People's Democratic Republic of Algeria
Ministry of Higher Education and Scientific Research

University M'Hamed BOUGARA – Boumerdes

Institute of Electrical and Electronic Engineering

Department of Power and Control

Final Year Project Report Presented in Partial Fulfilment of
the Requirements for the Degree of

MASTER

In **Control**

Option: **Control**

Title:

Hidden Markov Model-based Approach for Process Monitoring

Presented by:

- **BENABDALLAH Mounir**
- **LOUNAOUCI Mohamed Lamine**

Supervisor:

Dr. KOUADRI Abdelmalek

Registration Number:……/2019

Dedication

It is with immense gratitude and joy that I am dedicating this work to my beloved parents who have devotedly supported and encouraged me through my academic life. To my brother and sisters for believing in my abilities. To my uncle Nacereddine and my dear friend Ilyas who have been always my inspiration and support.

BENABDALLAH Mounir

ABSTRACT

Hidden Markov Models (HMMs) are a popular and ubiquitous tool for modelling a large range of time series data. It has been applied successfully to various complex problems, being especially effective with those requiring a huge amount of measured data, such as pattern recognition in speech, handwriting and even facial recognition. Since Fault detection and diagnosis is an important problem in process engineering recent studies are focusing on developing new techniques which are more accurate, sensitive to small faults, with no time delay and can monitor multi-mode process effectively. In order to satisfy these requirements, a huge data are needed and a suitable model to process these data is the HMM. The main objective of this work is to develop novel HMM-based approach to diagnose various operating modes of a process including Bayesian methods for mode selection. The mode in this work refers to process operational statuses such as normal or abnormal operating conditions.

Keywords: *Thesis*, *Fault Detection*, *Hidden Markov Model*, *Operating Mode*, *Multi-Mode Process*.

ACKNOWLEDGEMENT

With immense pleasure and deep sense of gratitude, we wish to express our sincere thanks to our supervisor **Dr. KOUADRI Abdelmalek**, with Signals and Systems Laboratory, Institute of Electrical and Electronic Engineering, University MHamed BOUGARA Boumerdes, without his motivation and continuous encouragement and assistance, this research would not have been successfully completed. His guidance was of a great help in all the time of research and writing of this thesis.

We wish to extend our profound sense of gratitude to **Mr. BENCHEIKH Fares** for willingly sharing his valuable knowledge and assistance during our research and also providing us with moral support and encouragement whenever required.

Last but not the least, We would like to thank **Our friends** for their constant encouragement and moral support with patience and understanding.

Place: Boumerdes

Date: 12/06/2019

BENABDALLAH Mounir

LOUNAOUCI Med Lamine

CONTENTS

List of Figures

List of Tables

LIST OF TERMS AND ABBREVIATIONS

LIST OF SYMBOLS

α_t The auxiliary forward variable

β_t The auxiliary backward variable

δ_t The highest probability of a path at time t

$\hat{P} \in R^{m \times l}$ Principal Loadings Matrix

$\hat{T} \in R^{n \times l}$ Matrix of principal scores, also written T_l

Λ Eigenvalues Diagonal Matrix

λ A model with given parameters

λ_i Singular value

μ_x The expected value (mean) of the random variable X

π_i initial state distribution

σ_x The standard deviation of the random variable X

A Transition probability matrix

a_{ij} Probability of transition from state i to j

B Emission probability matrix

$b_j(k)$ Probability of observing symbol k at state j

$diag(\lambda)$ The diagonal elements of the matrix λ, or a diagonal matrix with the vector λ forms its diagonal and the off-diagonal values are zero.

f Fault magnitude

l Number of retained components by PCA model

M Number of observations in the model

N Number of states in the model

O Set of discrete observations

O_t Discrete observation at time t

$P \in R^{n \times m}$ Loadings Projection Matrix

q_t State at time t

S_t State at time t

$T \in R^{n \times m}$ Scores or Features Matrix

$X \in R^{n \times m}$ Original data matrix with n observations and m variables

X_n Normalized data matrix

CHAPTER 1

General Introduction

In the process and manufacturing industries, there has been a large push to produce higher quality products, to reduce product rejection rates, and to satisfy increasingly critical safety and environmental regulations.

Process operations that were at one time considered acceptable are no longer adequate. To meet the higher standards, modern industrial processes contain a large number of variables operating under closed-loop control. The standard process controllers (PID controllers, model predictive controllers, ...etc) are designed to maintain satisfactory operations by compensating for the effects of disturbances and changes occurring in the process. While these controllers can compensate for many types of disturbances, there are changes in the process which the controllers cannot handle adequately. These changes are called faults [3].

More precisely, a fault in a system is defined as ”an unpermitted deviation of at least one characteristic property (feature) of the system from acceptable, usual and standard condition” [4]. Accordingly, faults might end up to a complete failure of the system. In chemical processes, such process shutdowns will cause product loss. Moreover, it takes extensive time and effort to return the process back to its normal operation.

A fault diagnosis problem consists of three main parts:

1. Fault Detection: to determine if the process is operating out of its normal conditions and decide on the presence of faults.

2. fault Isolation: to localize the fault and find the component which is the main cause of the fault.

3. Fault Identification: to identify the fault in terms of magnitude and other detailed properties [5].

All process monitoring techniques are based on available on-line observations from sensor measurements. The core idea is to find some underlying relation between observations and possible faults. The main task of Fault Detection system is to determine whether there are faults. In general, any fault detection subsystem cannot detect a variety of faults correctly 100 percent. Therefore, improving the correct fault detection rate and reduce the missed detection rate (which occurs without the fault is detected) and false alarm rate (not failure but police) has been the interesting topic in the area of fault detection and diagnosis [6]. In this procedure, several difficulties might occur. One example is the complexity to find model for process due to the complicated process behavior. Another example is the extraction of the critical information from high dimensional data sets.

Hidden Markov Models (HMMs) are sophisticated mathematical tools for considering the temporal information in both process model and history based fault diagnosis. Incorporating time domain information, HMMs will greatly assist a classifier to reduce false alarms.

In formulation of HMMs, some sort of prior knowledge of process temporal behavior is considered in the form of transition probability matrix. One obvious advantage of such knowledge is to diagnose between process modes which have significant overlaps [7].

In this thesis, our focus is on statistical/ probabilistic classifiers. More specifically, the main effort is to more precisely consider the temporal information. Our aim is to develop a robust fault detection system with high precision and less time delay, less Missed Detection Rate (MDR), and less False Alarms Rate (FAR), while various probability distributions are considered for observations in different operating modes (healthy and faulty mode).

Another important target of this research is to apply the developed fault detection strategies under industrial environments. The developed frameworks are tested on cement plant industry.

This thesis is organized as follows:

Chapter 2, the field of Fault Detection and Diagnosis (FDD) is introduced along with its terminology, available fault diagnosis techniques including process model based and process history based, each category will be reviewed.

Chapter 3, a review on fundamental mathematical tools of the thesis is provided. Most of chapters of the thesis are based on two main concepts: HMMs and Principal Component Analysis (PCA) which is a basic method in the framework of the multivariate analysis techniques. In this study, PCA technique is used for dimension reduction. Some previous applications of Hidden Markov Model (HMM) for assisting fault diagnosis methods are summarized. Next, the proposed model is discussed and tested for a simulated example, results are shown.

Chapter 4, the proposed Hidden Markov Model based fault detection is applied for the cement plant rotary kiln of Ain-ElKebira and results are analyzed.

CHAPTER 2

Review of Literature

2.1 Introduction

After several decades of development, the formation of fault diagnosis technology generally three types of methods, analytical model-based fault diagnosis method, signal processing based fault diagnosis method and knowledge-based fault diagnosis method[6].

To address such issues, two major areas have to be introduced:

1. Process model based fault diagnosis

2. Process history based fault diagnosis

The main concerns of model-based fault detection are: first, to find an appropriate mathematical model between process inputs and outputs, and then, generate some features through analysis of residuals, parameter estimates and state estimates. Comparison between the generated features and normal features will result in some symptoms, and eventually, fault diagnosis. In order to decide on details of the type of fault, model-based methods are usually accompanied by further classification and inference layers [4].

From the previous explanation on model based approaches, it is obvious that some prior knowledge about process is required. In contrast, in history based methods, only a large data set of process historical data is needed. The underlying idea of history based methods is to extract the critical information (features) from large data sets.

The data extraction might be either qualitative or quantitative. Unlike quantitative methods which try to extract the important quantitative information, qualitative methods

are usually based on the coding of knowledge and compact representation of trends [8]. Quantitative approaches can be divided into two main subcategories including non-statistical, e.g., neural networks, and statistical.

In general, in contrast to deterministic systems, in stochastic systems it is not possible to definitely determine the future state given the current information. Consequently, it is worth viewing the system in a probabilistic manner. In such approaches, various probability distributions are considered for classification of different operating conditions of the process. Some of the well known methods in this area are statistical classifiers, e.g., Bayes classifier with Gaussian density functions, Principal Component Analysis (PCA) and Partial Least Square (PLS) [9].

A summary of available fault diagnosis strategies, as stated in previous paragraphs, is presented in Figure 1.1.

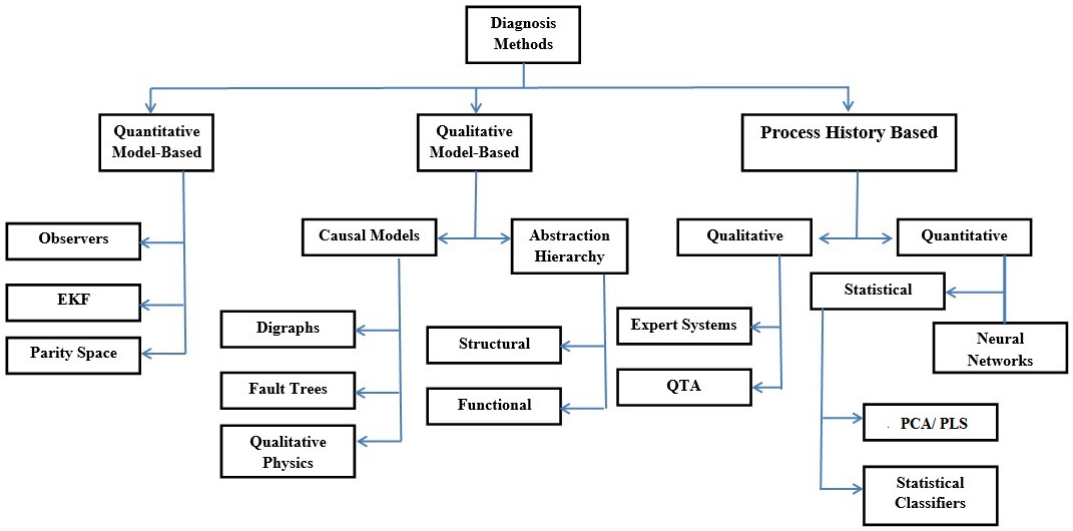

Fig. 2.1 Classification of Diagnostic Algorithms.

From this brief review on available process monitoring techniques, one could observe that all fault diagnosis techniques are based on the extracted information from observations. Among the underlying data contained in a signal, temporal information which contains the memory of operating mode transitions plays an important role. Many of the proposed methods are not robust enough to extract the temporal information from noisy signals. To deal with such an issue, one might consider feed back of outputs to the classifier or considering a window of observations. However, such

approaches will bring a lot of difficulties during the training [7].

2.2 Previous studies

As the treated field from faults and failures through reliability, safety and fault-tolerant systems is distributed over many different technological areas, the terminology used is not unique. Various efforts have been made to come to a standardization. Some useful terminology are stated in Appendix A.

The supervision of technical processes under normal operation or the quality control of products in manufacturing is usually performed by limit checking or threshold checking of some few measurable output variables $Y(t)$ like pressures, forces, liquid levels, temperatures, speeds, and oscillations. This means one checks if the quantities are within a tolerance zone $Y_{min} < Y(t) < Y_{max}$ [4] . If the tolerance zone is exceeded, an alarm is raised. Hence, the first task in supervision is:

1. Monitoring: Measurable variables are checked with regard to tolerances, and alarms are generated for the operator. The operator has to take appropriate counteractions after an alarm is triggered.

   However, if exceeding a threshold implies a dangerous process state, the counteraction should be generated automatically.

   This is a second task of supervision:

2. Automatic protection: In the case of a dangerous process state, the monitoring function automatically initiates an appropriate counteraction. Usually, the process is then commanded to a fail-safe state, which is frequently an emergency shutdown [4].

These classical methods of monitoring and automatic protection are suitable for the overall supervision of the processes. However, these methods will detect faults only after a considerable deviation of observations from expected behavior [10]. Therefore, more advanced monitoring techniques are developed for early detection of faults.

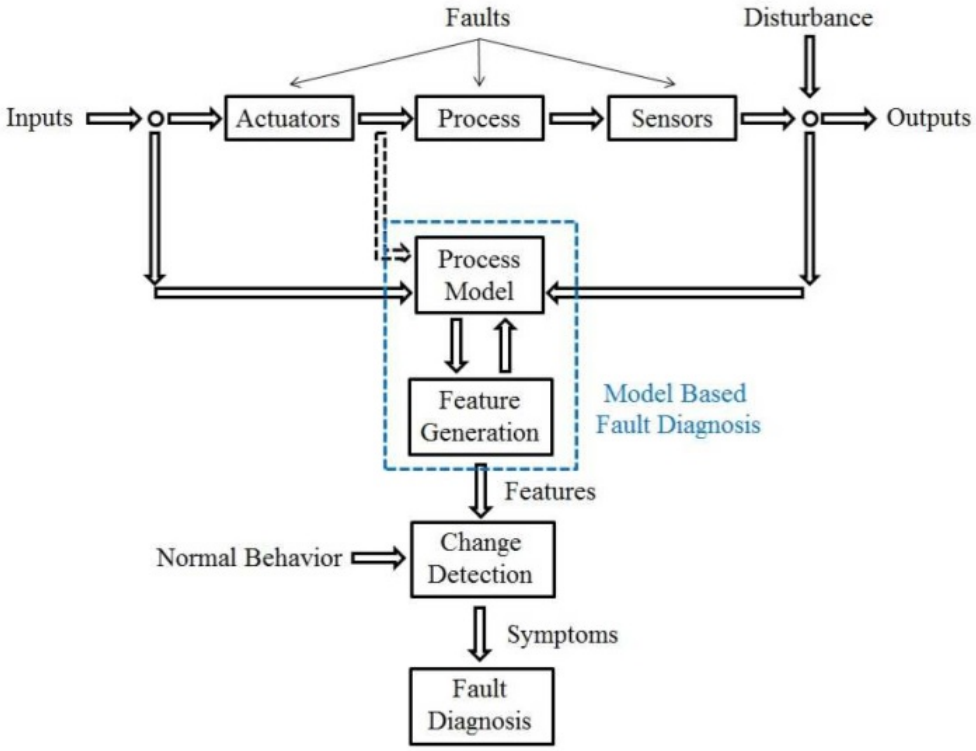

Fig. 2.2 A general schematic of model based fault diagnosis [1].

As illustrated in Figure 2.1, available fault diagnosis techniques are divided to three main categories including process model based and process history based.

2.2.1 Quantitative Model Based Fault Diagnosis

Most of the work on quantitative model-based approaches have been based on general input-output and state-space models as discussed below.

