Fault Detection and Isolation Based on Hidden Markov Models

by

Nima Sammaknejad

A thesis submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Process Control

Department of Chemical and Materials Engineering
University of Alberta

Abstract

A large volume of literature exists on fault detection and isolation for industrial processes. In a general view, these various methods may be divided into process model based and process history based fault diagnosis. In both classes, there has been a recent focus on extracting the temporal information corresponding to process transitions between various operating modes. In this context, Hidden Markov Models (HMMs) have been introduced and applied for process monitoring and diagnosis purposes. The main objective of this thesis is to develop novel HMM based approaches to diagnose various operating modes of a process. Mode in this thesis refers to process operational status such as normal operating condition or fault.

Many industrial processes work in various operating modes with an asymmetric temporal behavior, e.g., some of the modes may transit to each other much faster or slower than the other ones. To develop appropriate models for such processes with a better consideration of the transition periods, time varying HMMs, which are functions of some auxiliary scheduling variables, are introduced. The proposed procedure provides a more flexible HMM structure for process monitoring purposes. In addition, under the proposed framework, unlikely process jumps will be avoided while reducing the number of parameters to be estimated, and the computational cost will be reduced. This framework is also able to deal with missing observations in the training data set which might happen due to sensor failures, etc.

Industrial processes are often subject to irregular measurements or outliers. A simple approach to deal with this issue is to remove the irregular measurements from the training data set, and then, develop the model. However, such approaches might result in loss of information. Another approach is to use probability distributions which consider separate components for the regular and irregular data, e.g., mixture probabilistic models. The approach which is taken in this thesis is to consider vari-

ous Student t distributions for different operating modes of the process. Therefore, a scalar weight can be assigned according to the distributions in different modes, e.g., normal and faulty regimes, and effect of outliers in parameter estimation will be downweighted through the heavier tails of the distribution.

In some applications it is required to focus on main features of the signal rather than the details, i.e., there is a need to detect the slow trend of process changes rather than short period fluctuations. Qualitative Trend Analysis (QTA) is one of the proposed approaches in literature to extract such information. Here, we combine QTA, which is performed through triangular representation of signals, and HMMs to extract the key features and temporal information simultaneously. Some hierarchical procedure is introduced to improve the accuracy of the continuous to discrete mapping, and then, HMMs are trained for discrete observations of the different operating regimes, e.g., normal and abnormal conditions. It is shown that the proposed method is also able to isolate some specific types of faults.

For the industrial cases study of this thesis, which is concerned with sand deposition and pipeline plugging in the underflow line of an industrial scale Primary Separation Vessel (PSV), we show that a combination of data driven modeling based on HMMs and first principle knowledge provides an appropriate solution. An appropriate semi-empirical equation is used to estimate the critical velocity to move the solid bed inside slurry pipelines and avoid plugging. A HMM based approach is then proposed to modify the sensitivity of estimated velocities. The proposed method has been tested in on-line environment, and has demonstrated an acceptable performance.

In this thesis, various simulation and lab examples have been used to illustrate the efficiency of the proposed techniques. A comparison between the results in different chapters leads to the conclusion demonstrating effectiveness of the proposed approaches.

Acknowledgements

First and foremost, I would like to thank my supervisor Prof. Biao Huang for all the inspiration and help, not only as a research supervisor, but also as a great life mentor. I would like to greatly acknowledge his patience and encouragement during the times when I had difficulties understanding the concepts. Beside the advanced scholarly knowledge, the very organized and positive attitude toward various problems is what I have inherited from this journey under his supervision.

A significant part of this thesis comes from a tight collaboration with the oil sand industry, specifically, Syncrude Canada Ltd. I would like to thank Aris Espejo and Dr. Fangwei Xu who helped me to gain industrial research experience through providing the process knowledge and industrial data. I greatly appreciate Dr. Miao Yu's help and advice for on site implementation of our algorithm.

I would also like to thank a number of our current and former students in the Computer Process Control group including Mohammad Rashedi, Elham Naghoosi, Marziyeh Keshavarz, Ming Ma, Yaojie Lu, Mulang Chen and Aditya Tulsyan who I have truly enjoyed their friendship.

I would like to acknowledge both Syncrude Canada Ltd and National Science and Engineering Research Council (NSERC) of Canada for their financial support, and the department of Chemical and Materials Engineering at the university of Alberta for providing a pleasant environment to pursue my PhD.

In am thankful to my family for always supporting my decisions. They have always encouraged me to follow my dreams. Specifically, I am grateful to my mother, who without her emotional support, it was impossible to complete this journey. She has always been and will always remain as my hero.

In the end, I would like to thank my lovely companion Sepideh Rejaeirad who has stayed beside me in every step of this difficult journey. Without her heartful support, I really doubt if I could even complete a chapter of this thesis.

Contents

1 Introduction **1**

1.1 Motivation 1

1.2 Previous Studies 4

1.2.1 Process Model Based Fault Diagnosis 4

1.2.2 Process History Based Fault Diagnosis 6

1.2.3 HMM Based Fault Diagnosis 8

1.3 Thesis Outline 10

1.4 Published, Submitted and Under Preparation Materials 12

1.5 Main Contributions 13

2 Mathematical Fundamentals **15**

2.1 Expectation Maximization (EM) Algorithm 15

2.1.1 Monotonicty of the EM Algorithm 15

2.1.2 Convergence Properties of the EM Algorithm 18

2.1.3 Initialization of the EM Algorithm 19

2.1.4 Stopping Criteria for the EM Algorithm 20

2.1.5 Parameter Estimation for Mixture Densities Based on EM - An Example 21

2.2 Hidden Markov Models 23

2.2.1 An Illustrative Example 24

2.2.2 Basic Settings for HMMs 25

2.2.3 Three Fundamental Problems for HMMs 27

3 Operating Condition Diagnosis Based on HMM with Adaptive Transition Probabilities in Presence of Missing Observations **35**

3.1 Introduction 35

3.2 Problem Statement 38

3.3 Parameter Estimation Based on the Expectation Maximization Algorithm 42

3.3.1 Expectation Step . . . 42
3.3.2 Maximization Step . . . 46
3.4 Operating Mode Recognition . . . 49
3.5 Results and discussion . . . 51
3.5.1 A Numerical Case Study . . . 51
3.5.2 A Simulation Study . . . 55
3.5.3 An Industrial Case Study . . . 61
3.6 Conclusion . . . 61

4 Robust Diagnosis of Operating Mode Based on Time Varying Hidden Markov Models 62
4.1 Introduction . . . 62
4.2 Problem Statement . . . 65
4.3 Robust Parameter Estimation . . . 68
4.4 Operating Condition Diagnosis . . . 75
4.5 Results and Discussion . . . 75
4.5.1 Tennessee Eastman Process . . . 76
4.5.2 Experimental Evaluation . . . 81
4.6 Conclusion . . . 85

5 Adaptive Monitoring of the Process Operation Based on Symbolic Episode Representation and Hidden Markov Models 86
5.1 Introduction . . . 87
5.2 Data Preprocessing . . . 90
5.3 Triangular Representation . . . 90
5.4 State Recognition . . . 92
5.4.1 Parameter Estimation Based on the EM Algorithm . . . 94
5.5 Adaptive Fuzzification . . . 98
5.6 Data Classification for Multiple Discrete Observations . . . 100
5.6.1 Data Classification Based on BPNNs . . . 100
5.6.2 Data Classification Based on HMMs . . . 100
5.7 Data Classification Based on HMMs – A Moving Window Approach . 102
5.8 Simulation Case Study . . . 103
5.8.1 Comparison of the BPNN, HMM and HMM with Adaptive Fuzzification . . . 105
5.8.2 Detection of Various Types of Faults Based on HMM and HMM with Adaptive Fuzzification . . . 110

5.8.3 Detection of Size of the Faults Based on the HMM and Adaptive Fuzzification . . . 113
5.8.4 Data Classification Based on a Moving Window of Observations 119
5.9 Industrial Case Study . . . 124
5.10 Conclusion . . . 125

6 Applications to an Industrial Scale Oil Sands Primary Separation Vessel 126
6.1 Introduction . . . 126
6.1.1 Problem Statement . . . 126
6.1.2 Process Overview . . . 127
6.1.3 Definition of the Critical Minimum Velocity . . . 128
6.1.4 Importance of the Critical Minimum Velocity . . . 128
6.1.5 Solution Strategy . . . 129
6.2 Critical Minimum Velocity Estimation . . . 129
6.2.1 Carrier Fluid Density . . . 132
6.2.2 Carrier Fluid Viscosity . . . 132
6.2.3 Coarse Particle Diameter . . . 133
6.3 Soft Sensor Development . . . 134
6.3.1 Recursive Exponentially Weighted PLS (rPLS) . . . 135
6.4 Adaptive Sensitivity Levels for Critical Velocity Estimation . . . 136
6.4.1 Flow rate Mode Diagnosis . . . 137
6.4.2 HMM Training . . . 138
6.5 Results of the Proposed Method . . . 142
6.5.1 Soft Sensor Performance . . . 142
6.5.2 Critical Velocity Estimation . . . 143
6.6 PSV Operating Mode Diagnosis Based on Time Varying HMMs . . . 149
6.7 PSV Monitoring Based on Symbolic Episode Representation and HMMs 152
6.7.1 Upset and Normal Operating Conditions and Variable Selection 153
6.7.2 Training Data for the Normal and Upset Operations . . . 153
6.7.3 Adaptive Fuzzification and Decision Making . . . 154
6.8 Conclusion . . . 160

