \theta^{(n)})} \\ &= \frac{P(S_t = j|\Omega_t; \theta^{(n)}) p_{ij}^{(n)}}{P(S_{t+1} = i|\Omega_t; \theta^{(n)})} \end{aligned} \tag{3.5}$$

and

$$P(S_t = j|\Omega_T; \theta^{(n)}) = \sum_{i=1}^{k} P(S_t = j, S_{t+1} = i|\Omega_T; \theta^{(n)}) \tag{3.6}$$

Then, the smoothed probabilities can be expressed as

$$P(S_t = j|\Omega_T; \theta^{(n)}) = \sum_{i=1}^{k} \frac{P(S_{t+1} = i|\Omega_T; \theta^{(n)}) P(S_t = j|\Omega_t; \theta^{(n)}) p_{ij}^{(n)}}{P(S_{t+1} = i|\Omega_t; \theta^{(n)})} \tag{3.7}$$

Once the filtered probabilities are estimated, there is necessarily enough information to calculate the full log-likelihood function.

$$\ln L(\theta) = \sum_{t=1}^{T} \ln(f(y_t|\Omega_{t-1}; \theta^{(n)}) = \sum_{t=1}^{T} \ln \sum_{j=1}^{k} ((f(y_t|S_t = j, \Omega_{t-1}; \theta^{(n)}) P(S_t = j|\Omega_{t-1})) \tag{3.8}$$

This is simply a weighted average of the likelihood function in each state. The probabilities of states are considered as weights.

3.5.1.2. M-step

The new estimated model parameters $\theta^{(n+1)}$ is obtained by finding the set of parameters that maximizes the Equation 3.8. This new set of parameters is more exact value of the maximum likelihood estimates than the previous one. It serves as the set of parameters in the next iteration of the E-step. The estimated parameters are derived by taking the partial derivative of the log-likelihood function with respect to the specific parameter and then setting it to zero. Generally, this process is similar to the standard maximum likelihood estimation except that it has to be weighted by the smoothed probabilities since each observation y_t carries probability of coming from any of the k state.

3.6. State prediction

A function to predict the most probable state for the new observation is implemented in this analysis (see Appendix B).

The probabilities of being in state j at time $T+1$ on the basis of current information are computed by performing the filtering algorithm in the E-step of E-M algorithm. The filtered probabilities are

$$P(S_{T+1} = j|\Omega_{T+1};\theta) = \frac{f(y_{T+1}|S_{T+1} = j, \Omega_T;\theta)P(S_{T+1} = j|\Omega_T;\theta)}{f(y_{T+1}|\Omega_T;\theta)}$$

This is the Equation 3.4 where $t = T+1$. Then, the new observation at time $T+1$ is said to be in the state j if it has the highest probability.

3.7. Model selection

Model selection is a task of selecting the most suitable model for a given set of data based on the quality of the model. In this thesis framework, the Bayesian Information Criterion (BIC) is widely employed in the applied literature and proved to be useful in selecting the model among a finite set of models. It is also known as Schwarz Information Criterion (Schwarz et al., 1978).

$$\mathrm{BIC} = -2\ln(L(\hat{\theta})) + m \cdot \ln(T)$$

where $L(\hat{\theta})$ represents the maximized value of the likelihood function, T is the number of observations and m is the number of parameters to be estimated in the model. One benefit from using BIC is that this criterion heavily penalizes model complexity as it takes into account the number of parameters in the model. BIC attempts to reduce the risk of over-fitting.

3.8. Non-parametric analysis

The parametric test statistics outperform the non-parametric test if data belongs to some known distribution families. However, the parametric test is not properly performing well in detecting change point for an unknown underlying distribution (Sharkey and Killick, 2014). Applying the non-parametric analysis to the real-world process gives a real advantage to the analysis since data collected from application, in general, does not always have a well-defined structure and prefer such analysis that is not too restricted (Hawkins and Deng, 2010). For this reason, the non-parametric analysis is implemented in order to get a heuristic idea of the change point location in this thesis framework. The obtained result is also used for comparing with the result from the Markov switching autoregressive model.