In this approach, the first step is residual generation which consists in producing a signal that carries information about the faults with the aid of an analytical mathematical model of the system. It is important to note that for residual generation no assumptions about the time functions of the faults have to be made.

The residual generator is a dynamic system driven by the input and the output of the system of interest. Because an exact mathematical modelling of the system is impossible in practice, the effects of modelling uncertainties have to be taken into account with respect to which the residuals must be robust.

The numerous methods of analytical model-based residual generation that have been

developed during the last two and a half decades can be divided into three groups:

- Parameter estimation approach
- Observer-based approach
- Parity equations approach

There are close relationships among the different approaches, logical because all of them evaluate the same signals, namely the inputs and the outputs of the actual system of interest. However, depending on the situation, the one or other method might be more or less efficient and hence the approaches are often used in combination.

A general schematic of model based fault diagnosis is presented in Figure 2.2. It is clear that feature generation based on an appropriate process model is the key step in a model based fault diagnosis. The ultimate goal is to determine the presence of faults according to process mathematical model as well as inputs and outputs data (Equation 2.1) [10].

$$Y = f\{U, N, \theta, X\} \tag{2.1}$$

where in equation 2.1, U and $Y$are process inputs and outputs respectively, N represents non-measurable disturbances, θ is process parameters, and X indicates partially measurable process states.

Fault Detection with Parameter Estimation: The model in Equation 2.1 can be obtained from a first-principle analysis. The parameters of this static or dynamic model usually correspond to a real physical property, e.g., temperature, density, viscosity, etc. Therefore, abnormal deviations of such properties will appear in the corresponding parameters. In cases where direct measurement of physical properties is not possible, on-line estimation of parameters based on appropriate regression methods such as conventional/ recursive least squares will indirectly provide the information of the desired property. Based on the deviations of process parameters $(\Delta\theta)$, features and symptoms will be generated, and process faults will be diagnosed [10].

Fault Detection with Observers: There are cases where operating conditions of a

process depend on a an internal (non-measurable) state. In such cases designing an observer for such process states can assist in process monitoring. The process model in the structure of observer can be either static or dynamic. In this approach, first, residuals are generated. Then, ”some special testing methods” will be applied on the residuals, and ultimately, faults will be diagnosed [10]. For the case of multiple mode excited outputs, several methods including bank of observers excited by a single/ all modes are introduced in literature [11, 12].

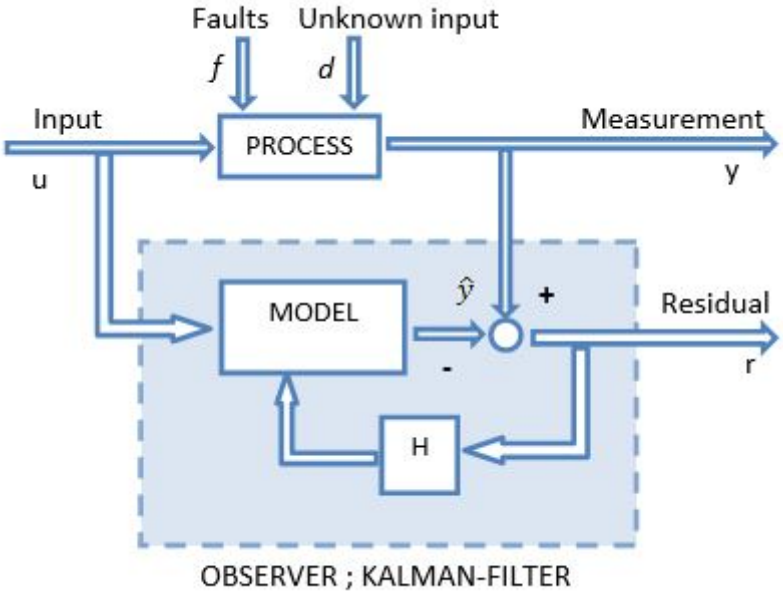

Fig. 2.3 Basic configuration of an observer-based residual generator: output observer of full order

Fault Detection with Parity Equations: The main idea of fault detection based on parity equations is to compare the actual response of a process with the predicted response of the process model, and generate the residuals. This will be followed by a linear transformation to reach the ultimate goal of fault detection and isolation. Parity equations can be developed based on both input and output errors [13].

2.2.2 Qualitative Model Based Fault Diagnosis

The model is usually developed based on some fundamental understanding of the physics of the process. This understanding is expressed in qualitative model functions centered

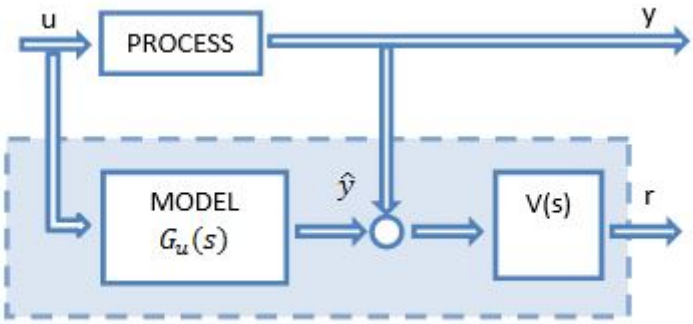

Fig. 2.4 Parity space approach in input-output format

around different units in a process. The need for a reasoning tool which can qualitatively model a system, capture the causal structure of the system in a more profound manner than the conventional expert systems and yet be not as rigid in nature as numeric simulation led to development of many methodologies to qualitatively represent knowledge, and to reason from them. The main developed models are Digraphs based causal models, Fault trees which are used in analyzing the system reliability and safety, qualitative physics and Abstraction hierarchy of process knowledge (based on decomposition in order to be able to draw inferences about the behavior of the overall system solely from the laws governing the behavior of its subsystems.

2.2.3 Process history Based Fault Diagnosis

The underlying idea of history based methods is to extract the critical information (features) from large data sets. The data extraction might be either qualitative or quantitative. Unlike quantitative methods which try to extract the important quantitative information, qualitative methods are usually based on the coding of knowledge and compact representation of trends [8].

In contrast to model based methods, in data based approaches, no prior information about the process is required. Such approaches are based on the extraction of critical information from process historical data. According to Figure 2.1, data based approaches are divided to two main categories including qualitative and quantitative methods.

Qualitative methods are based on either Expert Systems or Qualitative Trend Anal-

ysis (QTA). An Expert System usually consists of a knowledge base, which is appropriately coded, accompanied by an inference procedure based on input-output interfaces. The main advantage of an Expert System is the simplicity of development and analysis [8].

On the other hand, QTA is used for compact representation of significant events in a trend. Thus, these methods can reveal underlying abnormalities and facilitate the task of fault diagnosis. Consequently, one can view these methods as efficient data compression techniques [8].

In the rest of this section, we will focus on history based quantitative methods (Figure 2.1).

Neural Networks

Neural Network (NN) have been frequently used in chemical engineering applications [14, 15]. Many researchers have studied various aspects of NN based fault diagnosis in terms of network architecture, e.g., sinusoidal or radial basis, and the learning method (supervised/ unsupervised). In the case of supervised learning, a certain structure is considered for the network and only the unknown parameters including weights should be estimated. On the other hand, unsupervised neural networks have a time varying structure and can adapt according to recent network inputs. Back propagation neural networks are good examples of supervised learning. To focus on major events rather than the details, traditionally, neural networks are used in accompany with other feature extraction techniques [8, 16].

Principal Component Analysis (PCA)

Extracting dominant relations from a set of large highly correlated variables, and reducing the data set dimension is an important step for monitoring of industrial processes. PCA is frequently used in such cases due to the simple structure and decent performance [17]. The idea of PCA is to map a set of correlated observations to some latent variables. These latent variables, which are linear combinations of real variables, are independent of each other, and contain the significant information (variance) of ob-

servations [18].

In this context, two famous indicators have been introduced to diagnose an abnormal behavior. The first indicator, known as Hotellings T^2, checks variations of the latent variables. Consequently, it will detect an abnormal event only when variations of latent variables are greater than usual. The second indicator, which is the Square Prediction Error (SPE), also known as Q statistic, represents sum of squares of the residuals, and measures appropriateness of the fitness. In other words, T^2 represents the major variation in the data and Q represents the random noise [19]. An alarm will be generated when these indexes pass their standard limits.

Partial Least Square

PLS is similar to PCA in the sense that both methods try to deal with co-linearities in observation data set. However, in PLS, both input and output data are involved. PLS tries to find an outer relation between latent variables of input and output data [20]. Eventually, in this structure, variance of the latent variables in principal components and covariance between input-output latent variables will be maximized simultaneously. Similar to PCA, T^2 and Q can be defined for PLS, and used for process monitoring purposes [18].

Statistical Classifiers

These classifiers try to approximate various modes of a process using appropriate density functions. Accordingly, a new observation will be assigned to a certain class according to its distance from the means of various classes. It is obvious that estimation of the appropriate density function is the key step in such classifiers [8].

In cases where observations do not follow a well-defined distribution, non-parametric methods such as kernel density estimation should be used [21]. Otherwise, parametric distributions such as a mixture of multivariate Gaussian distributions can be applied [22, 23].

Bayesian Fault Diagnosis

Unlike statistical methods which try to ”determine” the true operating mode, in Bayesian approaches, the focus is on finding the ”probability” of current operating mode given all the available information. Consequently, In these approaches, a Bayesian hypothesis testing, which considers the prior (background) information of the modes, is used instead of the likelihood ratio test and other distance based techniques. Such prior information is assumed to be known from historical data or other sources [5].

2.3 Conclusion

In this chapter, the fundamental theory of fault detection and diagnosis has been provided. The various methods used to detect faults have been briefly described as they appear in the literature. And in order to obtain a background in the field of FDD, many definitions and aspects were explored.

In the next chapter, a detailed study of Hidden Markov Models will be provided with a simplified example. Also, dimension reduction tools will be introduced to be used in our proposed method that will be explained later.

CHAPTER 3

Theoretical part

3.1 Introduction

Modern control systems are becoming increasingly more complex, and issues of availability, cost efficiency, reliability, operating safety, and environmental protection concerns are gaining and receiving more attention. This requires a fault diagnosis system that is capable of detecting the occurrence of plant components, actuators and sensors faults (From the measured or unmeasured variables estimation [24]), identifying and isolating the faulty components [25].

Hidden Markov Models were originally reputed for their application in speech processing which has occurred mainly within the past few years [26]. A modeling is proposed by Xuefeng Jiang (2011) for facial expression recognition, which is based on combination of Hidden Markov Model (HMM) and K Nearest Neighbor (KNN) classifiers [27]. Recently, Hyokeun Lee et al. proposed a new initialization of Left-Right Banded (LRB) HMM to mitigate the confusion rate in hand gesture recognition. The proposed HMM achieves better robustness to the confusion miss-classification while maintaining the recognition accuracy [28].

For the application of HMM in fault detection, Sammaknejad (2015) provided robust approach for process operation mode diagnosis using HMMs with adaptive Transition Probability Matrix (TPM) in presence of missing observations [29]. Then, he presented a framework to deal with negative effects of outliers in time varying HMMs [30]. The proposed techniques have been applied to an Oil Sand Primary Separation Vessel [31]. They have some problems associated with HMMs [32] which are solved

using some algorithms as the forward-backward algorithm, the Viterbi algorithm [28], and the Baum-Welch algorithm [33], using standard techniques of statistical inference [34]. Stephen Adams et al. (2016) proposed a joint feature selection and parameter estimation algorithm that is presented for hidden Markov models (HMMs) and hidden semi-Markov models (HSMMs). New parameters and feature saliences are introduced to the model and used to select features that distinguish between states [35].

3.2 Hidden Markov Model

A Hidden Markov Model [36] is a tool for representing probability distributions over a sequence of observations [37]. Let us denote the observation at time t by the variable Y_t. This can be a symbol from a discrete alphabet, a real valued variable, an integer, or any other objects, as long as we can define a probability distribution over it. We assume that the observations are sampled at equally-spaced time intervals, so t can be an integer-valued time index.

3.2.1 Definitions

Definition 1: Suppose that $F\{C_t\}_{t\in N}$ is a Markov chain with state space $C = \{1, \cdots, r\}$, initial distribution $v_c(c \in S)$ and transition matrix $P = (p_{c\grave{c}})\ c, \grave{c} \in C$, and $S_t\ _{t\in N}$ is a stochastic process taking values in $S = \{1, \cdots, m\}$. The bivariate stochastic process $C_t,\ S_t\ _{t\in N}$ is said to be a Hidden Markov Model (HMM) if it is a Markov chain with transition probabilities [38]:

$$\begin{aligned} &P(C_t = c_t, S_t = s_t | C_{t-1} = c_{t-1}, S_{t-1} = s_{t-1}) \\ &= P(C_t = c_t, S_t = s_t | C_{t-1} = c_{t-1}) \\ &= q_{c_{t-1},c_t} \cdot g_{c_t,s_t}. \end{aligned} \tag{3.1}$$

Where $G = (g_{cs})\ c \in C, s \in S$ is a stochastic matrix [34].

Note that the conditional probability of the joint state (C_t, S_t) depends only on C_{t-1} and not S_{t-1}. A useful device for depicting the dependence structure of such model is

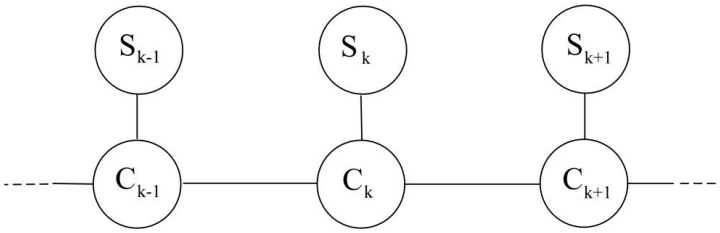

Fig. 3.1 Conditional independence graph of hidden Markov model.

the conditional independence graph. In such a graph the absence of an edge between two vertices indicates that the two variables concerned are conditionally independent given the other variables. Figure 3.1, displays the independence of the observations S_t given the states C_t occupied by the Markov chain, as well as the conditional independence of C_{t-1} and C_{t+1} given C_t, i.e. the Markov property.

Definition 2: The hidden Markov model gets its name from the two defining properties. First it is assumed that the observation at time t was generated by some process whose state S_t is hidden from the observer. Second, it assumes that the state of this hidden process satisfies the Markov property. That is, given the value of S_{t-1}, the current state S_t is independent of all the states prior to $t-1$. The outputs also satisfy a Markov property with respect to the states: given $S_t,\ Y_t$ is independent of the states and observations at all other time indices [32].

Hidden Markov Model can be characterized by five parameters:

1. **N** : the number of states in the model. The states are denoted as $S = \{s_1, s_2, \cdots, s_N\}$.

2. **M** : the number of distinct observation symbols per state. The observation symbols correspond to the physical output of the system being modelled. The symbols are denoted as

$$\mathbf{O} = \{o_1, o_2, \cdots, o_M\}$$

3. $\mathbf{A} = \{a_{ij}\}$: the state transition probability distribution, where

$$a_{ij} = P(q_{t+1} = S_j | q_t = S_i)\ , 1 \leq i, j \leq N \tag{3.2}$$

a_{ij} is the probability of going to state S_j at time $t+1$ given that at time t, the state is S_i.