7 Concluding Remarks and Future Directions 162
7.1 Concluding Remarks . . . 162
7.2 Future Directions . . . 165
7.2.1 Number of Operating Modes . . . 165

7.2.2 Uncertain/ Discrete Scheduling Variable 165
7.2.3 Latent Variable Models in Conjunction with HMMs 166
7.2.4 Conditional Random Fields 166
7.2.5 Missing Data Treatment 167
7.2.6 Model Switching Mechanism 168

Bibliography **169**

A Details of the Derivations in Chapter 3 **178**

List of Tables

3.1 System parameters to generate the simulation data 51
3.2 Estimated parameters from the EM algorithm 52
3.3 Estimated parameters for the CSTRs in series using the EM algorithm 58

6.1 Comparison between the performance of the fixed and recursive PLS soft sensors 143
6.2 Estimated parameters for the industrial case study using the EM algorithm 150
6.3 Parameters of the different modes from the historical data (interface level) 155
6.4 Parameters of the different modes from the historical data (underflow density) 156

List of Figures

1.1 A summary of available fault diagnosis methods in literature [1, 3] . . 3
1.2 A general schematic of model based fault diagnosis [1] 5
1.3 Dynamic Mixture Probabilistic Principal Component Analyzer [21] . 9

2.1 A three states Markov chain process (State 1=Sunny, State 2=Foggy and State 3=Rainy) [41] 25
2.2 Graphical illustration of Forward-Backward algorithm [40] 30

3.1 The diagram of operating mode transitions used for fault detection purpose of this chapter (normal modes: 1 to P, abnormal modes: $P+1$ to Q, faulty modes: $Q+1$ to M) 41
3.2 Validation data-set 53
3.3 Probability of the hidden operating modes for the validation data-set using the proposed method of this chapter 53
3.4 True and estimated hidden operating modes of the process for the validation data-set using the proposed method of this chapter 54
3.5 Probability of the hidden operating mode for the validation data-set based on a conventional multivariate hidden Markov model 54
3.6 True and estimated hidden operating modes of the process for the validation data-set based on a conventional multivariate hidden Markov model 55
3.7 Two CSTRs in series [67] 56
3.8 Different operating modes for the process variables (validation data-set) 57
3.9 The scheduling variable and disturbance to the process (validation data-set) 58
3.10 Probability of the hidden operating modes for the CSTRs in series based on the proposed method of this chapter and the validation data-set in Figure 3.8 59

3.11 True and estimated operating modes of the process for the CSTRs in series based on the proposed method of this chapter and the validation data-set in Figure 3.8 59
3.12 Probability of the hidden operating modes for the CSTRs in series using conventional HMMs and the validation data-set in Figure 3.8 . 60
3.13 True and estimated operating modes of the process for the CSTRs in series using conventional HMMs and the validation data-set in Figure 3.8 60

4.1 Graphical illustration of the proposed model in this chapter 68
4.2 TE Process [87] 76
4.3 Measured observations for monitoring of the TE process 77
4.4 Scheduling variable for monitoring of the TE process 78
4.5 Probability of the observations in Figure 4.3 given each operating mode based on the Gaussian distribution assumption 78
4.6 Estimated and true operating modes of the process based on the observations in Figure 4.3 and the Gaussian distribution assumption . . 79
4.7 Probability of the observations in Figure 4.3 given each operating mode based on the t distribution assumption 79
4.8 Estimated and true operating modes of the process based on the observations in Fig. 4.3 and t distribution assumption 80
4.9 Variations of the baseline transition probabilities for the presented results in Fig. 4.3 80
4.10 Variations of the degree of freedom while increasing the percentage of outliers in the TE process 81
4.11 Balls in tubes experiment 82
4.12 Changes in the fan speeds 1 to 3 after the change in the fan speed 4 . 83
4.13 Probability of the observations in Figure 4.12 given each operating mode based on the Gaussian distribution assumption 83
4.14 Estimated and true operating modes of the process based on the observations in Figure 4.12 and the Gaussian distribution assumption . 84
4.15 Probability of the observations in Figure 4.12 given each operating mode based on the t distribution assumption 84
4.16 Estimated and true operating modes of the process based on the observations in Figure 4.12 and t distribution assumption 85

5.1 The proposed process monitoring approach in this study 89
5.2 The procedure of adaptive fuzzification and state recognition 89

5.3 A sample episode for describing process trends [19] 91
5.4 Seven types of triangles [19] 91
5.5 'A' Triangle transformed to 9 sub-types using appropriate fuzzy rules and membership functions [19] 92
5.6 Combination of the fuzzy membership functions for different modes . 99
5.7 Optimal window size selection 104
5.8 Summary of the proposed methodology of this paper 104
5.9 CSTR reactors in series [67] 105
5.10 Normal and abnormal operations in the CSTRs in series - Feed flow rate 106
5.11 Normal and abnormal operations in the CSTRs in series - Output temperature and concentration 106
5.12 Triangular representation of output signals using fixed fuzzy membership functions - Output concentration (C_{A2}) 107
5.13 Triangular representation of output signals using fixed fuzzy membership functions - Output Temperature (T_2) 107
5.14 Results of the classification of normal and abnormal operating conditions based on fixed fuzzy membership functions ($NW = 5$) - BPNN method 108
5.15 Results of the classification of normal and abnormal operating conditions based on fixed fuzzy membership functions ($NW = 5$) - HMM method 108
5.16 Adaptive fuzzy membership functions for the output signals (Figure 5.11) in different operating conditions - Output concentration (C_{A2}) . 109
5.17 Adaptive fuzzy membership functions for the output signals (Figure 5.11) in different operating conditions - Output temperature (T_2) . . 109
5.18 Triangular representation of output signals in Figure 5.11 using adaptive fuzzy membership functions - Output concentration (C_{A2}) 110
5.19 Triangular representation of output signals in Figure 5.11 using adaptive fuzzy membership functions - Output temperature (T_2) 110
5.20 Results of the classification of normal and abnormal operating conditions in Figure 5.11 based on adaptive fuzzy membership functions ($N_W = 5$) 111
5.21 Normal, abnormal 1 and abnormal 2 operating conditions for the CSTRs in series - Feed flow rate 112
5.22 Normal, abnormal 1 and abnormal 2 operating conditions for the CSTRs in series - Output temperature and concentration 112

5.23 Triangular representation of normal, abnormal 1 and abnormal 2 output signals in Figure 5.22 using fixed fuzzy membership functions - Output concentration (C_{A2}) 113
5.24 Triangular representation of normal, abnormal 1 and abnormal 2 output signals in Figure 5.22 using fixed fuzzy membership functions - Output temperature (T_2) 114
5.25 Adaptive fuzzy membership functions for the output signals of the Figure 5.22 in different operating conditions - Output concentration (C_{A2}) 114
5.26 Adaptive fuzzy membership functions for the output signals of the Figure 5.22 in different operating conditions - Output temperature (T_2) 115
5.27 Triangular representation of normal, abnormal 1 and abnormal 2 output signals in Figure 5.22 using adaptive membership functions - Output concentration (C_{A2}) 115
5.28 Triangular representation of normal, abnormal 1 and abnormal 2 output signals in Figure 5.22 using adaptive membership functions - Output temperature (T_2) 116
5.29 Results of the classification of normal, abnormal 1 and abnormal 2 operating conditions in Figure 5.22 based on the HMM method ($NW = 5$) - Fixed fuzzy membership functions 116
5.30 Results of the classification of normal, abnormal 1 and abnormal 2 operating conditions in Figure 5.22 based on the HMM method ($NW = 5$) - Adaptive fuzzy membership functions 117
5.31 Normal, abnormal 1 and abnormal 2 operating conditions for the CSTRs in series - Feed flow rate 118
5.32 Normal, abnormal 1 and abnormal 2 operating conditions for the CSTRs in series - Output temperature and concentration 118
5.33 Triangular representation of normal, abnormal 1 and abnormal 2 output signals in Figure 5.32 using adaptive membership functions - Output concentration (C_{A2}) 119
5.34 Triangular representation of normal, abnormal 1 and abnormal 2 output signals in Figure 5.32 using adaptive membership functions - Output temperature (T_2) 120
5.35 Results of the classification of normal and abnormal operating conditions in Figure 5.32 based on adaptive fuzzy membership functions ($NW = 5$) 120
5.36 Normal and abnormal operations after reaching the desired set-point 121

