3. Methods

This chapter first starts by providing a survey of existing methods that address the problem of detecting changes in a system. Later on, general information about Markov chains, the simple Markov switching model feature, and model specification namely Markov switching autoregressive model are discussed. Thereafter, three sections are devoted to 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, the simulation technique is explained.

3.1. Survey of existing methods

Change point detection, Anomaly detection, Intrusion detection, or *Outlier detection* are terms that are closely related to one another. The main idea of these terms is to identify and discover events that are abnormal from the usual behavior. There are several methods to address these types of problem. A survey of existing methods has been done in the thesis, and some methods are presented in this section.

Valdes and Skinner (2000) employed a Bayesian inference technique, specifically a naive Bayesian network, to create an intrusion detection system on traffic bursts. Even though the Bayesian network is effective in detecting anomalies in some applications, there are some limitations that should be considered when using this method. As accuracy of a detection system depends heavily on certain assumptions, the system will have low accuracy if an inaccurate model is implemented (Patcha and Park, 2007).

Support vector machine (SVM) introduced in Cortes and Vapnik (1995) is a supervised learning algorithm to deal with a classification analysis problem by using the idea of separating hyperplanes. The main reason that SVM is used in anomaly detection is because of its speed and scalability (Sung and Mukkamala, 2003). Although this method is effective in identifying new kinds of anomalies, the method often has a higher rate of false alarms due to the fact that the SVM method ignores the relationships and dependencies between the features (Sarasamma et al., 2005).

Self-organizing maps (SOM) developed by Kohonen (1982) is a well-known unsupervised neural network approach for cluster analysis. SOM is efficient in handling large and high dimensional datasets. Nousiainen et al. (2009) used SOM for an anomaly detection in a server log data. The study presented an ability of the SOM method in detecting anomalies in the data, and also compared the results from the SOM method with a threshold based system. A disadvantage of the SOM is that initial weight vector affects a performance of the SOM, which leads to an unstable clustering result. Besides, if the anomalies in the data tend to form clusters, this method will not be able to detect these anomalies (Chandola et al., 2009).

Based on previous works, the Hidden Markov model or the Markov switching model has also been used in identifying changes and anomalies. One drawback from the method based on the Markov chain is that the method has a high computational cost, which is not scalable for an online change application (Patcha and Park, 2007). Apart from changes that can be detected in the data, some knowledge about the unobservable condition of the system can also be obtained. This additional information makes the method more appealing than the other methods. Therefore, the Markov switching model is implemented in this thesis framework.

3.2. Technical aspects

The thesis work has been carried out using the *R* programming language (R Core Team, 2014) for the purpose of data cleaning, preprocessing, and analysis. The Markov switching model was performed using the *MSwM* package. Various extensions and modifications were further implemented in the package e.g., handling predictor categorical variables, the state prediction function, and plots for visualizing the results (More details can be found in sec. A.2). For the E-divisive method, the *ecp* package was used.

3.3. Markov chains

A Markov chain is a random process which has a property that is given 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 T. When $T = \{0, 1, 2, ...\}$ the process is called a discrete-time process, and when $T = [0, \infty)$ it is called a continuous-time process. Let X_t be a sequence of values from a state space S. The process begins from one of these states and moves to another state. The move between states is called a step. The process of Markov chains is described here.

Definition 3.3.1. (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 the chain is in state i or the chain is in the ith state at the tth step.

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

Definition 3.3.2. (Grimmett and Stirzaker, 2001, p.214) A Markov 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. A *transition matrix* $\mathbf{P} = (p_{ij})$ is a matrix of transition probabilities

$$p_{ij} = P(X_t = j|X_{t-1} = i) \qquad \text{for all}\, t, i, j$$

Theorem. *(Grimmett and Stirzaker, 2001, p.215) The transition matrix* $\mathbf{P}$ *is a 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 row equal to one or* $\sum_j p_{ij} = 1$ *for all* i

Definition 3.3.3. (Grimmett and Stirzaker, 2001, p.227) The vector π is called a *stationary distribution* if π has entries $(\pi_j : j \in S)$ that satisfies

- $\pi_j \geq 0$ for all j, and $\sum_j \pi_j = 1$
- $\pi = \pi\mathbf{P}$, which is $\pi_j = \sum_i \pi_i p_{ij}$ for all j

Definition 3.3.4. (Grimmett and Stirzaker, 2001, p.220) A state i is called persistent (or recurrent) if

$$P(X_t = i\, \text{for some t} \geq 1|X_0 = i) = 1$$

Let $f_{ij}(t) = P(X_1 \neq j, X_2 \neq j, ..., X_t = j|X_0 = i)$ be the probability of visiting state j first by starting from i, takes place at tth step.

Definition 3.3.5. (Grimmett and Stirzaker, 2001, p.222) The mean recurrence time of a persistent state i is defined as

$$\mu_i = E(T_i|X_0 = i) = \sum_n n \cdot f_{ii}(n)$$

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

Definition 3.3.6. (Grimmett and Stirzaker, 2001, p.222) The period $d(i)$ of a state i 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 3.3.7. (Grimmett and Stirzaker, 2001, p.222) A state is called ergodic if it is non-null persistent and aperiodic.