4. $\mathbf{B} = b_j(k)$: the observation symbol probability distribution in state S_j, where

$$b_j(k) = P(O_k(t)|q_t = S_j),\ 1 \leq j \leq N, 1 \leq k \leq M \tag{3.3}$$

$b_j(k)$ is the probability of the k^{th} observation symbol given that the state at time t is S_j.

5. $\pi = \{\pi_i\}$: the initial state distribution, which is the probability of being in the i^{th} state at the initial time $t = 1$. Where

$$\pi_i = P(q_1 = S_i),\ 1 < i < N \tag{3.4}$$

The three main components of a HMM are the state transition probability matrix **A**, the measurement probability distribution matrix **B**, and the initial state probability distribution π. For convenience, a compact notation is used to indicate the complete parameter set of the model:

$$\lambda = (A, B, \pi) \tag{3.5}$$

According to characteristics of the observation symbols, there are two kinds of HMMs: Discrete Hidden Markov Model (DHMM) and Continuous Hidden Markov Model (CHMM) [26][39]. The observation symbols of DHMM are mentioned above, while, the observation symbols of CHMM are continuous, Gaussian distribution of which is assumed in each hidden state. In this study, only DHMM will be considered.

3.2.2 An illustrated example

Definition 3 (Discrete Markov Process): Consider a system which may be described at any time as being in one state from the set of m distinct states $S_1, S_2, \cdots, S_m$. At regularly spaced discrete times, the system undergoes a change of state (self transition is also possible) according to a set of probabilities associated with the state. The time

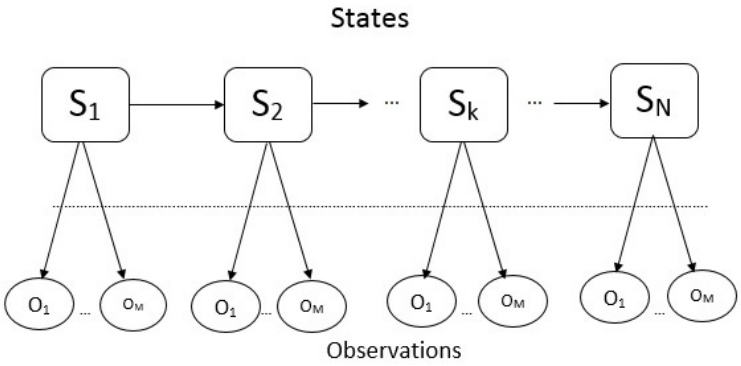

Fig. 3.2 Hidden Markov model with N-states and M-observation.

instants associated with the state changes are denoted as $t = 1, 2, \cdots$ and the actual state at time t is S_t.

In order to explain the concept of Markov models, the weather prediction example is used here [40]. Consider to have three types of weather, e.g., sunny, rainy and cloudy. Assuming that each weather lasts for a whole day, the goal is to predict tomorrows weather based on available historical data, i.e.,

$$P(q_t = S_j | q_{t-1} = S_i, q_{t-2} = S_k, ..., q_1 = S_1) \tag{3.6}$$

where q_t indicates the state (weather) at each sampling time (day), and $S_l,\ l \in \{1,2,3\}$ corresponds to sunny, cloudy and rainy weather respectively.

According to Equation 3.6, the more past history involved, the more complex will be the computation. Assuming to use only the past five states, $3^5 = 243$ statistics are required to do a future prediction. A first-order Markov assumption simplifies this computation as follows:

$$P(q_t = S_j | q_{t-1} = S_i, q_{t-2} = S_k, ..., q_1 = S_1) \approx P(q_t = S_j | q_{t-1} = S_i) \tag{3.7}$$

Furthermore, assuming stationarity, the right hand side of (3.7) is independent of time

and thus leads to the state transition probabilities characterized as follows:

$$A = [a_{i,j}]_{mm} = [P(q_t = S_j | q_{t-1} = S_i)]_{mm} \tag{3.8}$$

with the state transition elements having the properties $a_{i,j} \geq 0$ and $\sum_{j=1}^{m} a_{i,j} = 1$ since they obey standard stochastic constraints.

In this configuration, the closest previous state is considered to contain the historical information. The second or higher order Markov assumptions can be applied in a similar manner.

According to the first-order Markov assumption, the joint probability of a sequence of states can be computed as in Equation 3.9.

$$p(q_1, q_2, q_3, \cdots, q_t) = \prod_{K=1}^{t} P(q_k | q_{k-1}) \tag{3.9}$$

The proposed weather prediction Markov chain structure is illustrated in Figure 3.2. $a_{i,j}$ represents the probability of transiting from state i to state j.

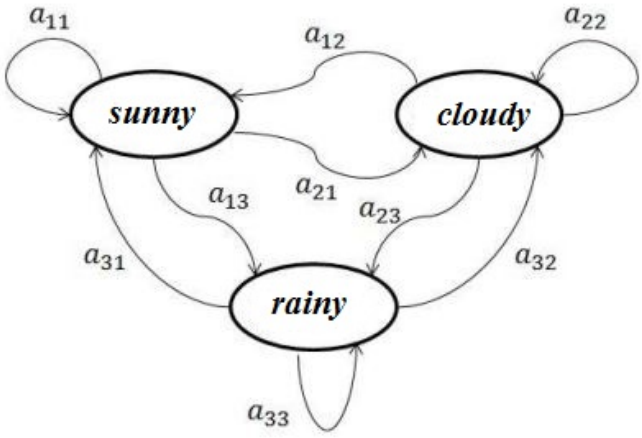

Fig. 3.3 A three states Markov chain process (State 1=Sunny, State 2=cloudy and State 3=Rainy).

This example can be further extended to explain hidden Markov models [41]. Suppose that you are locked in a room and you are asked about outside weather. Your only information about the outside weather is a person who brings the daily meal, whether he carries an umbrella or not. The main difference of the current example from the pre-

vious one is that the actual weather, or "state", is hidden from you now. Therefore, you need to indirectly infer about it. Assuming to have a set of observations $O_1, O_2, \cdots, O_t$, and using the Bayes rule, Equation 3.9 can be written as

$$P(q_1, q_2, ..., q_t|O_1, O_2, ..., O_t) = \frac{P(O_1, O_2, \cdots, O_t|q_1, q_2, \cdots, q_t)P(q_1, q_2, \cdots, q_t)}{P(O_1, O_2, \cdots, O_t)} \tag{3.10}$$

where in this scenario $O_{l,\ l\in\{1,\cdots,t\}}$ is either True (carrying an umbrella), or False (not carrying an umbrella). $P(q_1, q_2, \cdots, q_t)$ can be calculated using the same previous Markov model (Equation 3.9). $P(O_1, O_2, \cdots, O_t)$ is the prior probability of observing a particular sequence of umbrella events, e.g., {True, False, True, ...}. Assuming that O_i given q_i is independent of all O_j and q_j for $j \neq i = 1, \cdots, t$, then

$$P(O_1, O_2, \cdots, O_t|q_1, q_2, \cdots, q_t) = \prod_{K=1}^{t} P(O_k|q_{k-1}) \tag{3.11}$$

3.2.2.1 Basic Settings for HMMs

According to the previous example, fundamental underlying assumptions of HMMs can be summarized as follows [42]:

1. **Discrete state space assumption**: States can only accept discrete values, i.e., $q_t \in \{S_1, \cdots, S_N\}$.

2. **Markov assumptions**:

   - Given the state in the previous sample time $t-1$, current state at time t will be independent of previous states 1 to $t-2$, i.e., $q_t \perp q_i|q_{t-1},\ \forall i \leq t-2$.
   - Given the state at time t, the corresponding observation O_t is independent of all other states, i.e., $O_t \perp q_i|q_t,\ \forall i \neq t$.

Following the previous example and Markov assumptions, various elements of HMMs can be summarized as follows [2]:

1. **Number of states in the model (N)**:

   As previously mentioned, states at each sampling instant can only take a limited

number of discrete events, i.e., $q_t \in S = \{S_1, \cdots, S_N\}$. In many applications, these states correspond to a physical phenomena, e.g., in the weather prediction example, states correspond to sunny, foggy and rainy events.

2. **Number of observation symbols per state (M)**:

   If observations are discrete, they can only take a limited number of symbols,i.e., $O_t \in V = \{v_1, \cdots, v_M\}$. For example, in the weather prediction case study, observations in each state can only take two possible symbols, i.e., False and True. Observations can also take continuous values in general.

3. **State transition probability matrix** $(A = \{a_{ij}\})$: The state transition probabilities, as illustrated in Figure 2.1, can be defined as follows:

$$a_{ij} = P[q_t = S_j | q_{t-1} = S_i],\ 1 \leq i, j \leq N \tag{3.12}$$

   According to this definition, the following property should hold for the state transition probabilities:

$$\sum_{j=1}^{N} a_{i,j} = 1,\ a_{ij} \geq 0$$

4. **Emission probability matrix** $(B = \{b_j(k)\})$: Probability of observation symbols given various states can be defined using the transition probability matrix as follows:

$$b_j(k) = p[O_t = v_k | q_t = S_j],\ 1 \leq j \leq N,\ 1 \leq k \leq M \tag{3.13}$$

   According to this definition, the following constraint should be satisfied:

$$\sum_{j=1}^{M} b_j(k) = 1,\ b_j(k) \geq 0 \tag{3.14}$$

5. **Initial state distribution** $(\pi = \pi_i)$:

   The initial state distribution is specifically defined for the first sample time, i.e.,

$$\pi_i = p[q_1 = S_i],\ 1 \leq i \leq N \tag{3.15}$$

The following constraint should hold for the initial state distribution:

$$\sum_{j=1}^{N} \pi_i = 1,\ \pi \geq 0 \tag{3.16}$$

Having the above five parameters, a hidden Markov model is defined. For convenience, one might use the compact notation in Equation 3.16 to represent HMMs.

$$\lambda = (A, B, \pi) \tag{3.17}$$

3.2.2.2 Three Fundamental Problems for HMMs

In literature, three fundamental problems are introduced and addressed for HMMs as follows [2]:

- Problem 1 :

  Having the model $\lambda = (A, B, \pi)$, how to compute the probability of an observation sequence $O = O_1, O_2, \cdots, O_T$ given the model , $i.e., p(O|\lambda) =?$

  - Complicated-Hidden states.

  - Useful in sequence classification [43].

- Problem 2 :

  Having the model $\lambda = (A, B, \pi)$, how to find the optimal state sequence $Q = q_1, q_2, \cdots q_T$ which best explains the observation sequence $O = O_1, O_2, \cdots O_T$?

  - Optimality criterion.

  - Useful in recognition problems [43].

- Problem 3 :

  How to find model parameters in $\lambda = (A, B, \pi)$ such that $P(O|\lambda)$ is maximized ?

  In the rest of this section, we will address these three problems.

Solution to Problem 1

- Naive solution

A straight forward approach to address problem 1 is to marginalize $P(O|\lambda)$ over all possible state sequences [43], i.e.,

$$P\left(O|\lambda\right) = \sum_{allq} P\left(O, q|\lambda\right) = \sum_{q} P\left(O|q, \lambda\right) P\left(q|\lambda\right) \tag{3.18}$$

Using Markov assumptions, each term in the right hand side of Equation 3.18 can be written as

$$P\left(O|q, \lambda\right) = \prod_{t=1}^{T} P\left(o_t|q_t, \lambda\right) = b_{q_1}\left(o_1\right) b_{q_2}\left(o_2\right) \ldots b_{q_T}\left(o_T\right) \tag{3.19}$$

$$P\left(q|\lambda\right) = \prod_{t=1}^{T} P\left(q_t|q_{t-1}, \lambda\right) = \pi_{q_1} a_{q_1 q_2} a_{q_2 q_3} \ldots a_{q_{T-1} q_T} \tag{3.20}$$

Having the expression in Equation 3.18 and the above two equations, $p(O|\lambda)$ can be computed as

$$P\left(O|\lambda\right) = \sum_{allq} P\left(O, q|\lambda\right) = \sum_{q} P\left(O|q, \lambda\right) P\left(q|\lambda\right) \tag{3.21}$$

$$= \sum_{q_1, q_2, \ldots, q_T} \pi_{q_1} b_{q_1}\left(o_1\right) a_{q_1 q_2} b_{q_2}\left(o_2\right) \cdots a_{q_{T-1} q_T} b_{q_T}\left(o_T\right)$$

It can be proved that to compute the expression in Equation 3.21, $2T.N^T$ calculations are required, e.g., having only 5 states, for a sequence of 100 observations, 10^{72} calculations are required. Therefore, forward and backward procedures have been introduced to more efficiently compute $p(O|\lambda)$ [2].

- Efficient solution

  Forward algorithm:

  In the forward algorithm, the auxiliary forward variable defined as: $\alpha_t\left(i\right) = P\left(o_1, \ldots, o_t|q_t = S_i, \lambda\right)$ [44], represents the probability of observing the partial sequence $O_1, O_2, \cdots, O_t$ such that state q_t is S_i. This probability is calculated inductively as follows:

1. Initialization

$$\alpha_1(i) = \pi_i b_i(o_1), 1 \leq i \leq N \tag{3.22}$$

2. Induction

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i) a_{ij}\right] b_j(o_{t+1}), 1 \leq t \leq T-1, 1 \leq j \leq N \tag{3.23}$$

3. Termination

$$P(O|\lambda) = \sum_{i=1}^{N} \alpha_T(i) \tag{3.24}$$

The left side of Figure 3.3 illustrates one sample induction step where State i at time t can be reached from all the previous states at time $t-1$. The final termination step is to marginalize over all possible states at time T, i.e., q_T, which results in the probability $p(O|\lambda)$. As the result of such procedure, the computation complexity will be reduced to N^2T (total T observations and N^2 computations between each two consecutive observations) which is significantly lower than equation 3.21 [40].

Backward algorithm:

One can perform calculations, very similar to Equations 3.22 to 3.24, to find $p(O|\lambda)$ backward. To do this, the auxiliary backward probability $\beta_t(i)$ is defined as $\beta_t(i) = P(o_{t+1}, o_{t+2}, .., o_T|q_t = S_i, \lambda)$. To find $p(O|\lambda)$ the following steps are required:

1. Initialization

$$\beta_T(j) = 1, 1 \leq j \leq N \tag{3.25}$$

2. Induction

$$\beta_t(i) = \sum_{j=1}^{N} \beta_{t+1}(j) a_{ij} b_j(o_{t+1}), t = T-1, T-2, \cdots, 1, 1 \leq i \leq N \tag{3.26}$$

3. Termination

$$p(O|\lambda) = \sum_{i=1}^{N} \pi_i b_i(o_1) \beta_1(i) \tag{3.27}$$

The induction step of the backward algorithm is presented in the right hand side of Figure 3.4. Unlike the forward algorithm, in backward procedure, states of the next sample time $(t+2)$ are used to reach the states in the current sample time $(t+1)$. To address Problems 2 and 3, a combination of forward and backward algorithms (forward-backward algorithm) is required. Figure 3.4 presents how these two probabilities are connected [40].