5.37 Discritized observations for the output concentration in Figure 5.36 . 121
5.38 Discretized observations for the output temperature in Figure 5.36 . . 122
5.39 Normalized probability of the observation sequences in Figure 5.36 belonging to normal and abnormal regions using HMMs with fixed window of data ($N_W = N_{min} = 10$) 122
5.40 Normalized probability of the observation sequences in Figure 5.36 belonging to normal and abnormal regions using the BPNN approach ($N_W = N_{min} = 10$) 123
5.41 Normalized probability of the observation sequences in Figure 5.36 belonging to normal and abnormal regions using the BPNN approach ($N_W = N_{min} = 5$) 124
5.42 Normalized probability of the observation sequences in Figure 5.36 belonging to normal and abnormal regions using the proposed moving window approach ($N_W = 9, N_{min} = 5$) 124
5.43 Number of shifts in Figure 5.42 to find the optimal window 125

6.1 Three layers of the PSV unit 127
6.2 pressure loss verses average mixture velocity inside the slurry line . . 128
6.3 Solution strategy for on-line estimation of the critical minimum velocity 130
6.4 Middlings density ($\frac{g}{cm^3}$) soft sensor estimations verses lab data (scatter plot) 142
6.5 Middlings density ($\frac{g}{cm^3}$) soft sensor estimations verses lab data (time trend plot) 143
6.6 Sudden spike in the prediction of the critical minimum velocity due to the on-line analyzer unavailability 144
6.7 Modified critical velocity estimation results in the case of analyzer unavailability 144
6.8 On-line estimation of the critical velocity without adaptive sensitivity levels 145
6.9 On-line estimation of the critical velocity after applying adaptive sensitivity levels 146
6.10 A case of upset operation in 2013 data-set 146
6.11 Flow-rate operating modes for the data in Figure 6.10 147
6.12 Results of on-line testing of the algorithm in 2014 data set 148
6.13 A case of upset operating condition in the primary separation vessel from historical data 150
6.14 Validation data-set for the industrial case study 151

6.15 Operating modes of the process for the industrial case study based on the proposed method of Chapter 3 151
6.16 Operating modes of the process for the industrial case study based on conventional HMMs 152
6.17 Underflow density and the interface level of the PSV unit in an upset region 153
6.18 Interface level and underflow density - combination of normal and upset operating regions occurred in July 2011 154
6.19 Interface level and underflow density - combination of normal and upset operating regions occurred in July 2011. 155
6.20 Different states for the durations and magnitudes in the interface level signal 156
6.21 Different states for the durations and magnitudes in the underflow density signal 157
6.22 Adaptive fuzzy membership functions for the interface level signal . . 157
6.23 Adaptive fuzzy membership functions for the underflow density signal 158
6.24 Discretized observations of the interface level using appropriate fuzzy rules and membership functions 158
6.25 Discretized observations of the underflow density using appropriate fuzzy rules and membership functions 159
6.26 Overall classification of the process for different window sizes - $N_W = 15$ 159
6.27 Overall classification of the process for different window sizes - $N_W = 7$ 160

7.1 Various configurations of missing observations in the historical data . 167

List of Symbols

Symbol	Description
α	Forward variable
$\alpha(k)_{ij}$	Probability of transition from mode i to mode j at time k
β	Backward variable
μ_i	Mean vector in mode i
μ_m^i	Mean value of magnitudes in mode i
μ_d^i	Mean value of durations in mode i
$\mu_{d^i,m^i}^{sml,\ ave,\ lrg}$	Mean value of various fuzzy membership functions
μ_f	Carrier fluid viscosity
μ_l	Liquid viscosity
Σ_i	Covariance matrix in mode i
σ_i	Standard deviation in mode i
σ_{H_i}	Validity of scheduling variable in mode i
σ_m^i	Standard deviation of magnitudes in mode i
σ_d^i	Standard deviation of durations in mode i
$\sigma_{d^i,m^i}^{sml,\ ave,\ lrg}$	Standard deviations of fuzzy membership functions
θ	Unknown parameters
Γ	Gamma function
δ	Mahalanobis distance
ψ	Digamma function
π	Initial state distribution of the Markov chain model
Φ	Unknown parameters
ρ	Density
ρ_f	Carrier fluid density
ρ_l	Liquid density

ρ_s	Solid density
ν_i	Degree of freedom in mode i
$\omega_i(t)$	Weight of mode i at time t
ω_m	Settling velocity
τ_{ijk}	Posterior distribution for joint probability of modes i and j
A	Transition probability matrix
Ar	Archimedes number
a_{ij}	Probability of transition from state i to j
B	Symbol emission probability matrix
$b_j(k)$	Probability of observing symbol k at state j
C_f	Carrier fluid solid concentration
C_{Miss}	Missing data
C_{Obs}	Observed data
C_ν	Coarse particle concentration
D	Pipeline diameter
d	Coarse particle diameter
d_t	Duration of an episode at time t
F_t	Flow rate at time t
Fr	Froude number
g	Gravitational constant
H	Hessian matrix
H_k	Scheduling variable at time k
I_k	Operating mode at time k
M	Number of modes
m_t	Magnitude of an episode at time t
N	Number of observations in training data set
O	Set of discrete observations
O_t	Discrete observation at time t
P	Number of variables
Q	Q function
Q_C	Critical flow rate
q_t	State at time t

R_k	Scalar weight at time k
r_{ik}	Expected value of scalar weight in mode i at time k
S	Ratio of solid density to carrier fluid density
S_t	State at time t
U	Process inputs
V_C	Critical minimum velocity
x	Complete data set
Y	Observation data set
Y_k	Observation vector at time k
y	Observation data set
Z	Latent variables
z	Missing data set

Chapter 1

Introduction

1.1 Motivation

In general, a fault in a system is defined as "*an unpermitted deviation of at least one characteristic property of a variable from an acceptable behavior*". Accordingly, faults might end up to a complete failure of the system [1]. 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 components. First, determine if the process is operating out of its normal conditions and decide on the presence of faults, i.e., fault detection. The next is to localize the fault and find the component which is the main cause of the fault, i.e., fault isolation. The last task is to identify the fault in terms of magnitude and other detailed properties, i.e., fault estimation [2].

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. In this procedure, several difficulties might occur. One example is the complexity to find a first principle model for process due to the complicated process behavior. Another example is to find an appropriate framework to extract the critical information from high dimensional data sets. Dealing with unreliable measurements for robustness is another major issue for industrial applications. It will also be greatly advantageous if one can extract the temporal information of observations.

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 [1].

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 [3].

Quantitative approaches can be divided to 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 Squares (PLS) [3].

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

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 in Figure 1.1 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 [4].

Hidden Markov Models (HMMs) are sophisticated mathematical tools for consideration of the temporal information in both process model and history based fault diagnosis. Incorporating time domain information, HMMs will greatly assist a clas-

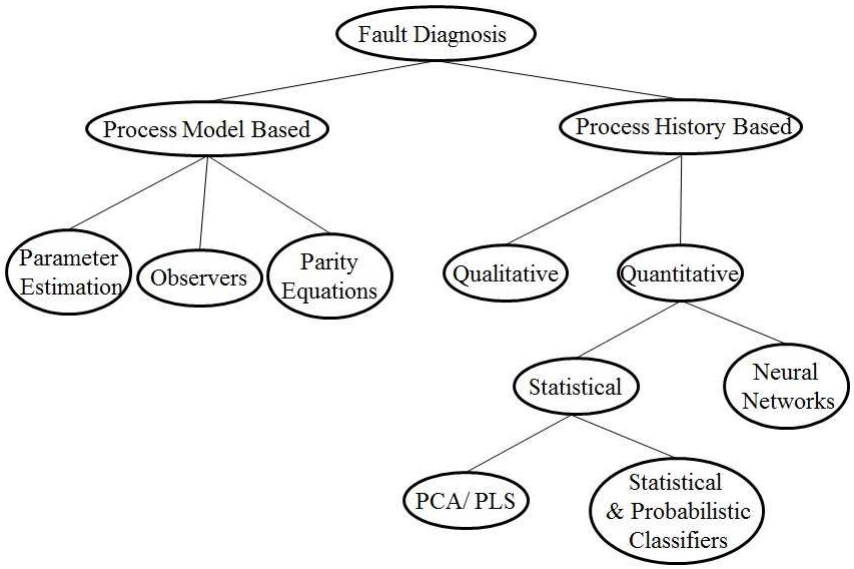

Figure 1.1: A summary of available fault diagnosis methods in literature [1, 3]

sifier 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 [4].

In this thesis, our focus is on statistical/ probabilistic classifiers (Figure 1.1). More specifically, the main effort is to more precisely consider the temporal information. Our aim is to improve the structure of transition probability matrix in HMMs, while various probability distributions are considered for observations in different operating modes. Also, in some chapters, HMMs are used to improve process history based fault diagnosis using qualitative signal representation.

Another important target of this research is to apply the developed fault detection strategies under industrial environments. The developed frameworks are tested on an important Primary Separation Vessel (PSV) of oil sands industry. PSV is an important unit in oil extraction process to separate the feed for various components. The underflow stream of the vessel usually contains more than 60 % (weight) sand, and therefore, is subject to sand deposition and plugging. The developed methods are tested both on-line and on the historical data of the unit.

The last objective of this research is to connect the two categories of process model based and history based fault diagnosis (Figure 1.1). By analyzing first principles in the underflow stream of the PSV, a model is developed to predict the minimum required velocity to avoid sand deposition. Then, HMMs are used to modify the sensitivity of the predicted velocity. More details of various thesis chapters will be

discussed next.