E-divisive

The *ecp*[1] is an extension package in R which mainly focuses on computing the non-parametric test for multiple change point analysis. This change point method is applicable to both univariate and multivariate time series. A fundamental idea for the method is based on the hierarchical clustering approach (James and Matteson, 2013).

The E-divisive method is an algorithm in the *ecp* package. This algorithm performs a divisive clustering in order to estimate the multiple change points. The E-divisive recursively partitions a time series and estimates a single change point in each iteration. Consequently, the new change point is located at each iteration and it divides the time series into different segments. The algorithm also used a permutation test to compute the statistical significance of an estimated change point. More details about the estimation is described on Matteson and James (2014).

3.9. Simulation

Since there is no annotation for the state of the CPU utilization in the data, an accuracy can not be computed after making a state prediction. One possible solution to test and verify how well the implemented predict function performs is to use the simulation technique. Therefore, a data consists of two predictor variables and one response variable with the known state is simulated. The actual models for each state are

$$y = \begin{cases} 10 + 0.6X1_t - 0.9X2_t + 0.5Y_{t-1} + \varepsilon_t^{(1)} & \varepsilon_t^{(1)} \sim N(0,1) \\ 2 + 0.8X1_t + 0.2Y_{t-1} + \varepsilon_t^{(2)} & \varepsilon_t^{(2)} \sim N(2,0.5) \\ -12 + 0.7X1_t + 0.2Y_{t-1} + \varepsilon_t^{(3)} & \varepsilon_t^{(3)} \sim N(1,1) \end{cases}$$

[1] https://cran.r-project.org/web/packages/ecp/index.html

The simulated data contains 500 observations which has a presence of three different states – Normal, Bad and Good. Figure 3.3 presents a plot of y over a period of time and the period where observations in the data belong to one of the state is shown in Figure 3.4.

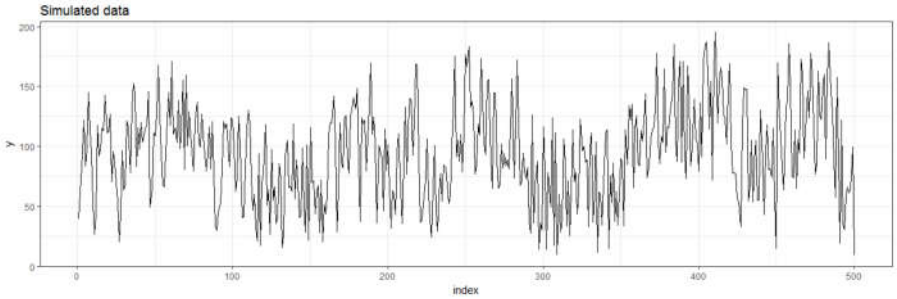

Figure 3.3.: Simulated data. The y variable is the response variable

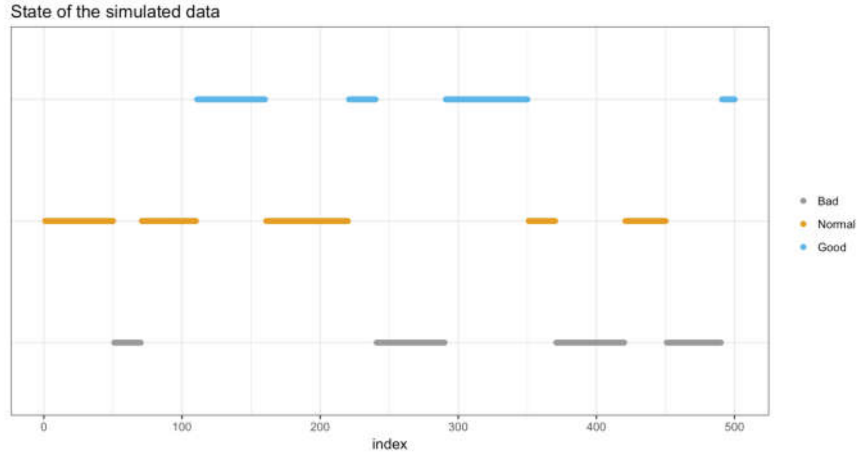

Figure 3.4.: The period in the time series when data is in each state