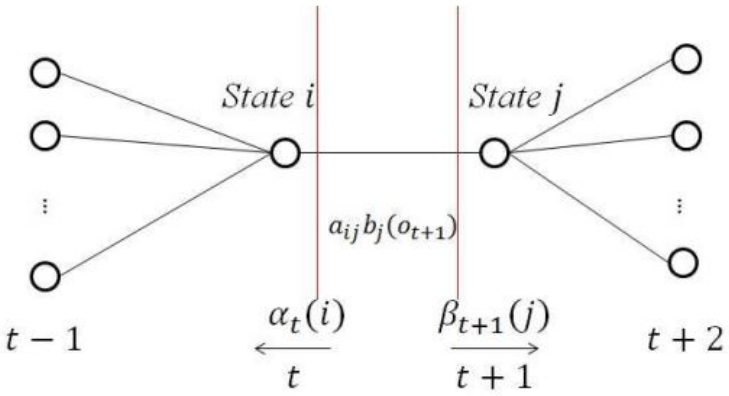

Fig. 3.4 Graphical illustration of Forward-Backward algorithm [2]

Solution to Problem 2 (Decoding)

The next important problem to be addressed is to find the "optimal" path of states given an observation sequence [2]. One approach to solve this problem is to choose the states which are individually more likely. The more general case is to find the optimal state sequence $Q = (q_1, q_2, \cdots, q_T)$ for a given observation sequence $O = (O_1, O_2, \cdots, O_T)$. The problem can be defined as the best state path to maximize $p(Q|O, \lambda)$ or equivalently $p(Q, O|\lambda)$ [40].

- **Viterbi algorithm:** is a dynamic programming technique, addresses this problem [2, 40]. This inductive technique keeps the best state sequence at each instance. Let us define the auxiliary variable $\delta_t(i)$ that represents the highest probability of a path at time t, which ends in state S_i, as follows:

$$\delta_t(i) = max_q P(q_1, q_2, .., q_t = i, O_1, O_2, \cdots, O_t|\lambda) \tag{3.28}$$

By induction, it is straightforward to show that

$$\delta_{t+1}(j) = max_i \left(\delta_t(i)\, a_{ij}\right) \cdot b_j(o_{t+1}) \tag{3.29}$$

The idea of the Viterbi algorithm is to keep the track of the maximized argument in Equation 3.29 for each t and j through the array $\varphi_t(j)$. The complete procedure is as follows:

1. Initialization:

$$\delta_1(i) = \pi_i \; b_i(o_1) \qquad 1 \leq i \leq N \tag{3.30}$$

$$\varphi_1(i) = 0 \tag{3.31}$$

2. Recursion:

$$\delta_t(i) = max_{1 \leq i \leq N} \left(\delta_{t-1}(i)\, a_{ij}\right) b_i(o_t) \tag{3.32}$$

$$\varphi_t(j) = argmax_{1 \leq i \leq N} \left(\delta_{t-1}(i)\, a_{ij}\right) \qquad 2 \leq t \leq T, 1 \leq j \leq N \tag{3.33}$$

3. Termination:

$$P^* = max_{1 \leq i \leq N}[\delta_T(i)] \tag{3.34}$$

$$q_T^* = argmax_{1 \leq i \leq N}[\delta_T(i)] \tag{3.35}$$

4. Backtracking:

$$q_t^* = \varphi_{t+1}\left(q_{t+1}^*\right) \qquad t = T-1, T-2, \cdots, 1 \tag{3.36}$$

It should be noted that steps of the Viterbi algorithm are very similar to the forward algorithm. However, except the marginalization step in Equation 3.23, a maximization (recursion step) is performed.

Solution to Problem 3 (Learning)

- Training HMM to encode observation sequence such that HMM should identify a similar observation sequence in future.

- Find $\lambda = (A, B, \pi)$, maximising $P(O|\lambda)$.
- General algorithm [43]:
  - Initialise: λ_o
  - Compute new model λ using λ_0 and observed sequence O.
  - Then $\lambda_0 \leftarrow \lambda$
  - Repeat steps 2 and 3 until:

$$logP\left(O|\lambda\right) - logP\left(O|\lambda_0\right) < d$$

Step 1 of Baum-Welch algorithm [40, 43]:

- Let $\zeta(i, j)$ be a probability of being in state i at time t and at state j at time $t + 1$, given λ and O sequence

$$\zeta\left(i,j\right) = \frac{\alpha_t\left(i\right) a_{ij} b_j\left(o_{t+1}\right) \beta_{t+1}\left(j\right)}{P\left(O|\lambda\right)} \tag{3.37}$$

$$= \frac{\alpha_t\left(i\right) a_{ij} b_j\left(o_{t+1}\right) \beta_{t+1}\left(j\right)}{\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_t\left(i\right) a_{ij} b_j\left(o_{t+1}\right) \beta_{t+1}\left(j\right)}$$

- Let be a probability of being in state i at time t, given observation O

$$\gamma_t\left(i\right) = \sum_{j+1}^{N} \varrho_t\left(i,j\right) \tag{3.38}$$

- $\sum_{t=1}^{T-1} \gamma_t\left(i\right)$ expected no. of transitions from state i
- $\sum_{t=1}^{T-1} \zeta_t\left(i\right)$ expected no. of transitions $i \rightarrow j$

Step 2 of Baum-Welch algorithm:

- $\hat{\pi} = \gamma_1\left(i\right)$ the expected frequency of state i at time $t = 1$
- $a_{ij} = \frac{\sum \zeta_t(i,j)}{\sum \gamma_t(j)}$ ratio of expected no. of transitions from state i to j over expected no. of transitions from state i

- $b_j = \frac{\sum_{t,o_t=k} \gamma_t(j)}{\sum \gamma_t(j)}$ ratio of expected no. of times in state j observing symbol k over expected no. of times in state j
- Baum-Welch algorithm uses the forward and backward algorithms to calculate the auxiliary variables α, β
- B-W algorithm is a special case of the EM algorithm:
  - E-step: calculation of ζ and γ
  - M-step: iterative calculation of $\hat{\pi}, \hat{a_{ij}}, b_j(k)$
- There are some practical issues concerning this algorithm [43]:
  - Can get stuck in local maxima
  - Numerical problems log and scaling

3.2.3 HMM Based Fault Diagnosis

As previously stated, HMMs are sophisticated mathematical tools to extract temporal information from historical data sets. In this section, we will review some of the previous studies where HMMs have been used to improve classic fault diagnosis methods.

3.2.3.1 Model Based Approaches in Conjunction with HMMs

Detection of faults based on a combination of dynamic process models such as ARX models and HMMs is a well-known technique for improving model-based detection of faults. This approach uses deviations of the estimated parameters of the process model as observations. Unlike traditional methods in which faults are assumed to be independent in each step, this approach considers faults to be "persistent" over time. Consequently, applying HMMs, previous information up to the current time is used as prior information for recent observations [7, 40].

3.2.3.2 Qualitative Trend Analysis in Conjunction with HMMs

Qualitative trend analysis is used to extract significant events from process data. Therefore, analysis of these major events as a time sequence can provide critical information about process status. To do this, one can first discretize a signal according to its extrema and inflection points. Therefore, the continuous signal will appear as a sequence of discrete observations. This sequence can then be used as the input of HMMs. After training the HMMs according to the historical data, they can be used for the purpose of fault diagnosis in an on-line application. The decision on operating condition of the process will be made based on the probability of a window of recent observations given various HMMs [45]. Problem can be extended to the case of multiple observations. However, another inference layer based on neural networks or other data driven tools will be required [46].

3.2.3.3 Statistical Classifiers in Conjunction with HMMs

In another recent study, a new approach for fault diagnosis of gear transmission systems is introduced. In this study, process behavior is modeled using a three state continuous-time homogeneous Markov process [47]. Different states of the Markov process correspond to various operating modes. Observations are assumed to follow multivariate Gaussian distributions. The ultimate target is to infer the current operating condition in an on-line application given the observations. An ”optimal control limit” is defined to isolate the modes. Parameter estimation is based on the Expectation Maximization (EM) algorithm [47].

Applications of HMMs in fault diagnosis are not limited to the above studies. One can refer to many other available articles for more information [48, 49, 50].

3.3 Principal Component Analysis

Principal component analysis (PCA) is a basic method in the framework of the multivariate analysis techniques. It has been successfully used in numerous areas includ-

ing data compression, feature extraction, image processing, pattern recognition, signal analysis, and process monitoring [51].

PCA is recognised as a powerful tool of statistical process monitoring and widely used in the process industry for fault detection and diagnosis due to its simplicity and efficiency in processing huge amount of process data [52, 53]. Recent development in the PCA technique is focused on achieving adaptive process monitoring using, for instance, recursive implementation of PCA, fast moving window PCA [51, 54, 55] or kernel PCA [56][57].

In this study, Pca technique is used for dimension reduction [58].

3.3.1 PCA: Mathematical Basis

The mean central idea of PCA is to reduce the dimensionality of a data table consisting of large number of interrelated variables, and to extract the important information from a data table which based on defining the direction of the maximum variance containing the maximum amount of information about the process, and the direction having a minimum variance along them contains almost no information to retain. To represent it as a set of new orthogonal variables called principle component (PCs), which are uncorrelated, and which are ordered so that the first few component contain most of the information in the original data set. Since a change of basis applied, PCA is clearly assuming a linear relationship between the original data [59].

The GOALS of PCA is:

1. Extract the most information from the data table.
2. Reduce the size of the data table by keeping only the important information.
3. Simplify the description of the data table.
4. Analyze the structure of the observation and variables [60].

The data table to be analyzed by PCA comprises N of observation (instances) described by m variables and it is represented by matrix $X \in R^{N \times m}$. A standard PCA

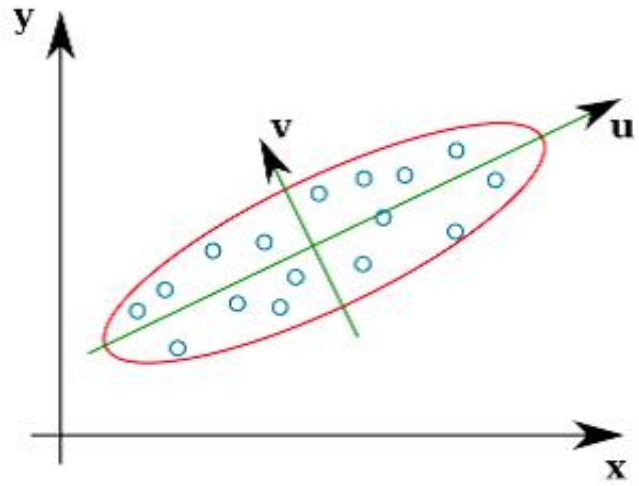

Fig. 3.5 Illustration of PCA in 2D

approach for fault diagnosis consists of three steps and can be briefly formulated as follow:

- Collection and normalization: In this step, N samples for each sensor are first collected and recorded in a data matrix X, $X \in R^{N\times m}$, then shifted to zero mean, and scaled to unit variance. Let the scaled data be

$$X_n = \begin{bmatrix} {x^T}_1 \\ \vdots \\ {x^T}_N \end{bmatrix}, X_n \in R^{N\times m} \tag{3.39}$$

  with $x_i \in R^m,\ i = 1, \ldots, N$ denoting a sample vector of the m sensors.

- Computation of singular values, corresponding singular vectors :

  - First, the covariance matrix is formed,

$$cov(X) \approx \frac{1}{N-1} X^T X \tag{3.40}$$

  - Then, by means of, for example, an SVD (singular value decomposition), the principal components and the associated singular vectors are computed

as

$$\frac{1}{N-1}X^T X = P^T \Lambda P \tag{3.41}$$

Where P consists of orthonormal covariance matrix eigenvectors and Λ is m by m diagonal matrix containing the covariance matrix eigenvalues sorted in descending order. Once the number of principal components l is selected, P and Λ matrices can be decomposed into

$$\Lambda = \begin{bmatrix} \hat{\Lambda} & 0 \\ 0 & \tilde{\Lambda} \end{bmatrix} \tag{3.42}$$

$$P = \begin{bmatrix} \hat{P} & \tilde{P} \end{bmatrix} \tag{3.43}$$

So,

$$cov(X_n) = P^T(diag[\lambda_1, \lambda_2, \cdots, \lambda_l | \lambda_{l+1}, \lambda_{l+2}, \cdots, \lambda_m])P \tag{3.44}$$

Such that $\hat{\Lambda} \in R^{l \times l}, \tilde{\Lambda} \in R^{(m-l)\times(m-l)}, \hat{P} \in R^{m \times l}$, $\tilde{P} \in R^{m(m-l)}$, and l is the retained Principal Components (PCs).

- Computing the scores matrix T : The scores are computed by the following linear transformation

$$T = X_n P \tag{3.45}$$

T can be decomposed as

$$T = (\hat{T}\ \tilde{T}) \tag{3.46}$$

this leads to

$$\hat{T} = X_n \hat{P} \tag{3.47}$$

In our case, $\hat{T}$ is used to extract our observables to built our Markov model.

3.3.2 PCA Model Dimension identification:

The selection of the number of retained component plays a major role in the goodness of PCA model. Many researches have been carried out in order to determine the optimal number of principal components. However, the number of retained components that is guaranteed to peak the best number of PCs still out of reach, even though a considerable number of methods are already presented in literature. Some of the widely used methods are presented in this section.

1. Kaiser Criteria:

   The most used method to estimate the number of component in PCA and factor analysis consists of taking the components corresponding to those eigenvalues larger than one. Thus any PC with eigenvalue less than one contain less information than one of the original variables and so is not worth retaining. The major problem with Kaiser always retain 1/3, 1/5 or 1/6 of the total components. However deletion will occur if Kaiser rule is used and if those eigenvalues less than one. It is therefore advisable to choose cut off lower than one. Jolliffe proposed a value of 0.7 to deal with this significant problem.

2. Scree plot:

   Scree plot is a graphical way to determine the number of retained principal components. It was discussed and named by Cattell (1966) [53]. It consist of plotting the eigenvalues of the covariance/correlation matrix in decreasing order which involve looking a plot of λ_K (eigenvalues) versus K (the index of the eigenvalues) and deciding at which value of K the slope of line joining the plotted points are steep to the left of K, and not steep to the right. This value of K that defines an elbow in the graph is then considered to be the number of components to be retained [53],[61]. Rather than λ_K versus K we can plot $\log(\lambda_K)$ against K, this known as the Log Eigen-Values (LEV).

3. Cumulative percent of variance (CPV):

CPV is one of the most obvious and simplest criteria for choosing the number of nontrivial [61] components that corresponds to a certain percentage of the total variance based on the eigenvalues of the covariance/correlation matrix to find the amount of variance preserved by l-dimension out of total of available measurement as:

$$CPV(\%) = \frac{\sum_{i=1}^{l} \lambda_i}{\sum_{i=1}^{m} \lambda_i} \times 100 \tag{3.48}$$

Choosing those l PCs turns out matter of choosing the percentage of total variance we want to retain. Usually, a percentage between 80 - 85% is used in some context while some references encourage to use percentage greater than 85%. While we want to retain few PCs as possible and in same time we would like to account for as much variance as possible. Then the decision becomes a balance between the amount of parsimony and comprehensiveness of the model. We must decide what balance will serve our purpose.