1.2 Previous Studies

Direct observation of process variables through on-line measurement is a well known method for monitoring of industrial processes. Such methods include limit checking of key process variables, or trend (first derivative) checking. However, these methods will detect faults only after a considerable deviation of observations from expected behavior [5]. Therefore, more advanced monitoring techniques are developed for early detection of faults.

As illustrated in Figure 1.1, available fault diagnosis techniques are divided to two main categories including process model based and process history based. In this chapter, first, each category will be reviewed. Next, previous applications of HMMs for assisting fault diagnosis methods are summarized.

1.2.1 Process Model Based Fault Diagnosis

A general schematic of model based fault diagnosis is presented in Figure 1.2. 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 1.1) [5].

$$Y = f\{U, N, \theta, X\} \tag{1.1}$$

where in Equation 1.1, U and Y are process inputs and outputs respectively, N represents nonmeasurable disturbances, θ is process parameters, and X indicates partially measurable process states.

This general category can be divided to three sub-categories [1]:

1- Fault detection with parameter estimation

2- Fault detection with observers

3- Fault detection with parity equations

Fault Detection with Parameter Estimation

The model in Equation 1.1 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

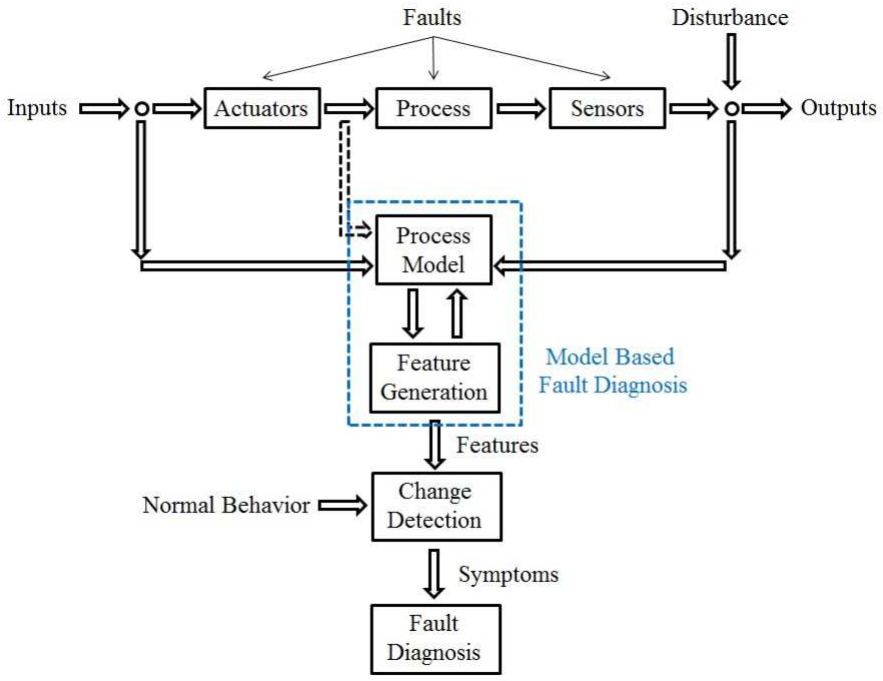

Figure 1.2: A general schematic of model based fault diagnosis [1]

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 [5].

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 [5]. For the case of multiple mode excited outputs, several methods including bank of observers excited by a single/ all modes are introduced in literature [6, 7].

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 [8].

1.2.2 Process History Based Fault Diagnosis

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 1.1, 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 Analysis (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 [3].

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 [3].

In the rest of this section, we will focus on history based quantitative methods (Figure 1.1).

Neural Networks

Artificial Neural Networks (NN) have been frequently used in chemical engineering applications [9, 10]. Many researchers have studied various aspects of NN based fault diagnosis in terms of network architecture, e.g., sigmoidal 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 [3, 11].

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 [12]. 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 observations [13].

In this context, two famous indicators have been introduced to diagnose an abnormal behavior. The first indicator, known as Hotelling's 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" [14]. An alarm will be generated when these indexes pass their standard limits.

Partial Least Squares (PLS)

PLS is similar to PCA in the sense that both methods try to deal with collinearities 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 [15]. 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 [13].

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 [3].

In cases where observations do not follow a well-defined distribution, non-parametric

methods such as kernel density estimation should be used [16]. Otherwise, parametric distributions such as a mixture of multivariate Gaussian distributions can be applied [17, 18].

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 [2].

1.2.3 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.

Model Based Approaches in Conjunction with HMMs

Fault detection based on a combination of dynamic process models, e.g., ARX models, and HMMs is a well known technique to improve model based fault detection. This approach uses deviations of the estimated parameters of the process model as observations. In contrast to classic methods where faults are assumed as independent events at each time 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 [4].

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 [19]. 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 [20].

PCA in Conjunction with HMMs

In two very recent studies, Probabilistic PCA (PPCA) is used in conjunction with HMMs for the purpose of fault diagnosis [21, 22]. PPCA has the advantage of consideration of uncertainties over regular PCA. Addition of dynamics of process transitions through the HMM framework brings a very general structure to do both fault detection and diagnosis tasks simultaneously [21]. The general structure of the proposed model is presented in Figure 1.3.

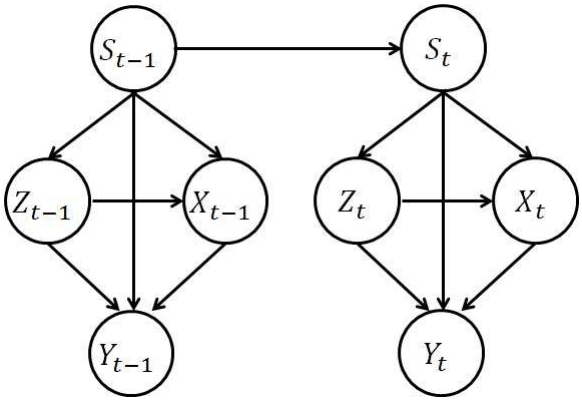

Figure 1.3: Dynamic Mixture Probabilistic Principal Component Analyzer [21]

where in this figure S represents the states (operating modes), X indicates the latent variables and Z is the mixture of Gaussians indicator. The main role of HMMs is to consider the transition from S_{t-1} to S_t through a transition probability matrix. The goal is to infer the true operating mode S_t when receiving a new observation Y_t. Expectation Maximization (EM) algorithm, which will be reviewed in the next chapter, is used for parameter estimation [21]. In another study, the same authors improved their model to handle outliers. Robust distributions, such as Student t, were used for this purpose [22].

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 [23]. 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 EM algorithm [23].

Applications of HMMs in fault diagnosis are not limited to the above studies. One can refer to many other available articles for more information [24, 25, 26].

1.3 Thesis Outline

After reviewing background on previous applications of HMMs in fault diagnosis, the remainder of this thesis is organized as follows:

In Chapter 2, 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 the EM algorithm. Applications of HMMs have been explained in previous sections. The EM algorithm is an iterative procedure for maximum likelihood estimation of parameters. In presence of missing variables, it provides a more numerically robust solution since conventional optimization methods will usually reach a local solution, specially, when the likelihood surface is not concave. Also, application of the EM algorithm has advantages in handling probabilistic constraints and guarantees of convergence. More details will be discussed in Chapter 2.

Chapter 3 provides a new process modeling and and monitoring method in presence of faulty and missing data based on HMMs. A certain structure is proposed for the Markov chain model. First, transition probabilities are considered to be time varying as a function of an underlying scheduling variable. This flexible structure assists the model to adapt to new operating conditions. Second, the proposed structure of the transition probability matrix imposes a logical transition between various operating modes, i.e., the process can only transit from normal to faulty modes after passing some intermediate modes. Thus, undesirable jumps are avoided. This consideration significantly reduces the number of parameters to be estimated from optimization. Observations, which are randomly missed in the training data set, are assumed to follow multivariate Gaussian distributions in each mode. Due to the presence of missing observations and unknown operating regimes, the maximum likelihood parameter estimation problem is solved under the EM framework. The goal of this

study is to infer the true current operating mode given all the available information. A filtering algorithm is used for this purpose. In addition to some simulation studies, the model is tested on the historical data of the PSV unit. The proposed method shows a significant improvement over available conventional techniques in literature.

Chapter 4 is a robust extension of the proposed method of Chapter 3. In the proposed model, various operating modes can transit to each other following a similar time varying Markov chain structure as in Chapter 3. However, observations in various operating modes of the HMM are considered to follow different multivariate Student t distributions which have heavier tails in comparison to the Gaussian version. The idea is to weight the covariance matrix of the Gaussian distribution in each mode according to percentage of outliers. This weight is a probabilistic function of a degree of freedom which is estimated according to the data quality in each mode. Consequently, the negative effect of outliers during parameter estimation will be downweighted and a more accurate diagnosis is achieved. Due to the existence of unknown weights and operating modes, the parameter estimation problem is solved under the EM framework. The ultimate goal is to infer the current operating mode given all the available information in an on-line application. The model is tested on simulation and real life lab experiments and shows a superior performance over other available methods in presence of lower quality data.