Definition 3.3.8. A chain is called irreducible if it is possible to go from every state to every other states.

Theorem. *If all states in an irreducible Markov chain are ergodic, the chain is said to be ergodic.*

Theorem. *(Manning et al., 2008) If there is an aperiodic finite state, an irreducible Markov chain is the same thing as an ergodic Markov chain.*

3.4. Markov switching model

A Markov switching model is a switching model where the shifting back and forth between the states or 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 normally observable, but possible to be observed through another stochastic process which generates the sequence of observation (Rabiner and Juang, 1986). The transition time between two states is random. In addition, the state are assumed 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 describe changes in the dynamic behavior. A general structure of the model can be drawn 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 implies a conditional dependency.

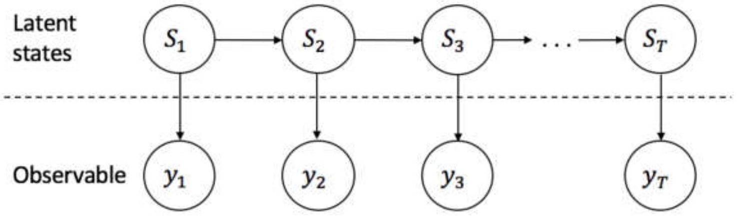

Figure 3.1.: Model structure

The process is given by (Hamilton, 1989)

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

where,

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

X_t is a design matrix, also known as model matrix, containing values of predictor variables of the time series at time t

β_{S_t} are a column vector of coefficients in state S_t, where $S_t \in \{1, ..., k\}$

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

Equation 3.1 is the simplest form for the switching model. To aid understanding, the baseline model is assumed to have only two states ($k = 2$) in this discussion. 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}. A diagram showing a state-transition is shown in Figure 3.2. Note that these probabilities are independent of t.

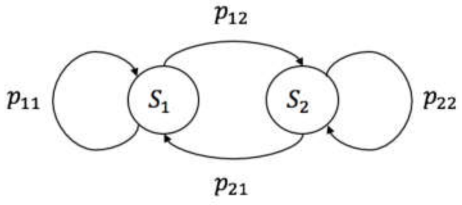

Figure 3.2.: State-transition diagram

Since the whole process S_t is unobserved, the initial state where $t = 0$ 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 commonly set $P(S_0 = i) = 0.5$. Alternatively, by presuming an ergodic Markov chain (Hamilton, 2005), a stationary distribution is

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

which is simply from solving the system of equations $\pi = \pi\mathbf{P}$.

Proof. Let $\pi = (\pi_1, \pi_2)'$ and $\mathbf{P} = \begin{bmatrix} p_{ii} & 1 - p_{ii} \\ 1 - p_{jj} & p_{jj} \end{bmatrix}$
Definition 3.3.3 of a stationary distribution,

$$\pi = \pi\mathbf{P} \tag{3.2}$$

and

$$\pi_1 + \pi_2 = 1 \tag{3.3}$$

From 3.2,

$$\begin{aligned} \pi_1 &= \pi_1 p_{ii} + \pi_2(1 - p_{jj}) \\ \pi_2 &= \pi_1(1 - p_{ii}) + \pi_2 p_{jj} \end{aligned}$$

Therefore,

$$\pi_2 = \frac{\pi_1(1 - p_{ii})}{1 - p_{jj}} \tag{3.4}$$

Substitute 3.4 into Equation 3.3, then

$$\pi_1 = \frac{1 - p_{jj}}{2 - p_{ii} - p_{jj}}$$

□

A coefficient of a predictor variable in the Markov switching model can have either different values in different state or a constant value in all state. The variable which have the former behavior is said to have a *switching effect*. Likewise, the variable which have the same coefficient in all states is the variable that does not have a switching effect, or said to have a *non-switching effect*.

A generalized form of Equation 3.1 can be defined as (Perlin, 2015)

$$y_t = X_t^{ns}\alpha_t + X_t^{s}\beta_{S_t} + \varepsilon_t \tag{3.5}$$

where,

X_t^{ns} contains all predictor variables that have non-switching effect of the time series at time t

α_t are non-switching coefficients of the time series at time t

X_t^{s} contains all predictor variables that have the switching effect of the time series at time t

β_{S_t} are switching coefficients in state S_t, where $S_t \in \{1, ..., k\}$

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

3.4.1. Autoregressive (AR) model

An autoregressive model is one type of time series models used to describe a 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 (i.e., the number of prior values used in the model) is denoted by p.

Definition 3.4.1. An autoregressive model of order p or AR(p) model can be written as

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

where c is a constant, ϕ_i are coefficients in the autoregression and ε_t follows a normal distribution with zero mean and variance σ^2.

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

3.4.2. Markov switching autoregressive model

A Markov switching autoregressive model is an extension of a basic Markov switching model where observations are drawn from an autoregressive process. The model relaxes the conditional independence assumption by allowing an observation to depend on both past observation and a current state (Shannon and Byrne, 2009). Basically, this is the combination between the Markov switching model and the autoregressive model.

Definition 3.4.2. The first order Markov switching autoregressive model is

$$y_t = X_t \beta_{S_t} + \phi_{1,S_t} y_{t-1} + \varepsilon_t$$