4. Parallel Analysis (PA):

   Parallel analysis is a method was proposed by Horn in 1965 [61] as a modification of catells scree diagram to alleviate the component indeterminacy problem. The procedure is based on plotting and comparing the eigenvalues of the covariance/correlation matrix of the original data set with those of a random uncorrelated data set matrix from normally distributed population. If the associated eigenvalue is larger than those generated from random sample this principle component considered to be significant and retainable. The values above the intersection between two graphs represent the process information retained components and those under the intersection are considered to be noise. It is usually required to repeat the generation of pseudo-eigenvalues for K times for may occur since the method is based on generating randomly sampled data set [59, 61]. Horn was proposed solution for this issue by taking the states that a number of correlation matrices with the same size as the data set under study are constructed, their pseudo-eigenvalues are averaged and compared to the eigenvalue of the real data

correlation matrix. The algorithm of PA as stated in [62] is summarized in the following few lines:

- Calculate the eigenvalues of the covariance matrix: $\lambda = \{\lambda_1, \lambda_2, \cdots, \lambda_m\}$.
- Generate $k - sets$ of random data with the same size as $X(Y_i \in R^{n \times m}, i = 1, 2, \cdots, k)$.
- Compute the eigenvalues Y_i and store them in $\Xi \in R^{k \times m}$ (row-wise). With ξ_{ij} the j^{th} pseudo-eigenvalue for the i^{th} experiment.

$$\Xi = \begin{bmatrix} \xi_{11} & \cdots & \xi_{1m} \\ \vdots & \ddots & \vdots \\ \xi_{k1} & \cdots & \xi_{km} \end{bmatrix}$$

  Do either:

  - Compute the average eigenvalue: $\hat{\Xi} = [\xi_1, \xi_2, \cdots, \xi_m]$.
  - Compute the $\alpha - percentile$ for each column: $\hat{\Xi} = [\xi_{1\alpha}, \xi_{2\alpha}, \cdots, \xi_{m\alpha}]$ With α is usually selected between 0.95 and 0.99 serves to fix the tendency of the method to overestimate the number of components.

- Retain the components for which: $\lambda_i > \hat{\Xi}_i$. [59]

5. Variance of the reconstruction error (VRE):

The variance of the reconstruction error (VRE) was first developed by Dunia and QIn [61]. This techniques used to determine the number of principal components based on the best reconstruction of the process variable. This approach has an important characteristic that the VRE index reach a minimum value corresponding to the best reconstruction. The reconstruction error is a function of the number of PCs when the PCA model used to reconstruct faulty sensors, the minimum found in VRE calculation determine the number of PCs. This is because VRE is decomposed in two subspaces: the Principle Components Subspace (PCS) and the Residual Subspace (RS) [63][64]. The PCS has a tendency to increase with the number of PCs, and that in the RS has a tendency to decrease with the number of

PCs. As result the VRE always has a minimum value that determine the optimal number of PCs.

Principle of reconstruction: The principle of reconstruction estimate the i^{th} variable from the remaining variables in measurement vector $X \in R^{n \times m}$. Qin and Dunia [63] consider that the sensor measurement is corrupted with a fault along a direction ξ_i, thus, the actual sensor measurement is

$$x = x^* + f\xi_i$$

With x^* is the normal sample vector, f is the fault magnitude and $||\xi|| = 1$. Where ξ_i corresponds to the i^{th} column of an identity matrix. The task of sensor reconstruction is to find an estimate for x^* along the direction of upset ξ, i.e to best correct the effect of the upset in other word we try to find f_i such that:

$$x_i = x + f_i\xi_i$$

Is the most consisting with the PCA model or x_i has the minimum model error the solution can be easily found using least square method. There exist several different approaches of reconstruction that yield exactly the same solution [65][64].

The VRE procedure for selecting the number of PCs can be summarized as follows:

- Build a PCA model using normal data from all variables.
- Reconstruct each variable using other variables and calculate the VRE.
- Calculate the Total VRE.
- The number of PCs that gives the minimum VRE is selected, which corresponds to the best reconstruction.

3.4 Proposed Method

In this chapter, a proposed model will be discussed and the explained method will be used to detect the fault in a process. The procedure will be as follow:

3.4.1 Work Procedure

1. **Model construction**: The first step in the procedure is to initiate a suitable model for our diagnosis process, with initial and transition probabilities. In our case the proposed model is shown in figure 3.6, with two states and 'l' observables.

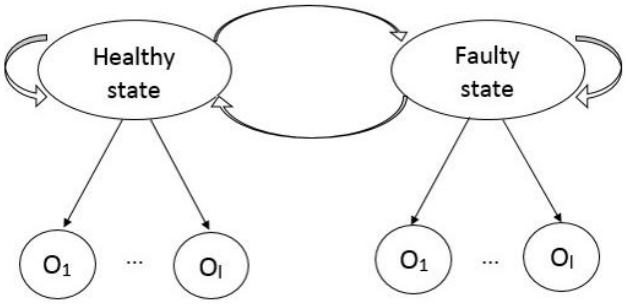

Fig. 3.6 Proposed model

Observables are extracted from the training data, using principal component analysis.

- Data from sensors are collected and recorded in a data matrix called X, $X \in R^{n \times m}$.
- Data matrix is normalized (shifted to zero mean, and scaled to unit variance). Let the normalized data be

$$X_n = \Big[x_{n1} \dots x_{nm} \Big] \ , X_n \in R^{n \times m} \tag{3.49}$$

with $x_{ni} \in R^{n \times 1},\ i = 1, \cdots, m$ denoting a (scaled) sample vector of the m sensors with n samples.

- Computation of singular values and corresponding singular vectors.

- Calculation of 'l' the number of retained PCs using one of the techniques discussed above .
- Computation of the scores matrix $\hat{T}$.

2. **Training phase**: This step is too important since different information will be introduced to the constructed model. The more information involved the more accurate will be the model.
   - Using the $hist$ function with a suitable number of $bins$, we calculate the joint probabilities of our obsevables as follow:

$$P_{T_1,T_2,...,T_l}(t_1, t_2, ..., t_l)$$

$$= P(T_1 = t_1, T_2 = t_2, ..., T_l = t_l) \tag{3.50}$$

   Since observations are independent, equation 3.50 will be

$$P_{T_1,T_2,...,T_l}(t_1, t_2, ..., t_l)$$

$$= P_{T_1}(t_1) \times P_{T_2}(t_2) \times \cdots \times P_{T_l}(t_l) \tag{3.51}$$

   - Now we inject different types of faults in the healthy data and we repeat the same steps in order to calculate the joint probabilities for faulty data.

3. **Testing phase**:
   - Now, new data are collected from sensors and normalized.
   - Observables are extracted, and their corresponding joint probabilities are computed.
   - Calculating

$$P(H_{t+1}/O_{t+1}) = \frac{P(O_{t+1}/H_{t+1}) \times P(H_{t+1})}{P(O_{t+1})} \tag{3.52}$$

$$P(F_{t+1}/O_{t+1}) = \frac{P(O_{t+1}/F_{t+1}) \times P(F_{t+1})}{P(O_{t+1})} \tag{3.53}$$

- Checking the maximum between the two calculated probabilities.
- Constructing a fault indicator that satisfies the the condition bellow :

  $S = 0$ if $P(H_{t+1}/O_{t+1}) > P(F_{t+1}/O_{t+1})$;

  $S = 1$, elsewhere.

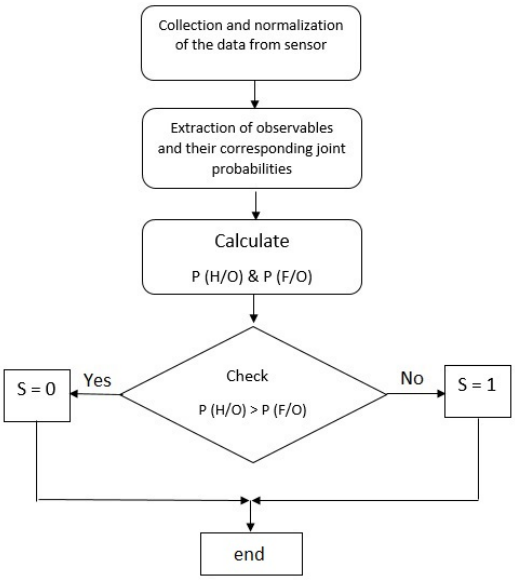

Fig. 3.7 Procedure's Flowchart

3.5 A Numerical Example

In this section, a numerical example is used to test the accuracy and effectiveness of the discussed procedure in detecting different types of faults. The example is expressed as follows:

$$x_1 \sim N(1, 0.01^2);$$

$$x_2 \sim N(4, 0.01^2);$$

$$x_3 = 2x_1 - x_2;$$

$$x_4 = x_1 - 3x_3 + e_1;$$

$$x_5 = x_2 + x_4 + x_1 + e_2;$$

$$x_6 = x_3 + x_4 - x_1 + e_3;$$

in which $e_i \sim N(0, 0.0001^2)$, i = 1,2,3, are white noises. Variable x_1 influences variable x_3, x_4 and x_5. Therefore, failure occurring in variable x_1 will affect the quality.
The training data-set is stored in the data matrix

$$X = [X1\ X2\ X3\ X4\ X5\ X6] \in R^{6000\times 6}.$$

The used data sets are described in Table 3.1, and the simulated faults in Table 3.2.

Table 3.1 Data sets used in the simulated example.

Data sets	Size	description
Training set	$X_0 \in R^{6000\times 6}$	Normal operation data, used to construct FD scheme
Testing set	$X_0 \in R^{4000\times 6}$	Normal operation data, used to test the FD scheme
Simulated faults (6 sets)	$X_0 \in R^{2000\times 6}$	Sensor fault simulations

Table 3.2 Simulated faults.

Fault	Faulty variables	Description of the fault
Fault(1)	X1	Single, Abrupt fault (500 - 2000)
Fault(2)	X2	Single, Abrupt additive fault, (500 - 1500)
Fault(3)	X3	Single, Linear drift (500 - 1500)
Fault(4)	X4	Single, Drift fault (500-900), Abrupt change (100-1400)
Fault(5)	X5	Single, Intermittent fault, changing intervals and amplitudes
Fault(6)	X1, X2, X6	Multiple, additive, abrupt change: X1 (800-1100), X6 (900-1500), Drift fault: X2 (500-700)

3.5.1 Model construction:

For the simulated example, the proposed model consists of 2-states representing healthy and faulty operating modes. To conclude the number of features (the number of retained principal components), we do the following:

- Normalization:

  The mean and standard deviation values of the generated training data-set are the

vectors $\mu_x \in R^{1\times 6}$ and $\sigma_x \in R^{1\times 6}$, respectively

$$\mu_x = [0.9998\ 3.9999\ -2.0002\ 8.0005\ 14.0004\ 6.0004]$$

$$\sigma_x = [0.0100\ 0.0100\ 0.0223\ 0.0581\ 0.0638\ 0.0446]$$

These two vectors are used to shift and scale the data matrix X to zero mean and unity variance. The new matrix obtained throughout this operation is denoted by X_n which represents the normalized data matrix.

- Computation of singular values:

  The covariance matrix (6×6) of the normalized data can be obtained using equation 3.40. Then using Singular Value Decomposition (SVD), the eigen values and their corresponding eigen vectors are computed:

$$\Lambda = diag(4.9636\ \ 1.03625\ \ 0.0001\ \ 6.62e-06\ \ 2.40e-16\ \ 7.27e-17)$$

- Dimension identification:

  Over various methods discussed in the previous section, Kaiser criterion is used in this example for its simplicity.

  The number of retained principal Components is 2, representing a total of variance of 99.99 %.

- Observables extraction:

  The scores matrix T is calculated using equation 3.45. Then $\hat{T}$ is obtained using equation 3.47, to be used as observables to built the Markov model.

3.5.2 Model Training:

- Using the $hist$ function with a suitable number of $bins$, we calculate the joint probabilities of our obsevables using equation 3.51.

- The same work is repeated in order to calculate the joint probabilities for faulty data.

After constructing the Hidden Markov Model with the 2 extracted features, it has been tested and the performance of the proposed scheme is evaluated: (i) fault Detection Time Delay (DTD): (time required for indicating the fault after its occurrence).
(ii) False Alarms Rate (FAR):

$$FAR(\%) = \frac{N_{N.F}}{N_N} \times 100 \tag{3.54}$$

$N_{N.F}$: is the number of normal samples detected as faults.

N_N: is the number of normal samples.

(iii) Missed Detection Rate (MDR):

$$MDR(\%) = \frac{N_{F.N}}{N_F} \times 100 \tag{3.55}$$

$N_{F.N}$: is the number of faulty samples detected as normal.

N_F: is the number of faulty samples.

The objective of a reliable monitoring scheme is to achieve shortest delays for detection, moreover, lowest FAR and MDR [52].

The FAR for the training and testing sets is represented in Table 3.3.

Table 3.3 False alarms rate for the training and testing sets

Data set	Training set	Testing set
False Alarms Rate	1.183	1.225

3.5.3 Model Testing:

The model is tested using some simulated faults which are presented in Table 3.2.

- Figures 3.8 to 3.13 contain 2 plots:
  - The first shows the change in the two retained observables (T1,T2).
  - While the second one represents the fault indicator.

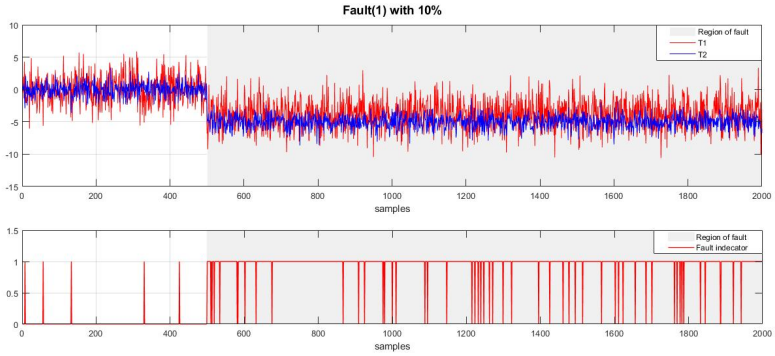

Fig. 3.8 Fault (1) with 10 % amplitude

- The area with the shading shows the region of the occurrence of the fault.
- Fault (1) represented in figures 3.8 to 3.10 with different amplitudes (10%, 15% and 20%). The fault from (500 - 2000).