Chapter 5 presents a novel framework for classification of process trends based on a combination of Qualitative Trend Analysis (QTA) and HMMs. First, continuous time signals are discretized as some symbolic observations using the method of triangular representation. Due to the large difference between magnitudes and durations of the triangles in various modes, time varying fuzzy membership functions are used for the purpose of discretization. As the consequence of this fuzzification, continuous signals appear as some discrete episodes. Next, various HMMs are trained for the multiple discrete observations of normal and faulty regimes. In an on-line application, the forward-backward algorithm is used to calculate the probability of observations within a most recent moving window given each HMM. This probability is then used for the purpose of classification. In addition to a simulation example, the method is tested on the historical data of the PSV unit and shows a better performance over available techniques.

In comparison to the HMM based data classification frameworks in Chapters 3 and 4, the discritization step of Chapter 5 brings the following advantages: 1) The important qualitative and quantitative information of a signal will be captured and used for classification purposes. 2) Since this approach focuses on key features of the signals rather than details, the high frequency noise should be removed and the

method is less sensitive to noise. 3) The sequence of symbols is an appropriate input for HMMs to extract the temporal information.

In Section 7 of Chapter 5, an optimal approach to select a more informative window of observations for the purpose of trend classification is introduced. Selection of large window sizes will cause a large delay in trend classification due to the existence of old data. On the other hand, small window sizes will classify the observations without appropriate consideration of the previous memory. Therefore, an optimal framework for window size selection is required.

Chapter 6 includes the industrial case study on fault detection and monitoring of the tailings pipeline of a PSV in an oil sand industry. PSV unit is a cone shape vessel to separate the feed stream including bitumen aggregates, water, coarse sand and fines to three separate layers. The bitumen floats over a weir circling the top and makes the froth layer. Middlings layer, which contains bitumen aggregates, fines and water, is removed from the middle of the vessel. The underflow layer, which is removed from the bottom through tailings stream, contains coarse sands and water. Due to the high solid concentration and sand deposition, this stream may face plugging or "sanding".

In the chapter discussed above, a combination of semi-empirical equations and data driven methods is used to estimate the critical (optimal) flow rate to avoid sanding in the tailings stream. First, according to the physical properties of the tailings, the appropriate semi-empirical equation for estimation of the critical velocity is selected. Since one of the key variables has some measurement inaccuracies, a soft sensor is developed to improve on-line measurements. Next, HMMs are used to adaptability change the sensitivity of critical velocity estimations, i.e., to generate more sensitive predictions while operating near abnormal conditions. The algorithm has been tested on-line, and according to the available process alarm reports, shows successful predictions.

Chapter 7, which concludes the thesis, also presents the future work. Some practical improvements, to be further considered in future studies, are introduced.

1.4 Published, Submitted and Under Preparation Materials

Materials of this thesis, as summarized in the outline section, have been previously presented in following publications (in the same order as thesis chapters):

1. N. Sammaknejad, B. Huang, W. Xiong, A. Fatehi, F. Xu, and A. Espejo (2015). "Operating Condition Diagnosis Based on HMM with Adaptive Transition Probabilities in the Presence of Missing Observations". *AIChE Journal*, 61(2), 477-492. (***Chapter 3***)

2. N. Sammaknejad, B. Huang, and Y. Lu (2015). "Robust Diagnosis of Operating Mode Based on Time Varying Hidden Markov Models". *IEEE Transactions on Industrial Electronics.* DOI: 10.1109/TIE.2015.2478743. (***Chapter 4***)

3. N. Sammaknejad, B. Huang, A. Fatehi, Y. Miao, F. Xu, and A. Espejo (2014). "Adaptive Monitoring of the Process Operation Based on Symbolic Episode Representation and Hidden Markov Models with Application Toward an Oil Sand Primary Separation". *Computers and Chemical Engineering*, 71, 281-297. (***Chapter 5 - Except Section 7***)

4. N. Sammaknejad, and B. Huang (2014). "Process Monitoring Based on Symbolic Episode Representation and Hidden Markov Models - A Moving Window Approach". *Proceedings of the* 5^{th} *International Symposium on Advanced Control of Industrial Processes (ADCONIP).* Hiroshima, Japan. (***Chapter 5 - Section 7***)

5. N. Sammaknejad, B. Huang, R. S. Sanders, Y. Miao, F. Xu, A. Espejo (2015). "Adaptive Soft Sensing and On-line Estimation of the Critical Minimum Velocity with Application to an Oil Sand Primary Separation Vessel". *Proceedings of the IFAC* 9^{th} *International Symposium on Advanced Control of Chemical Processes (ADCHEM).* Whistler, Canada. (***Chapter 6 - Short Version***)

6. To be submitted as N. Sammaknejad, B. Huang, R. S. Sanders, Y. Miao, F. Xu, A. Espejo. "Adaptive Prediction of Critical Minimum Velocity of Slurry Flow with Application to an Oil Sand Primary Separation Vessel". *Journal of Process Control.* (***Chapter 6 - Complete Version***)

1.5 Main Contributions

The main contributions of this thesis can be summarized as follows:

1. Development of a time varying HMM structure for the modeling and monitoring of industrial processes.

2. Providing appropriate mathematical procedures based on the EM algorithm to deal with irregular data such as outliers and missing observations.

3. Data compression and classification through the combination of HMMs and Qualitative Trend Analysis (QTA) based on adaptive triangular representation.

4. An optimal search procedure to find the more informative observations in a recent window of data for process monitoring based on HMMs.

5. Combining first-principle knowledge and data driven methods for on-line monitoring of an industrial scale primary separation vessel through critical minimum velocity estimation.

Chapter 2

Mathematical Fundamentals

In this chapter, a review on fundamental mathematical tools of this thesis will be provided. As previously explained, the Expectation Maximization (EM) algorithm and Hidden Markov Models (HMMs) are the two major tools which will be used throughout the thesis.

2.1 Expectation Maximization (EM) Algorithm

A maximum likelihood estimation problem searches for the set of parameters, for which, observations are most likely to occur. The main contribution of the EM algorithm is to solve this problem through maximization of a lower bound of the likelihood function, also known as the Q-function, iteratively. In presence of missing variables, or other similar ill conditions, EM is a safe algorithm which guarantees the convergence to at least a local optimum [27].

Each iteration of the EM algorithm consists of two steps. First, in the Expectation (E) step, expected value of the log likelihood of complete data including the missing and observed parts given observations and parameters in the previous iteration is computed, i.e., the missing information are integrated out. Next, result of the expectation step, which is also known as the Q-function, is maximized over unknown parameters (M Step). This iterative procedure is repeated until some convergence criterion is satisfied [27].

2.1.1 Monotonicty of the EM Algorithm

In the original paper of the EM algorithm, it has been proved that the incomplete (observed) data likelihood function is non-decreasing after each EM iteration [28, 29].

To prove this, the incomplete (observed) data log likelihood function can be de-

fined as in Equation 2.1. y is the observed data set and Φ represents the set of unknown parameters.

$$L(\Phi) = log\ g(y|\Phi) \tag{2.1}$$

The complete data set x includes both the observed (y) and missing (z) information, i.e., $x = (y, z)$.

The conditional density of x given y and Φ can be defined as,

$$k(x|y, \Phi) = \frac{f(x|\Phi)}{g(y|\Phi)} \tag{2.2}$$

where $f(x|\Phi)$ represents the complete data likelihood function.

From Equation 2.2, one can simply write Equation 2.1 as,

$$L(\Phi) = log\ f(x|\Phi) - log\ k(x|y, \Phi) \tag{2.3}$$

Taking the expected value of both sides of Equation 2.3 with respect to the conditional distribution of missing variables given observations and parameters in the previous iteration, we have that:

$$E\left\{L(\Phi)\right\} = E\left\{(log\ f(x|\Phi))\,|y, \Phi^{(k)}\right\} - E\left\{(log\ k(x|y, \Phi))\,|y, \Phi^{(k)}\right\} \tag{2.4}$$

$$= Q(\Phi|\Phi^{(k)}) - H(\Phi|\Phi^{(k)})$$

where $\Phi^{(k)}$ represents the set of parameters obtained from the previous k^{th} iteration,

$$Q(\Phi|\Phi^{(k)}) = E\left\{(log\ f(x|\Phi))\,|y, \Phi^{(k)}\right\} \tag{2.5}$$

and

$$H(\Phi|\Phi^{(k)}) = E\left\{(log\ k(x|y, \Phi))\,|y, \Phi^{(k)}\right\} \tag{2.6}$$

From Equation 2.4, it can be concluded that

$$L(\Phi^{(k+1)}) - L(\Phi^{(k)}) = \left\{Q(\Phi^{(k+1)}|\Phi^{(k)}) - Q(\Phi^{(k)}|\Phi^{(k)})\right\} - \left\{H(\Phi^{(k+1)}|\Phi^{(k)}) - H(\Phi^{(k)}|\Phi^{(k)})\right\} \tag{2.7}$$

In the maximization step of the EM algorithm, the Q-function in the right hand side of Equation 2.7 is maximized such that

$$Q(\Phi^{(k+1)}|\Phi^{(k)}) \geq Q(\Phi|\Phi^{(k)}) \tag{2.8}$$

Therefore, $L(\Phi^{(k+1)}) \geq L(\Phi^{(k)})$ will hold if $H(\Phi^{(k+1)}|\Phi^{(k)}) - H(\Phi^{(k)}|\Phi^{(k)}) \leq 0$. For any arbitrary Φ,

$$H(\Phi|\Phi^{(k)}) - H(\Phi^{(k)}|\Phi^{(k)}) = E\left\{\left(log\ \frac{k(x|y,\Phi)}{k(x|y,\Phi^{(k)})}\right)|y,\Phi^{(k)}\right\} \tag{2.9}$$

$$\leq log\ E\left\{\left(\frac{k(x|y,\Phi)}{k(x|y,\Phi^{(k)})}\right)|y,\Phi^{(k)}\right\} = log\ \int_X k(x|y,\Phi)dx = 0$$

where Equation 2.9 is a consequence of Jensen's inequality [29].

Therefore, the log likelihood function $L(\Phi)$ will be increased, which results in an increasing likelihood function $g(y|\Phi)$. If this sequence is bounded, it should monotonically converge to some upper bound [29].

Generalized EM (GEM) Algorithm

In Equation 2.8, one could observe that $\Phi^{(k+1)}$ is estimated such that $Q(\Phi|\Phi^{(k)})$ is globally maximized over Φ. However, in cases where global maximization is not feasible, according to Equation 2.7, satisfying Equation 2.10 is sufficient to ensure the increasing likelihood sequence [29].

$$Q(\Phi^{(k+1)}|\Phi^{(k)}) \geq Q(\Phi^{(k)}|\Phi^{(k)}) \tag{2.10}$$

GEM algorithm based on one Newton-Raphson step is an example of a likelihood increasing sequence [30]. In this configuration the parameters are updated as follows:

$$\Phi^{(k+1)} = \Phi^{(k)} + a^{(k)}\delta^{(k)} \tag{2.11}$$

where

$$\delta^{(k)} = -\left[\frac{\partial^2 Q(\Phi|\Phi^{(k)})}{\partial\Phi\partial\Phi^T}\right]^{-1}_{\Phi=\Phi^{(k)}}\left[\frac{\partial Q(\Phi|\Phi^{(k)})}{\partial\Phi}\right]_{\Phi=\Phi^{(k)}} \tag{2.12}$$

and $0 < a^{(k)} \leq 1$.

In the case of $a^{(k)} = 1$, the sequence becomes the same as iterations of the Newton-Raphson procedure when solving Equation 2.13.

$$\frac{\partial Q(\Phi|\Phi^{(k)})}{\partial\Phi} = 0 \tag{2.13}$$

However, for the GEM sequence to be satisfied, $a^{(k)}$ should be selected such that Equation 2.10 holds. This will impose other constraints on selection of $a^{(k)}$ [30].

2.1.2 Convergence Properties of the EM Algorithm

The EM algorithm can be considered as a point to set map [29], i.e.,

$$M(\Phi^{(k)}) = arg\ max\ _{\Phi\in\Omega}\ Q(\Phi|\Phi^{(k)}) \tag{2.14}$$

where Ω is the parameter space of Φ.

When $\Phi^{(k)}$ converges to some Φ^*, if $M(\Phi)$ is continuous, Φ^* must satisfy Equation 2.15.

$$\Phi^* = M(\Phi^*) \tag{2.15}$$

where Φ^* is a fixed point of this map.

By a Taylor series expansion near Φ^* and using the property in Equation 2.15, the mapping in Equation 2.14 can be written as [31],

$$\Phi^{(k+1)} - \Phi^* = \frac{\partial M(\Phi^*)}{\partial \Phi^*}(\Phi^{(k)} - \Phi^*) \tag{2.16}$$

and thus,

$$\|\Phi^{(k+1)} - \Phi^*\| = \|\frac{\partial M(\Phi^*)}{\partial \Phi^*}\|\|(\Phi^{(k)} - \Phi^*)\| \tag{2.17}$$

with $\frac{\partial M(\Phi^*)}{\partial \Phi^*} \neq 0$.

Since the term $\|(\Phi^k - \Phi^*)\|$ in the right hand side of Equation 2.17 is to the power of one, the EM algorithm is "almost surely" a first order algorithm.

This first order convergence has been mentioned in literature as a major drawback [31]. Accordingly, some people have argued that superlinear (quasi-Newton) and second order (Newton) methods should be prefered to EM [32].

Although convergence to true parameters might be slow when using the EM algorithm, convergence of the likelihood is quite fast. This simply means that after a few number of iterations, the EM algorithm is able to estimate model parameters such that they can perfectly represent the data set [31].

In addition to this, the EM algorithm has two main advantages which makes it a popular choice for maximum likelihood estimation problems [31]:

1. The EM algorithm provides a condition to automatically satisfy probabilistic constraints of mixture models. Other optimization techniques either modify each step to keep the parameters inside the desired domain, or transform the constrained optimization to unconstrained with appropriate parameterization. Both approaches will require more computation cost or algorithm developments.

2. As previously explained in Equations 2.1 to 2.9, the EM algorithm guarantees an increasing likelihood sequence and monotonic convergence without step size parameters or line searches. The update formula for the Newton's method is presented in Equation 2.18.

$$\Phi^{(k+1)} = \Phi^{(k)} + H(\Phi^{(k)})^{-1}\frac{\partial l}{\partial \Phi^{(k)}} \tag{2.18}$$

where $H(\Phi^{(k)})$ is the Hessian matrix.
Unless the iterative process is close to a solution, it is possible that inverse of the Hessian matrix become indefinite, and iterations might diverge. Other advanced techniques such as Quasi-Newton or Levenberg-Marquardt are either not appropriate for a constrained optimization or require a form of parallel search to achieve the optimums. On the other hand, for many problems, it has been proved that the iterative procedure of the EM algorithm can be converted to the form of Equation 2.19.

$$\Phi^{(k+1)} = \Phi^{(k)} + P(\Phi^{(k)})\frac{\partial l}{\partial \Phi}|_{\Phi=\Phi^{(k)}} \tag{2.19}$$

where $P(\Phi^{(k)})$ is a positive definite matrix.
In the form of Equation 2.19, the EM algorithm can be viewed as a gradient ascent algorithm where the positive definite projection matrix $P(\Phi^{(k)})$ can change at each iteration as a function of $\Phi^{(k)}$.

As a general comment, EM is a conservative algorithm with guaranteed fast convergence of the likelihood. However, parameter convergence might be slow due to the first order behavior. For ill conditioned problems and problems with missing variables, EM plays an important role in design of training algorithms [31]. There have been some studies in literature to address the slow convergence behavior of the EM algorithm. Switching between estimated parameters from the EM and maximum of exact gradient of the likelihood function according to the proportion of missing data is one approach to deal with such an issue [33].

2.1.3 Initialization of the EM Algorithm

From the previous review on theory of the EM algorithm, it is clear that selection of the initial values is a very important step in this iterative procedure. Appropriate selection of initial values can increase the speed of convergence. More importantly, it assists the model to avoid local optimums and reach a global solution [34].

One practical approach to tackle this problem is to start from different initial values, which are randomly selected from a uniform or other distributions, and then, select the answer with the largest likelihood. This approach might require a high computational cost [34]. In multivariate mixtures, selection of initial values based on PCA mapping is a popular method to decide on the initial values [35]. Initialization based on a primary clustering is another well known technique for mixture distributions [36, 37].

Due to the wide range of available methods, one can conclude that selection of the appropriate method of initialization is subjective to the problem at hand. In this thesis, for each specific problem, we will propose the appropriate initialization method which is usually a combination of various approaches in this section.

2.1.4 Stopping Criteria for the EM Algorithm

When the value of the stopping criteria becomes smaller than a specified constant, EM iterations will be terminated. Therefore, magnitude of this constant will directly affect the sensitivity of estimated parameters. Different criteria have been introduced in literature for this purpose including relative change of estimated parameters, relative change of log likelihood, and Aitken's criterion [34].

- **Relative Change of Estimated Parameters**: This stopping criterion is based on the relative change of estimated parameters in two consecutive iterations. At each iteration, the parameter with the largest relative change is selected to create the stopping criterion, i.e., in a set of M available parameters, $\phi_{j,\ 1\leq j\leq M}$ is selected such that:

$$max\ \left(\frac{|\phi_j^{(k+1)}-\phi_j^{(k)}|}{|\phi_j^{(k)}|}\right)_{1\leq j\leq M} < \epsilon \tag{2.20}$$

  where ϵ is a specified constant [29].

- **Relative Change of Log Likelihood**: As previously stated, EM algorithm might have a slow convergence in parameters. However, convergence of the likelihood is usually fast. Therefore, although it might not indicate the actual convergence, one can define the stopping criterion based on the log likelihood function. Using the definition of $L(\Phi^{(k)})$ in Equation 2.1, this stopping criterion can be defined as [38]

$$\frac{|L(\Phi^{(k+1)})-L(\Phi^{(k)})|}{|L(\Phi^{(k)})|} < \epsilon \tag{2.21}$$

where the likelihood function can be obtained from a marginalization over all possible hidden variables.