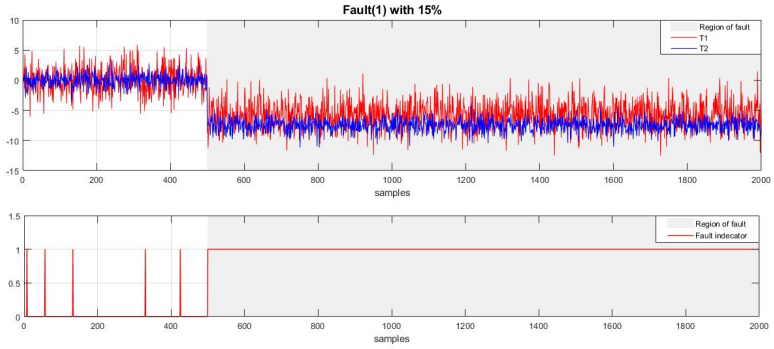

Fig. 3.9 Fault (1) with 15 % amplitude

- Figures bellow from 3.11 to 3.13 corresponds to fault (6):
- The Detection Time Delay, False Alarms Rate and Missed Detection Rate are computed and recorded in Table 3.4. All faults with different amplitudes are presented in the table bellow:

Table 3.4 DTD, FAR and MDR for the simulated faults

Faults	10% Fault			15% Fault			20% Fault		
amplitude	DTD(s)	FAR(%)	MDR(%)	DTD(s)	FAR(%)	MDR(%)	DTD(s)	FAR(%)	MDR(%)
Fault(1)	0	1	3.2967	0	1	0	0	1	0
Fault(2)	0	1.4	0	0	1.4	0.29	0	1.4	0
Fault(3)	215	1.3	97.6074	158	1.3	92.60	18	1.3	82.7173
Fault(4)	82	1.4	81.2968	100	1.4	49.2519	100	1.4	31.6708
Fault(5)	1	1.5	26.3682	0	1.5	14.262	0	1.5	10.9453
Fault(6)	15	1.2	39.6896	7	1.2	13.3038	7	1.2	2.6608

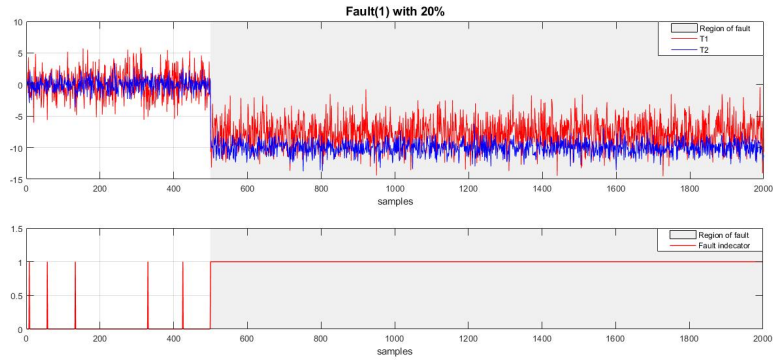

Fig. 3.10 Fault (1) with 20 % amplitude

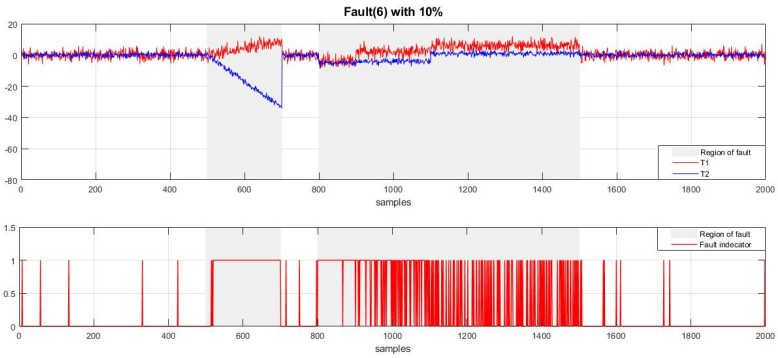

Fig. 3.11 Fault (6) with 10 % amplitude

3.5.4 Discussion

From the result shown in figures 3.8 to 3.13, tables 3.3 and 3.4, the false alarm rate is acceptable however, those values are literally high due to the fact that the used data are generated randomly.

From Table 3.4, the faults in X1 and X2 are well detected in comparison with other faults since X1 and X2 are dominant variables so they affect T1 and T2 strongly, as it is shown in figures 3.8 - 3.10 (the upper plot).

The Single sensor fault of intermittent type (fault(5)) is promptly detected, unlike fault(3) and fault(4) are detected after a long time and consequently with high missed detection rate.This is due to their drift type with a small slope.

The delay time detection for the multiple sensors fault is relatively acceptable. In addition, the missed detection rate was high for fault(6) with 10% of magnitude, than, it rapidly decreased when the fault amplitude increased.

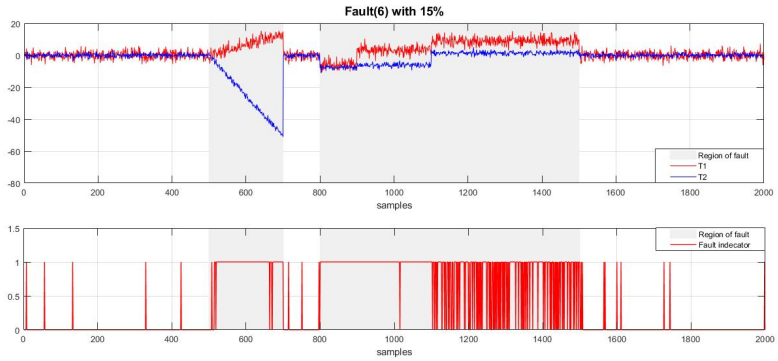

Fig. 3.12 Fault (6) with 15 % amplitude

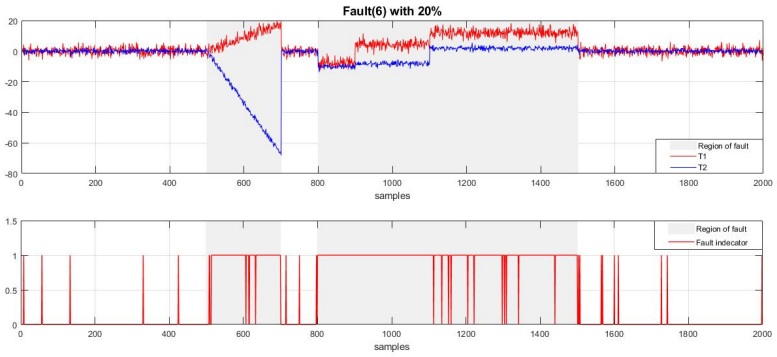

Fig. 3.13 Fault (6) with 20 % amplitude

3.6 Conclusion

In this chapter, A detailed study about HMMs was provided with an illustrated example, the three main problems with Hidden Markov Models were introduced and solutions are discussed in detailed.

Reduction of dimensionality and its aim were briefly discussed and how statistical multivariate monitoring is used in order to extract significant information from data sets. The proposed monitoring scheme was explained in detail and it has been tested using a numerical example, results were introduced.

In the next chapter, the application of HMM based fault detection in real industrial process will be discussed.

CHAPTER 4

The application of HMM based Fault Detection in Industrial Process

4.1 Introduction

In this chapter the proposed method is tested in industrial environment. The industrial case study is the detection of abnormal operating condition (Faulty mode) in a rotary kiln of a cement plant. A detailed description of the cement plant is provided.

The application is also described in details, by indicating types and sizes of data sets collected to construct and evaluate the performance of the monitoring scheme.

4.2 Process Description

Cement production plant is composed of five areas, beginning with the quarry where the limestone rocks are dug out and crushed, comes after the raw meal area where limestone, clay and iron ore mix are grinded to some fineness. Then the material is fed to the cook area where material undertakes many physical and chemical operations, such as drying, dehydration, decomposition of carbonates. Some reactions are solid others involves the liquid phase (sintering) and third are achieved while cooling.

Almost the physical transformations happen in the preheat tower whereas the main chemical reactions happen in the rotary kiln under high temperature gradient. The product of this area is called clinker. Clinker will be mixed to other additives and undertakes finish grinding in the subsequent area to produce the cement. The fifth area is dedicated to packing and expedition.

Ain El Kebira cement plant in the Algerian east, where the work is conducted, is a dry process consisting of four cyclone levels in the suspension preheater and short rotary kiln of 5.4 (m) shell diameter (without brick and coating) and 80 (m) length, with 3 incline. The kiln is spun up to 2.14 (rpm) using two 560 (kws) asynchronous motors and producing clinker of density varying from 1300 to 1450 (kg/m3) under normal conditions. Two natural gas burners are used, the main one in the discharge end and the secondary in the first level of preheater tower.

The main signals that monitor the kiln system are described in Appendix B. They were collected from the historian of the plant [52].

4.3 Application of the proposed monitoring scheme

In this section the proposed technique is applied for monitoring the industrial cement plant. Its different variables are measured to construct a reliable monitoring scheme and evaluate its detection performance.

This work is based on the real-time data collected by process computers. Even though the data collected contain noise processing and measurement noise, it is not necessary for the application to pre-process and filter data, since PCA is a noise decoupling method.

A schematic diagram is provided in figure 4.1 to summarize our work:

Table 4.1 Data sets used for this application.

Data sets	Size	Sampling interval (s)	description
Training set	$X_0 \in R^{768\times16}$	20	Normal operation data, used to construct FD scheme
Testing set	$X_0 \in R^{11000\times16}$	1	Normal operation data, used to test the FD scheme
Process fault	$X_0 \in R^{2084\times16}$	1	Normal/Faulty operation, process fault
Sensor faults (6 sets)	$X_0 \in R^{1500\times16}$	1	Sensor fault simulations

Table 4.1 lists the different data sets used in order to construct and test the proposed fault detection scheme then evaluate and compare its performance.

Set 1, Training data: This set consists of 16 variables listed in Appendix B. About

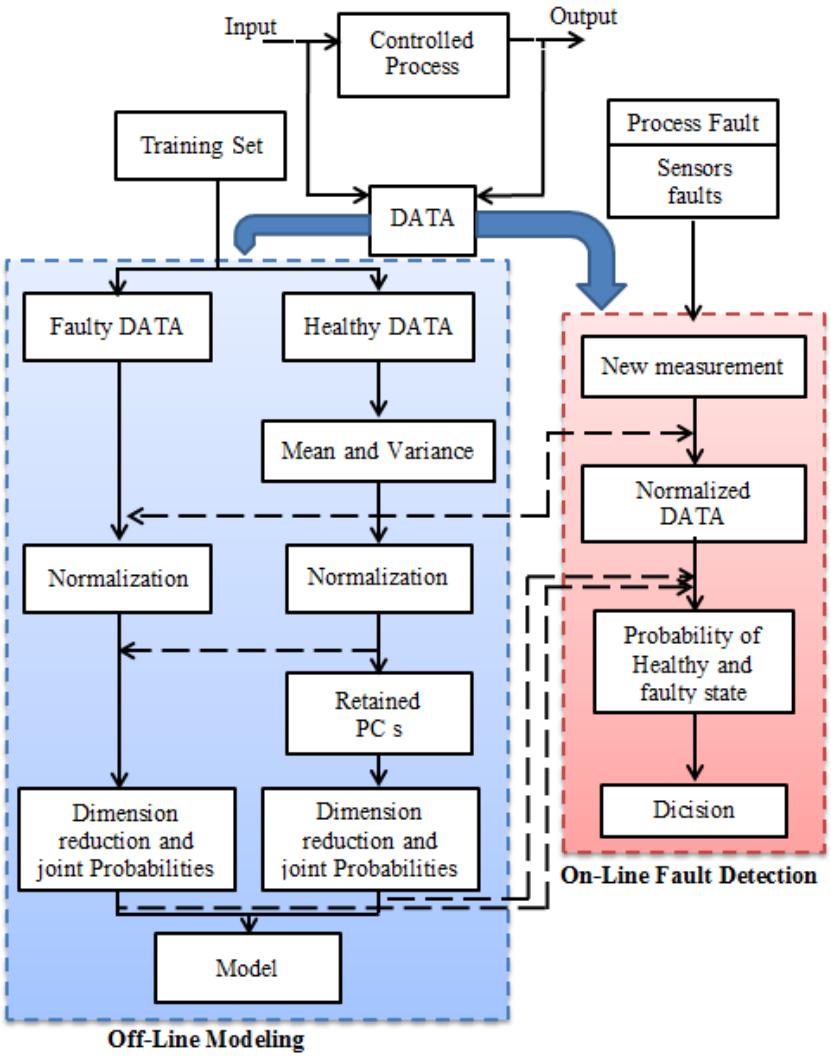

Fig. 4.1 Schematic diagram

770 observation samples (more than 4 h) are collected during healthy operation, with a sampling rate of 1 sample each 20 s.

This set is used to construct the proposed fault detection scheme:

- Extraction of principal components.
- Computing the joint probabilities of healthy state.
- Building the Hidden Markov Model.

Set 2, Testing data: The testing data set consists of 11 000 samples, collected from the plant during healthy operation with a sampling interval of 1 s [52]. This data set is used to test the accuracy and noise rejection of the proposed fault detection scheme by

calculating the False Alarms Rate.

Set 3, Faulty process data: The faulty data is collected from the Cement plant for more than half of an hour (2084 s) with a sampling interval of 1 s. We use this data to check the ability of the proposed detection scheme to detect the fault.

Set 4, Seven simulated sensor/actuator faults: This set contains 7 simulated sensor faults occurring in the rotary kiln process. It includes abrupt, random, intermittent, and slow drift additive faults that might be single as well as multiple faults. Each simulation covers 1500 s of process operation.

The original data is taken during healthy operation of the rotary kiln, and then faults are introduced from 500^{th} sample to 1000^{th} sample. Thus each simulation has three regions (healthy, faulty, healthy). Each region lasts for 500 s. The simulation is done by adding a fault of magnitude 2 % to the data collected during normal (healthy) operation. Table 4.2 lists the 7 sensor faults, with type and magnitude of each fault.

Sfault(1) is a random fault of mean 0 and variance 0.05 added to variable 16, which represents the speed of the induced draft fan installed in top of tower II for drawing gas from preheater cyclones to filters. This fault simulates a loss of accuracy in the speed sensor. It can also represent a vibration in the fan resulting in a small oscillation in the speed without changing its average value.

Sfault(2) is an abrupt bias of 2% in feed of natural gas (variable 16) to the secondary burner which can happen due to many reasons: (i) a partial blockage in the natural gas tubes, (ii) leakage of natural gas, or (iii) just a bias error in sensor readings. Therefore, it has a bad impact on performance and safety of the process.

Sfault(3) is a positive drift in the speed of the first blowing fan located in the cooling unit (variable 13), where the measured speed increased linearly from 0% to 2% in 500 s.

Sfault(4) is a negative drift in the speed of the third blowing fan similar to Sfault3. This fault may change the speed of cooling the clinker and result in a bad quality product.

Sfault(5) and **Sfault6** represent multiple abrupt faults of magnitude 2%, occurring simultaneously at different sensors in positive and negative directions, respectively.

Sfault(7) is a positive intermittent fault in the speed of the exhauster fan of tower I, this actuator/sensor fault has a changing interval (non-periodic) with changing amplitude from 4.5% to 5.5%. These changes in speed can be caused by the actuator due to malfunctions in the drives semiconductors or vibrations on the fans shaft, they can also be caused by the speeds sensor (tachometer). This fault is introduced in the intervals: 500-580, 610-660, 700-740, 800-830, 870-900, and 975-1000 with amplitudes 5.5%, 4.5%, 5%, 5.5%, 5%, and 4.5% respectively.

Table 4.2 Simulated sensor faults, introduced at 500-1000 s.

Fault	Faulty variables	Fault magnitude	Description of the fault
SFault(1)	10	(0, 5 %)	Additive random fault, with mean 0, and variance 0.05
SFault(2)	16	-2%	Abrupt additive fault, bias b = - 0.02
SFault(3)	13	+2 %	Additive fault: Linear drift from 0% to 2%; slope $K_s = 4 \times 10^{-5}$
SFault(4)	14	-2%	Additive fault: Linear drift from 0% to 2%; $slope = -4 \times 10^{-5}$
SFault(5)	8, 11, 15	[+,-, +]2%	Abrupt additive fault 2% (multiple)
SFault(6)	4, 5, 6, 9, 12	[+,+,+,-,-]2%	Additive fault: Linear drift from 0% to $\pm 2\%$ (multiple); $slope = \pm 4 \times 10^{-5}$
SFault(7)	7	+4.5% -5.5%	Additive fault: Intermittent fault, changing intervals and amplitudes

4.3.1 Model construction

The proposed model consists of 2-states corresponding to 2-operating modes: Healthy mode and Faulty mode. the 1^{st} set with 768 samples is used to construct our model and calculate the joint probabilities.

To conclude the number of observables (the number of retained principal components), the procedure described in figure 4.1 will be followed.

- Normalization:

  The mean and standard deviation values of the healthy data (set 1) are calculated, (Mean $\in R^{1\times 16}$ and Std $\in R^{1\times 16}$). These two vectors are used to shift and scale the data matrix X to zero mean and unity variance.

The new matrix obtained throughout this operation is denoted by X_n which represents the normalized data matrix, where its (16×16) covariance matrix can be obtained using equation 3.40. Then using SVD, The singular values and their corresponding eigen vectors are computed.

- Dimension identification:

  Several techniques for determining the relevant number of principal components have been proposed which help to separate the informative sub-space from the residual one. The Variance of Reconstructed Error (VRE) is chosen in our work and applied to select the optimum number of components to keep for building the HMM. Unlike most other methods reported in the literature, the VRE method has a guaranteed minimum over the number of scores corresponding to the best reconstruction [61]. Therefore, it avoids the arbitrariness of other methods with monotonic indices [63].

- Observables extraction:

  using VRE technique where the number of retained PCs is 5 which explain a total of 66 % of variance, The scores matrix T is calculated using equation 3.45. Then $\hat{T}$ is obtained using equation 3.47, to be used in extracting our observables for built the Markov model.

4.3.2 Model Training:

The joint probabilities of the 5 extracted obsevables are calculated for both healthy data and faulty data.

After constructing the Hidden Markov Model with the 5 extracted observables, it has been tested.

The performance of the proposed scheme is evaluated : (i) fault detection time delay, (ii) False Alarms Rate (FAR), and (ii) Missed Detection Rate (MDR).

The FAR is calculated under healthy operation of a cement plant. It is evaluated for the training and testing sets. Results are presented in fig. 4.2 - 4.3 and table 4.3.

These false alarms are due to the transient state during a change in the process

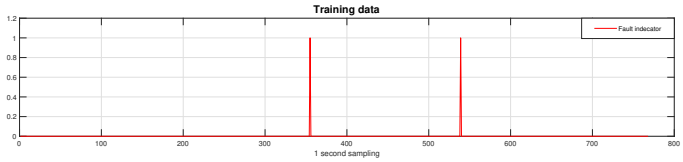

Fig. 4.2 Fault indicator monitoring results of normal (healthy) process operation, Far for Training data set

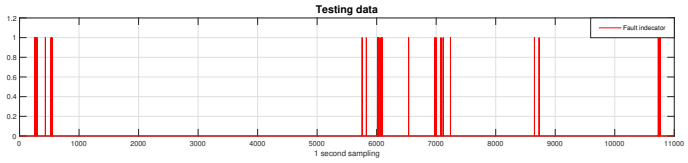

Fig. 4.3 Fault indicator monitoring results of normal (healthy) process operation, FAR for Testing data set

dynamics, where some set-points are changed by the operators.

4.3.3 Model Testing:

The model is tested using a real process fault and some simulated faults which are presented in Table 4.2.

4.3.3.1 Real Process Fault Detection

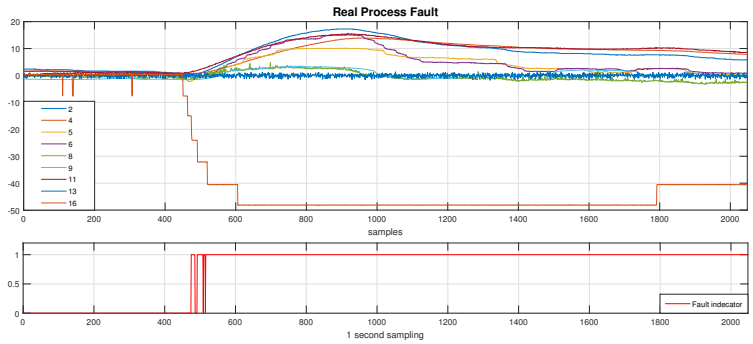

Fig. 4.4 Some affected variables by the real process fault , Fault indicator monitoring results of a real involuntary process fault in the cement rotary kiln.

Figure 4.4 contain 2 plots:

- The first shows the deviation of Some affected variables by the real process fault.
- While the second one represents the fault indicator.

Table 4.3 FAR for the training and testing sets Using the proposed technique

Data set	Training set	Testing set
False Alarm Rate (%)	0.2604	0.3090

Results

The proposed fault detection scheme indicates the occurrence of the real process fault at its early stages, before it evolves and spreads to other process and increases its effects. This fault was detected at a convenient time as shown in fig. 4.4. The figure shows the change in Some affected variables by the real process fault and the fault indicator monitoring results of the real involuntary process fault in the cement rotary kiln. The fault was detected at 475s with Missed Detection Rate of 0.3178 %.

4.3.3.2 Simulated sensor fault detection

In this section, the performance of the fault detection scheme is evaluated in detecting single as well as multiple sensor faults of abrupt, random, intermittent, and drift types.

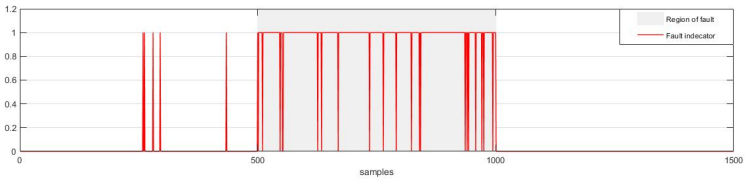

Fig. 4.5 Fault indicator monitoring results of single sensor fault of random type

Fig. 4.6 Fault indicator monitoring results of single sensor fault of Drift type

- The area with the shading shows the region of the occurrence of the fault.
- The injection of faults is from (500 - 1000).
- Figures from 4.5 to 4.8 represents fault indicators for some of the simulated faults.

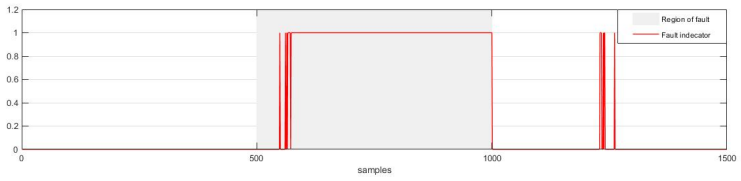

Fig. 4.7 Fault indicator monitoring results of Multiple sensor fault of Drift type

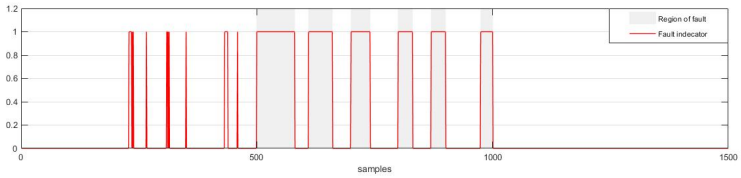

Fig. 4.8 Fault indicator monitoring results of single sensor fault of Intermittent type

Table 4.4. lists the results of Delay Time Detection, False Alarms Rate and The missed Detection Rate for the 7 simulated faults described in Table 4.2.

Results

Table 4.4 Detection Time Delay, False Alarm Rate and Miss Detection Rate for the Simulated Faults.

Faults	DTD (s)	FAR (%)	MDR (%)
Sfault(1)	0	0.5	3.6
Sfault(2)	0	0	6.4
Sfault(3)	348	0	88.4
Sfault(4)	30	0.5	6
Sfault(5)	0	0.2	29
Sfault(6)	48	0.4	12.6
Sfault(7)	0	1.2	0

For random fault (SFault1) ,abrupt fault in single sensor (SFault2) and intermittent fault (SFault7) or multiple sensors (SFault5), the faults are promptly detected regardless of their small amplitudes.

However, the detection time of the drift faults in single sensor (SFault3, SFault4) or multiple sensors (SFault6) was delayed, which means the faults are not detected until they reach a certain level. Taking into consideration that these faults are simulated with slope of $\pm 4 \times 10^{-5}$, the faults are successfully detected by the model after reach amplitude of 1%. The intermittent fault SFault7 is accurately detected in all faulty region.

4.4 Discussion

The industrial process of the cement plant rotary kiln used for our work contains sixteen (16) variables which represents different sensors (mentioned in Appendix .B). The aim of the work is to develop a monitoring technique to detect different types of faults in the process (sensors and actuator faults).

Due to the complexity of the industrial process, it can't be modeled using mathematical models, so process history based fault diagnosis is needed. The proposed technique of fault detection based on Hidden Markov Model is applied to the rotary kiln process.

The Hidden Markov Model consists of states and observations, in this case the states represents the operating modes while observations are extracted from the data collected from sensors. To extract those observables, joint probabilities of all variables are calculated, this leads to N^m value, where N is the number of sample used for training the model and m is the number of variables.

In our case, huge calculations must be done using Set 1 for the training phase, the joint probabilities will be 768^{16} and this is unacceptable because this leads to a large time complexity algorithm that needs a large station to do this work and this is not available. So, we need a tool to reduce the number of variable while keeping most of the information and accuracy.

Principal Component Analysis is used in our work to achieve this purpose. PCA is used to convert the data from the original space to an artificial space organized so that, some of the components are containing more information (variation) than other ones, keeping only the principal components will lead to a dimension reduction. The new dimensions corresponds to the chosen technique used for dimension identification (techniques introduced in chapter .3, section .2). In our case we needed a minimum number of PCs in order to be able of training our model and calculating the joint probabilities, this was achieved using The variance of the reconstruction error (VRE) technique, which gave us a minimum number of PCs that is 5, while other methods like kaiser gave as a minimum number of retained PCs of 6.

After extracting the scores matrix, the joint probability matrix is computed with

suitable number of $bins$, this latter has a great role in the model training, the larger number of $bins$, the smaller is the distance between probabilities, the larger is the joint probability matrix, this leads to a high precision when computing the conditional probabilities in the testing phase as mentioned in the description of the scheme (chapter .3, section .3). In our work, the maximum number of $bins$ we could use is 50 due to the absence of computation station, and this value is too small and can not be considered as a sufficient one.

This is why drift faults and random faults with small amplitudes can not be well detected because if the number of $bins$ takes small values, probabilities for faulty data will be too close to the ones of healthy data or it will be the same, so when the process checks the nature of a collected data, it may consider it as healthy data while it is a faulty one those miss-classifications leads to a high values of FAR, while the inverse will lead to high values of MDR. This will justify the obtained results, why these types of faults are not detected as expected in comparison with other types. The probabilistic model is accurate and more precise than other techniques but, it need a huge data computation.

4.5 Conclusion

The developed method have been applied on the cement production plant. The proposed model uses some historical data collected from cement rotary kiln to build a suitable model for the on-line fault detection.

The proposed scheme works effectively since it reduces the number of false alarms and detects the abnormal changes in cement rotary kiln process efficiently and effectively.

CHAPTER 5

General Conclusion

With the improvement of technology and automation in industry, large-scale, high risk and multi-variable characteristics are presented in the production process. This makes it extremely essential for control systems to be more robust to ensure and guarantee the safety and reliability of production process.

Fault detection role is to determine whether there has been an error. Early detection can give a valuable alert about the issues that arise, and suitable moves to be taken can prevent serious damages and reduce accident rate. Over the past few centuries, data-driven based fault detection approaches have been extensively researched due to their ease compared to model-based techniques (the non-requirement of a mathematical model).

We first provided a fundamental theory of fault detection and diagnosis, and we briefly described various methods used for fault detection and diagnosis exploring many definitions and aspects in the domain of fault detection.

Then we launched HMM's as a tool that enables us to accurately and usefully define the real-life system, Then we addressed three canonical questions concerning HMMs and we showed how they can be resolved in a tractable manner. This gave rise to the three well known algorithm that are the Forward-Backward algorithm, the Viterbi algorithm and the Baum-Welch algorithm. In addition, Principal Component Analysis was described and it has been used as dimension reduction tool.

The main focus of this thesis is the development of a probabilistic model based on Hidden Markov Model for the purpose of fault detection and isolation. HMMs, by their nature, provide an appropriate framework to extract temporal information related

to process transitions between various modes. The proposed model was used to provide a monitoring for a cement rotary kiln system. A methodology for process monitoring was presented in this work with static transition probabilities and tested on a numerical example and the cement plant rotary kiln process. The application of the proposed scheme on the industrial process gives some good results, it was able to detect the real process fault as well as the simulated faults. Some of them was detected with zero time delay and approximately zero False alarms. Unfortunately, the proposed scheme suffers in detecting random and drift type faults with small amplitudes.

As a future work, the development of an adaptive Hidden Markov Model with dynamic Transition probabilities with large number of modes. Also, we recommend that the use of the Hidden Markov Model be extended to other industrial process where other kinds of faults may occur.

References

[1] Rolf Isermann. Model-Based Fault Detection and Diagnosis - Status and Applications. *IFAC Proc. Vol.*, 37(6):49–60, jun 2004.

[2] L.R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE*, 77(2):257–286, 1989.

[3] Leo H. Chiang, Evan L. Russell, and Richard D. Braatz. *Fault Detection and Diagnosis in Industrial Systems*. Advanced Textbooks in Control and Signal Processing. Springer London, London, 2001.

[4] Rolf Isermann. *Fault-Diagnosis Systems*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006.

[5] Anna Pernestl, Hannes Wettig, Tomi Silander, Mattias Nyberg, and Petri Myllymki. A bayesian approach to learning in fault isolation. *CEUR Workshop Proceedings*, 406, 01 2008.

[6] Wei Zhang. *Fault Detection*. 2010.

[7] Padhraic Smyth. Hidden Markov models for fault detection in dynamic systems. *Pattern Recognit.*, 27(1):149–164, jan 1994.

[8] Venkat Venkatasubramanian, Raghunathan Rengaswamy, Surya N. Kavuri, and Kewen Yin. A review of process fault detection and diagnosis part III: Process history based methods. *Comput. Chem. Eng.*, 27(3):327–346, 2003.

[9] Venkat Venkatasubramanian, Raghunathan Rengaswamy, and Surya N. Kavuri. A review of process fault detection and diagnosis part II: Qualitative models and search strategies. *Comput. Chem. Eng.*, 27(3):313–326, 2003.

[10] Rolf Isermann. Process fault detection based on modeling and estimation methods-A survey. *Automatica*, 20(4):387–404, 1984.

[11] R. Clark. A Simplified Instrument Failure Detection Scheme. *IEEE Trans. Aerosp. Electron. Syst.*, AES-14(4):558–563, jul 1978.

[12] Alan S Willsky. A survey of design methods for failure detection in dynamic systems. *Automatica*, 12(6):601–611, nov 1976.

[13] J Gertler. Fault detection and isolation using parity relations. *Control Eng. Pract.*, 5(5):653–661, may 1997.

[14] Kajiro Watanabe, Ichiro Matsuura, Masahiro Abe, Makoto Kubota, and D. M. Himmelblau. Incipient fault diagnosis of chemical processes via artificial neural networks. *AIChE J.*, 35(11):1803–1812, nov 1989.