- **Aitken Acceleration Based Stopping Criterion**: It is another stopping criterion based on relative log likelihood changes. In comparison to 2.21, some memory of the log likelihood in $(k-1)^{th}$ iteration is further considered. According to the definition in Equation 2.1, let us define

$$l^{(k)} = L(\Phi^{(k)}) \tag{2.22}$$

First, the variable $l^{(k)}$ in Equation 2.22 is mapped on some new variable as follows:

$$l_A^{(k+1)} = l^{(k)} + \frac{1}{(1-c^{(k)})}\left(l^{(k+1)} - l^{(k)}\right) \tag{2.23}$$

where $c^{(k)} = \dfrac{l^{(k+1)} - l^{(k)}}{l^{(k)} - l^{(k-1)}}$.
Next, the stopping criterion is defined based on this new variable, i.e.,

$$|l_A^{(k+1)} - l_A^{(k)}| < \epsilon \tag{2.24}$$

This algorithm has shown a satisfactory performance for problems with the main concern of log likelihood convergence [29].

2.1.5 Parameter Estimation for Mixture Densities Based on EM - An Example

This section provides an example for the application of EM algorithm to estimate the parameters of a mixture model. This problem is one of the very well known applications of the EM algorithm in literature [39].

In the mixture structure, likelihood of an observation (y) given parameters (Φ) is assumed to follow the model in Equation 2.25.

$$p(y|\Phi) = \sum_{i=1}^{M} \alpha_i p_i(y|\phi_i) \tag{2.25}$$

where $\Phi = (\alpha_1, ..., \alpha_M, \phi_1, ..., \phi_M)$ such that $\sum_{i=1}^{M} \alpha_i = 1$, and p_i is a density function with parameter ϕ_i.

Therefore, having the independence assumption between observations $(Y = \{y_i\}_{i=1}^{N})$, the incomplete data log likelihood function can be derived as

$$L(\Phi) = log\ p(Y|\Phi) = log\ \prod_{i=1}^{N} p(y_i|\Phi) = \sum_{i=1}^{N} log\left(\sum_{j=1}^{M} \alpha_j p_j(y_i|\phi_j)\right) \tag{2.26}$$

Optimization of Equation 2.26 is difficult due to the existence of "log of the sum". An alternative is to define some missing identities ($I = \{I_i\}_{i=1}^{N}$, $I_i \in \{1, ..., M\}$) which indicate the mixture component that has generated the data. Thus, the complete data log likelihood function can be expressed as

$$L_C(\Phi) = log\ p(Y, I|\Phi) = \sum_{i=1}^{N} log\left(\alpha_{I_i} p_{I_i}(y_i|\phi_{I_i})\right) \tag{2.27}$$

where Equation 2.27 is obtained based on the chain rule of probability, independence assumption of observations given model identities and independence assumption of model identities. It should be obvious that $\alpha_{I_i} = p(component\ I_i)$.

Given the structure of density function, Equation 2.27 can be optimized to estimate the unknown parameters. However, the identities (I_i), which are considered as random variables in the current format, should be known to proceed the optimization.

First, the posterior distribution of missing variables given observations and parameters in the previous iteration, i.e., the "old" parameters, is required. According to the Bayes rule:

$$p(I_i|y_i, \Phi^{old}) = \frac{\alpha_{I_i}^{old} p_{I_i}(y_i|\phi_{I_i}^{old})}{p(y_i|\Phi^{old})} = \frac{\alpha_{I_i}^{old} p_{I_i}(y_i|\phi_{I_i}^{old})}{\sum_{k=1}^{M} \alpha_{I_k}^{old} p_{I_k}(y_i|\phi_{I_k}^{old})} \tag{2.28}$$

Considering that the component identities are randomly drawn, it can be concluded that

$$p(I|Y, \Phi^{old}) = \prod_{i=1}^{N} p(I_i|y_i, \Phi^{old}) \tag{2.29}$$

Next, computing the expected value of Equation 2.27 with respect to model identities, the missing information is integrated out, and the Q-function is calculated as follows:

$$Q(\Phi|\Phi^{old}) = \sum_{I} L_C(\Phi) p(I|Y, \Phi^{old}) \tag{2.30}$$

$$= \sum_{I} \sum_{i=1}^{N} log\left(\alpha_{I_i} p_{I_i}(y_i|\phi_{I_i})\right) \prod_{j=1}^{N} p(I_j|y_j, \Phi^{old})$$

After some simplification, Equation 2.30 can be written as

$$Q(\Phi|\Phi^{old}) = \sum_{l=1}^{M} \sum_{i=1}^{N} log(\alpha_l) p(l|y_i, \Phi^{old}) + \sum_{l=1}^{M} \sum_{i=1}^{N} log\left(p_l(y_i|\phi_l)\right) p(l|y_i, \Phi^{old}) \tag{2.31}$$

Since the terms containing α_l and ϕ_l are not related, they are maximized independently.

To maximize the expression of α_l, the constraint $\sum_{l=1}^{M} \alpha_l = 1$ should be satisfied. Therefore, Lagrange multiplier λ should be introduced, i.e.,

$$\frac{\partial \left(\sum_{l=1}^{M} \sum_{i=1}^{N} log(\alpha_l) p(l|y_i, \Phi^{old}) + \lambda(\sum_{l=1}^{M} \alpha_l - 1) \right)}{\partial \alpha_l} = 0 \tag{2.32}$$

Taking the derivative with respect to both α_l and λ, and solving the set of linear equations, the following update formula is obtained:

$$\alpha_l = \frac{1}{N} \sum_{i=1}^{N} p(l|y_i, \Phi^{old}) \tag{2.33}$$

where $l \in \{1, ..., M\}$.

In order to obtain ϕ_l, various distributions can be considered for the observations in each mixture component, e.g., a multivariate Gaussian distribution ($\phi_l = \{\mu_l, \Sigma_l\}$), i.e.,

$$p_l(y|\mu_l, \Sigma_l) = \frac{1}{(2\pi)^{d/2} |\Sigma_l|^{1/2}} e^{\frac{-(y-\mu_l)^T \Sigma_l^{-1} (y-\mu_l)}{2}} \tag{2.34}$$

Taking the derivative of the second term in Equation 2.31 with respect to μ_l and Σ_l, and solving the equations, the following update formulas will be obtained:

$$\mu_l = \frac{\sum_{i=1}^{N} y_i p(l|y_i, \Phi^{old})}{\sum_{i=1}^{N} p(l|y_i, \Phi^{old})} \tag{2.35}$$

$$\Sigma_l = \frac{\sum_{i=1}^{N} p(l|y_i, \Phi^{old})(y_i - \mu_l)(y_i - \mu_l)^T}{\sum_{i=1}^{N} p(l|y_i, \Phi^{old})} \tag{2.36}$$

The estimated parameters at each iteration will be used as the initial guess for the next iteration. This iterative procedure is repeated until some convergence criterion is satisfied [39].

2.2 Hidden Markov Models

Although initial studies on HMMs were in late 1960s and early 1970s, the main applications started from late 1980s. In general, HMMs provide appropriate statistical frameworks to model real-world signals [40]. Signal models have various applications.

They can be used to enhance corrupted signals by optimally removing the noise. Also, having the appropriate signal model, one can infer the source of the signal. Such models have shown an excellent performance in recognition and identification applications. Signal models are either deterministic or stochastic. Deterministic models assume some known structures, e.g., sum of exponentials or sine wave for the signals, and try to determine the appropriate corresponding parameters. Statistical models, on the other hand, consider the signal as a parametric random process. Hidden Markov models and Gaussian processes are some examples of this category. Appropriate learning procedures are required to estimate the parameters [40].

2.2.1 An Illustrative Example

In order to explain the concept of Markov models, the weather prediction example is used here [41]. Consider to have three types of weather, e.g., sunny, rainy and fuggy. Assuming that each weather lasts for a whole day, the goal is to predict tomorrow's weather based on available historical data, i.e.,

$$p(q_t = S_j | q_{t-1} = S_i, q_{t-2} = S_k, ..., q_1 = S_1) \tag{2.37}$$

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

According to Equation 2.37, 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:

$$p(q_t = S_j | q_{t-1} = S_i, q_{t-2} = S_k, ..., q_1 = S_1) \approx p(q_t = S_j | q_{t-1} = S_i) \tag{2.38}$$

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 2.39.

$$p(q_1, q_2, q_3, ..., q_t) = \prod_{k=1}^{t} p(q_k | q_{k-1}) \tag{2.39}$$

The proposed weather prediction Markov chain structure is illustrated in Figure 2.1. a_{ij} represents the probability of transiting from state i to state j.