[15] J. C. Hoskins, K. M. Kaliyur, and David M. Himmelblau. Fault diagnosis in complex chemical plants using artificial neural networks. *AIChE J.*, 37(1):137–141, jan 1991.

[16] David Mautner Himmelblau. *Fault Detection and Diagnosis in Chemical and Petrochemical Processes, Bd. 8 der Serie Chemical Engineering Monographs. VonD. M. Himmelblau, herausgegeben vonS. W. Churchill, Elsevier Scientific Publishing Company, Amsterdam - New York 1978. 1. Aufl., X*, volume 51 of *Chemical engineering monographs*. Elsevier Scientific Pub. Co., jul 1979.

[17] Shen Yin, Steven X. Ding, Xiaochen Xie, and Hao Luo. A Review on Basic Data-Driven Approaches for Industrial Process Monitoring. *IEEE Trans. Ind. Electron.*, 61(11):6418–6428, nov 2014.

[18] Shen Yin, Xianwei Li, Huijun Gao, and Okyay Kaynak. Data-Based Techniques Focused on Modern Industry: An Overview. *IEEE Trans. Ind. Electron.*, 62(1):657–667, jan 2015.

[19] D. Garcia-Alvarez, M.J. Fuente, P. Vega, and G. Sainz. Fault Detection and Diagnosis using Multivariate Statistical Techniques in a Wastewater Treatment Plant.*

*This work was supported in part by the national research agency of Spain (CICYT) through the project DPI2006-15716-C02-02 and the regional governmen. *IFAC Proc. Vol.*, 42(11):952–957, 2009.

[20] Paul Geladi and Bruce R. Kowalski. Partial least-squares regression: a tutorial. *Anal. Chim. Acta*, 185:1–17, 1986.

[21] Emanuel Parzen. On Estimation of a Probability Density Function and Mode. *Ann. Math. Stat.*, 33(3):1065–1076, sep 1962.

[22] Keinosuke Fukunaga. *Introduction to statistical pattern recognition, Second Edition*. Academic Press, 2 edition, 1990.

[23] Sylvain Verron, Teodor Tiplica, and Abdessamad Kobi. Fault detection of univariate non-Gaussian data with Bayesian network. In *2010 IEEE Int. Conf. Ind. Technol.*, pages 94–99. IEEE, 2010.

[24] Rolf Isermann. Process fault detection based on modeling and estimation methods-A survey. *Automatica*, 20(4):387–404, 1984.

[25] Nader Meskin and Khashayar Khorasani. Fault detection and isolation multivehicle unmanned systems. 2011.

[26] Bo Tang, Xiaobin Tan, and Baoqun Yin. Continuous-time hidden markov models in network simulation. *2008 IEEE Int. Symp. Knowl. Acquis. Model. Work. Proceedings, KAM 2008*, pages 667–670, 2008.

[27] Xuefeng Jiang. A facial expression recognition model based on HMM. *Proc. 2011 Int. Conf. Electron. Mech. Eng. Inf. Technol. EMEIT 2011*, 6:3054–3057, 2011.

[28] Soo Woong Lee and B. V.K. Vijaya Kumar. Application of pattern-output viterbi algorithm to algebraic soft-decision decoding over partial response channels. *IEEE Commun. Lett.*, 14(5):483–485, 2010.

[29] Nima Sammaknejad, Biao Huang, Weili Xiong, Alireza Fatehi, Fangwei Xu, and Aris Espejo. Operating condition diagnosis based on HMM with adaptive transi-

tion probabilities in presence of missing observations. *AIChE J.*, 61(2):477–493, feb 2015.

[30] Nima Sammaknejad, Biao Huang, and Yaojie Lu. Robust Diagnosis of Operating Mode Based on Time-Varying Hidden Markov Models. *IEEE Trans. Ind. Electron.*, 63(2):1142–1152, 2016.

[31] Nima Sammaknejad, Biao Huang, Alireza Fatehi, Yu Miao, Fangwei Xu, and Aris Espejo. Adaptive monitoring of the process operation based on symbolic episode representation and hidden Markov models with application toward an oil sand primary separation. *Comput. Chem. Eng.*, 71:281–297, dec 2014.

[32] Zoubin Ghahramani. An Introduction to Hidden Markov Models and Bayesian Networks. *Int. J. Pattern Recognit. Artif. Intell.*, 15(01):9–42, feb 2001.

[33] Zhang Xueying, Wang Yiping, and Zhao Zhefeng. A hybrid speech recognition training method for HMM based on genetic algorithm and Baum Welch algorithm. *Second Int. Conf. Innov. Comput. Inf. Control. ICICIC 2007*, pages 25–28, 2008.

[34] Mohamed Zraiaa. *Hidden Markov Models : A Continuous-Time Version of the Baum-Welch Algorithm*. Number September. 2010.

[35] Stephen Adams, Peter A. Beling, and Randy Cogill. Feature Selection for Hidden Markov Models and Hidden Semi-Markov Models. *IEEE Access*, 4:1642–1657, 2016.

[36] M. Soruri, S. Hamid Zahiri, and J. Sadri. A new approach of training Hidden Markov Model by PSO algorithm for gene Sequence Modeling. In *2013 First Iran. Conf. Pattern Recognit. Image Anal.*, pages 1–4. IEEE, mar 2013.

[37] Mohd Izhan Mohd Yusoff, Ibrahim Mohamed, and Mohd Rizam Abu Bakar. Hidden Markov models: An insight. *Conf. Proc. - 6th Int. Conf. Inf. Technol. Multimed. UNITEN Cultiv. Creat. Enabling Technol. Through Internet Things, ICIMU 2014*, pages 259–264, 2015.

[38] Padhraic Smyth. Hidden Markov models for fault detection in dynamic systems. *Pattern Recognit.*, 27(1):149–164, jan 1994.

[39] Zhijun Li, Jiang Zhong, Cunwu Han, and Dehui Sun. Process fault detection based on continuous hidden Markov model. *Proc. - 2017 Chinese Autom. Congr. CAC 2017*, 2017-Janua(1):2755–2760, 2017.

[40] Nima Sammaknejad. *Fault Detection and Isolation Based on Hidden Markov Models*. Thesis, University of Alberta, 2015.

[41] Eric Fosler-lussier. Markov models and hidden markov models: A brief tutorial. 01 1999.

[42] Leon Gu. EM and HMM.

[43] Catherine Sweeney-Reed. Hidden Markov Models.

[44] Tapas Kanungo. Hidden Markov Models. *Cent. Autom. Res. Univ. Maryl.*, 1999.

[45] James C. Wong, Karen A. McDonald, and Ahmet Palazoglu. Classification of process trends based on fuzzified symbolic representation and hidden Markov models. *J. Process Control*, 8(5-6):395–408, oct 1998.

[46] James C. Wong, Karen A. McDonald, and Ahmet Palazoglu. Classification of abnormal plant operation using multiple process variable trends. *J. Process Control*, 11(4):409–418, aug 2001.

[47] Rui Jiang, Jing Yu, and Viliam Makis. Optimal Bayesian estimation and control scheme for gear shaft fault detection. *Comput. Ind. Eng.*, 63(4):754–762, dec 2012.

[48] Jie Ying, T. Kirubarajan, K.R. Pattipati, and A. Patterson-Hine. A hidden Markov model-based algorithm for fault diagnosis with partial and imperfect tests. *IEEE Trans. Syst. Man Cybern. Part C (Applications Rev.*, 30(4):463–473, 2000.

[49] Alireza Ghasemi, Soumaya Yacout, and M-Salah Ouali. Parameter Estimation Methods for Condition-Based Maintenance With Indirect Observations. *IEEE Trans. Reliab.*, 59(2):426–439, jun 2010.

[50] Wee Chin Wong and Jay H. Lee. Fault Detection and Diagnosis Using Hidden Markov Disturbance Models. *Ind. Eng. Chem. Res.*, 49(17):7901–7908, sep 2010.

[51] Mustapha Ammiche, Abdelmalek Kouadri, and Abderazak Bensmail. A Modified Moving Window dynamic PCA with Fuzzy Logic Filter and application to fault detection. *Chemom. Intell. Lab. Syst.*, 177:100–113, 2018.

[52] Azzeddine Bakdi, Abdelmalek Kouadri, and Abderazak Bensmail. Fault detection and diagnosis in a cement rotary kiln using PCA with EWMA-based adaptive threshold monitoring scheme. *Control Eng. Pract.*, 66(March):64–75, 2017.

[53] I T Jolliffe. *Principal Component Analysis*, volume 98 of *Springer Series in Statistics*. Springer-Verlag, New York, 2002.

[54] E Ding, P Deng, W Gui, P Zhang, S Ding, and A Naik. On the application of PCA technique to fault diagnosis. *Tsinghua Sci. Technol.*, 15(2):138–144, 2010.

[55] Binzhang Lu, Yuhong Zhao, and Zhenhua Mao. Fault diagnosis method based on moving window PCA. In *2009 Chinese Control Decis. Conf.*, Springer Series in Statistics, pages 185–188, New York, jun 2009. IEEE.

[56] Raphael Tari Samuel and Yi Cao. Fault detection in a multivariate process based on kernel PCA and kernel density estimation. In *2014 20th Int. Conf. Autom. Comput.*, pages 146–151. IEEE, sep 2014.

[57] Hanen Chaouch. Fault detection and isolation of ECG process using kernel principal components. In *2017 Int. Conf. Eng. MIS*, number i, pages 1–6. IEEE, may 2017.

[58] Fengxi Song, Zhongwei Guo, and Dayong Mei. Feature selection using principal component analysis. *Proc. - 2010 Int. Conf. Syst. Sci. Eng. Des. Manuf. Informatiz. ICSEM 2010*, 1:27–30, 2010.

[59] Yacine ARIBI and Hamza GUERMI. *PCA-Based Approach for Fault Detection In Cement Rotary Kiln*. PhD thesis, 2016.

[60] Hervé Abdi and Lynne J. Williams. Principal component analysis. *Wiley Interdiscip. Rev. Comput. Stat.*, 2(4):433–459, jul 2010.

[61] BENKARA Bochra Amina BOUNOUA Wahiba. *Online Monitoring Scheme Using PCA through Kullback-Leibler Divergence Analysis Technique for Tennessee Eastman Process Fault Detection*. thesis, 2017.

[62] John L. Horn. A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30(2):179–185, jun 1965.

[63] S.Joe Qin and Ricardo Dunia. Determining the number of principal components for best reconstruction. *J. Process Control*, 10(2-3):245–250, apr 2000.

[64] Sergio Valle, Weihua Li, and S. Joe Qin. Selection of the Number of Principal Components: The Variance of the Reconstruction Error Criterion with a Comparison to Other Methods . *Ind. Eng. Chem. Res.*, 38(11):4389–4401, nov 1999.

[65] Ricardo Dunia and S. Joe Qin. Joint diagnosis of process and sensor faults using principal component analysis. *Control Eng. Pract.*, 6(4):457–469, 1998.

[66] CEI IEC 61508-1 Fuctional safety of electrical saefty reated systems. 1997.

[67] Vamshi Krishna Kandula. Fault detection in process control plant as using principal component analysis. (December), 2011.

[68] Alkan ALKAYA. *Novel Data Driven-Based Fault Detection for Electro Mechanical and Process Control Systems* . PhD thesis, Çukurova University, 2012.

[69] Rolf Isermann. *Fault-Diagnosis Systems: An Introduction from Fault Detection to Fault Tolerance*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006.

Appendices

Appendix A

Terminology

The IFAC-Technical Committee SAFE PROCESS has made an effort to come to accepted definitions:

- **Fault**: A fault is an unpermitted deviation of at least one characteristic property (feature) of the system from the acceptable, usual standard condition.

  Remarks:

  - A fault is a state within the system.
  - The unpermitted deviation is the difference between the fault value and the violated threshold of a tolerance zone for its usual value.
  - A fault is an abnormal condition that may cause a reduction in, or loss of the capability of a functional unit to perform a required function [66].
  - There exist many different types of faults, e.g. design fault, manufacturing fault, assembling fault, normal operation fault (e.g. wear), wrong operation fault (e.g. overload), maintenance fault, hardware fault, software fault, operators fault (Some of these faults are also called errors, especially if directly caused by humans).
  - A fault in the system is independent of whether the system is in operation or not, a fault may not effect the correct functioning of a system (like a small rent in an axle).
  - A fault may initiate a failure or a malfunction.
  - Frequently, faults are difficult to detect, especially if they are small or hidden faults may develop abruptly (step-wise) or incipiently (drift-wise).

A suitable modelling of faults is important for the right functioning of fault-detection methods. Also faults that occur in process plant can be classified into three types:

1. Sensor faults: They represent incorrect reading from the sensors this can be due to broken wires, lost contact with the surface etc. In which case the reading shown by the sensor is not related to the value of the measured physical parameter. This can for instance be a gain reduction, a biased measurement or increased noise. [67, 68].
2. Actuator faults: They represent partial or complete loss of control action. Total actuator fault can occur for instance, as a result of a breakage cut or burned wiring short cuts, or the presence of outer body in the actuator. In case of partially failed actuator only part of the normal actuation is produced, it can be result of hydraulic, pneumatic leakage, reduced input voltage or increased resistance.
3. Component faults: These faults are those faults that are not able to be considered as sensor faults or actuator faults. They occur due to structural damages of the components. The dynamical behavior of the system can be changed because of these faults [69]. And these are the most frequently encountered types in fault family to deal with.

There are other types of faults but, these are the most frequently encountered types in fault family to deal with.

- **Failure**: A failure is a permanent interruption of a systems ability to perform a required function under specified operating conditions.
  Remarks:
  - a failure is the termination of the ability of a functional unit to perform a required function [66].
  - A failure is an event.
  - A failure results from one or more faults.

- Usually a failure arises after the start of the operation or by increasingly stressing the system.

- **Malfunction**:

  A malfunction is an intermittent irregularity in the fulfillment of a systems desired function.

  Remarks:

  - A malfunction is a temporary interruption of a systems function.
  - A malfunction is an event.
  - A malfunction results from one or more faults.
  - Usually a malfunction arises after the start of the operation or by increasingly stressing the system.

Appendix B

Process variables of the cement plant

Signal	Description	Units
1	Depression of gases in outlets of cyclone one, tower I	mbar
2	Temperature of gases in outlets of cyclone one, tower I	$^{\circ}C$
3	Depression of gases in outlets of cyclone two, tower I	mbar
4	Temperature of gases in outlets of cyclone two, tower I	$^{\circ}C$
5	Temperature of gases in outlets of cyclone three, tower I	$^{\circ}C$
6	Temperature of gases in outlets of cyclone three, tower I	$^{\circ}C$
7	Speed of the exhauster fans of tower I	r.p.m
8	Temperature of the material entering the kiln from tower I	$^{\circ}C$
9	Temperature of gas in the outlet of the smoke filter: tower I	$^{\circ}C$
10	Speed of the exhauster fans of tower II	r.p.m
11	Temperature of gases in outlets of cyclone one tower II	$^{\circ}C$
12	Temperature of gases in outlets of cyclone two tower II	$^{\circ}C$
13	Speed of the cooling fan I	r.p.m
14	Speed of the cooling fan III	r.p.m
15	Flow of fuel (natural gas) to the main burner	m^3/h
16	Flow of fuel (natural gas) to the secondary burner (pre-calcination level)	m^3/h