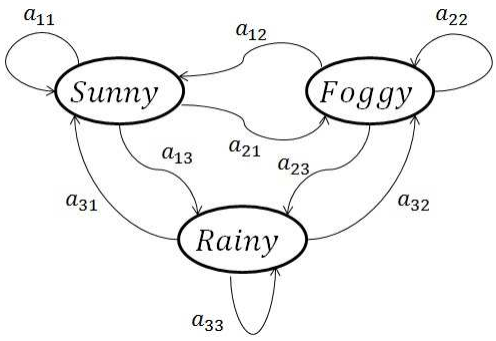

Figure 2.1: A three states Markov chain process (State 1=Sunny, State 2=Foggy and State 3=Rainy) [41]

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 previous 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 $O_1, O_2, O_3, ..., O_t$, and using the Bayes rule, Equation 2.39 can be written as

$$p(q_1, q_2, q_3, ..., q_t|O_1, O_2, O_3, ..., O_t) = \frac{p(O_1, O_2, O_3, ..., O_t|q_1, q_2, q_3, ..., q_t)\ p(q_1, q_2, q_3, ..., q_t)}{p(O_1, O_2, O_3, ..., O_t)} \tag{2.40}$$

where in this scenario $O_{l,\ l\in\{1,...,t\}}$ is either True (carrying an umbrella), or False (not carrying an umbrella). $p(q_1, q_2, q_3, ..., q_t)$ can be calculated using the same previous Markov model (Equation 2.39). $p(O_1, O_2, O_3, ..., O_t)$ 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 $j \neq i = 1, ..., t$, then $p(O_1, O_2, O_3, ..., O_t|q_1, q_2, q_3, ..., q_t) = \prod_{k=1}^{t} p(O_k|q_k)$.

2.2.2 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., $q_t \in \{S_1, ..., S_N\}$.

2. **Markov assumptions:**

(a) 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., $q_t \perp q_i|q_{t-1},\ \forall i \leq t-2$.

(b) 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 [40]:

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., $q_t \in S = \{S_1, ..., S_N\}$. 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., $O_t \in V = \{v_1, ..., v_M\}$. 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:

$$a_{ij} = p\left[q_t = S_j | q_{t-1} = S_i\right],\ 1 \leq i, j \leq N \tag{2.41}$$

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

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

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:

$$b_j(k) = p\left[O_t = \nu_k | q_t = S_j\right],\ 1 \leq j \leq N,\ 1 \leq k \leq M \tag{2.42}$$

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

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

5. **Initial state distribution ($\pi = \{\pi_i\}$):**
   The initial state distribution is specifically defined for the first sample time, i.e.,

$$\pi_i = p\left[q_1 = S_i\right],\ 1 \leq i \leq N \tag{2.43}$$

   The following constraint should hold for the initial state distribution:

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

Having the above five parameters, a hidden Markov model is defined. For convenience, one might use the compact notation in Equation 2.44 to represent HMMs.

$$\lambda = (A,\ B,\ \pi) \tag{2.44}$$

2.2.3 Three Fundamental Problems for HMMs

In literature, three fundamental problems are introduced and addressed for HMMs as follows [40]:

- ***Problem 1***:
  Having the model $\lambda = (A,\ B,\ \pi)$, how to compute the probability of an observation sequence $O = O_1, O_2, ..., O_T$ given the model λ, i.e., $p(O|\lambda)$=?

- ***Problem 2***:
  Having the model $\lambda = (A,\ B,\ \pi)$, how to find the optimal state sequence $Q = q_1, q_2, ..., q_T$ which best explains the observation sequence $O = O_1, O_2, ..., O_T$?

- ***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

A straightforward approach to address problem 1 is to marginalize $P(O|\lambda)$ over all possible state sequences [40], i.e.,

$$p(O|\lambda) = \sum_{all\ Q} p(O, Q|\lambda) = \sum_{all\ Q} p(O|Q, \lambda)p(Q|\lambda) \tag{2.45}$$

Using Markov assumptions, each term in the right hand side of Equation 2.45 can be written as

$$p(O|Q,\lambda) = \prod_{t=1}^{T} p(O_t|q_t,\lambda) = b_{q_1}(O_1).b_{q_2}(O_2). \ ... \ .b_{q_T}(O_T)$$

$$p(Q|\lambda) = \prod_{t=1}^{T} p(q_t|q_{t-1}) = \pi_{q_1} a_{q_1 q_2} a_{q_2 q_3} \ ... \ a_{q_{T-1} q_T}$$

Having the expression in Equation 2.45 and the above two equations, $p(O|\lambda)$ can be computed as

$$p(O|\lambda) = \sum_{all \ Q} p(O|Q,\lambda) p(Q|\lambda) \tag{2.46}$$

$$= \sum_{q_1, q_2, \ ..., \ q_T} \pi_{q_1} b_{q_1}(O_1) a_{q_1 q_2} b_{q_2}(O_2), \ ..., \ a_{q_{T-1} q_T} b_{q_T}(O_T)$$

It can be proved that to compute the expression in Equation 2.46, $2T.N^T$ 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 $p(O|\lambda)$ [40].

In the forward algorithm, the auxiliary probability $\alpha_t(i)$, which represents the probability of observing the partial sequence $O_1, O_2, ..., O_t$ such that state q_t is S_i, is introduced ($\alpha_t(i) = p(O_1, O_2, ..., O_t, q_t = S_i|\lambda)$) [43]. This probability is calculated inductively as follows:

- Initialization:

$$\alpha_1(i) = \pi_i b_i(O_1), \ 1 \leq i \leq N \tag{2.47}$$

- Induction:

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i) a_{ij} \right] b_j(O_{t+1}), \ 1 \leq t \leq T-1, \ 1 \leq j \leq N \tag{2.48}$$

- Termination

$$p(O|\lambda) = \sum_{i=1}^{N} \alpha_T(i) \tag{2.49}$$

The left side of Figure 2.2 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 2.46.

One can perform calculations, very similar to Equations 2.47 to 2.49, to find $p(O|\lambda)$ backward. To do this, the auxiliary backward probability $\beta_t(i)$ is defined as $\beta_t(i) = p(O_{t+1}, O_{t+2}, ..., O_T|q_t = S_i, \lambda)$ [43]. To find $p(O|\lambda)$ the following steps are required:

- Initialization:

$$\beta_T(i) = 1,\ 1 \leq i \leq N \tag{2.50}$$

- Induction:

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(O_{t+1}) \beta_{t+1}(j),\ t = T-1, T-2, ..., 1,\ 1 \leq i \leq N \tag{2.51}$$

- Termination

$$p(O|\lambda) = \sum_{i=1}^{N} \pi_i b_i(O_1) \beta_1(i) \tag{2.52}$$

The induction step of the backward algorithm is presented in the right hand side of Figure 2.2. 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 2.2 presents how these two probabilities are connected.

Solution to Problem 2

The next important problem to be addressed is to find the "optimal" path of states given an observation sequence [40]. One approach to solve this problem is to choose the states which are *individually* more likely. In such an approach, the probability of state S_i at time t given the observation sequence O and the model λ can be defined as follows:

$$\gamma_t(i) = p(q_t = S_i|O, \lambda) \tag{2.53}$$

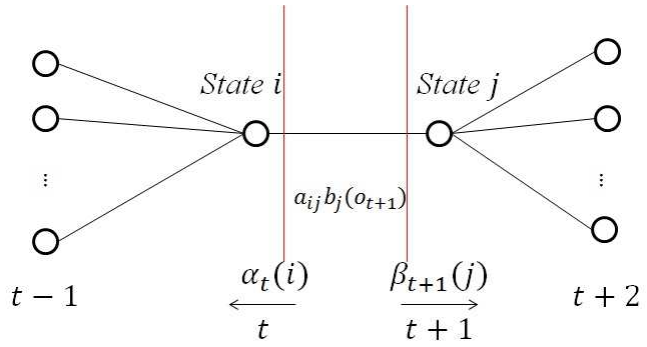

Figure 2.2: Graphical illustration of Forward-Backward algorithm [40]

Then, through an optimal search, the most likely individual state q_t at time t is selected, i.e.,

$$q_t = argmax_{1 \leq i \leq N} \left[\gamma_t(i)\right], \ 1 \leq t \leq T \tag{2.54}$$

Using the chain rule of probability, and definitions of the forward and backward probabilities, it is easy to write $\gamma_t(i)$ as follows:

$$\gamma_t(i) = \frac{\alpha_t(i)\beta_t(i)}{\sum_{i=1}^{N} \alpha_t(i)\beta_t(i)} \tag{2.55}$$

According to this definition it should be clear that $\sum_{i=1}^{N} \gamma_t(i) = 1$.

Obviously, the optimality criterion in Equation 2.54 searches for the optimal state at current sample time rather than the *optimal state sequence*. Therefore, one might consider pairs of states (q_t, q_{t+1}) or triples of states (q_t, q_{t+1}, q_{t+2}) instead the most current state. The more general case is to find the optimal state sequence $Q = \{q_1, q_2, ..., q_T\}$ for a given observation sequence $O = \{O_1, O_2, ..., O_T\}$. The problem can be defined as the best state path to maximize $p(Q|O, \lambda)$ or equivalently $p(Q, O|\lambda)$. The *Viterbi Algorithm*, which is a dynamic programming technique, addresses this problem [40].

Let us define the highest probability of a path at time t, which ends in state S_i, as follows:

$$\delta_t(i) = max_{q_1, q_2, ..., q_{t-1}} \ p\left[q_1, q_2, ..., q_t = i, O_1, O_2, ..., O_t|\lambda\right] \tag{2.56}$$

By induction, it is straightforward to show that

$$\delta_{t+1}(j) = \left[max \ \delta_t(i)a_{ij}\right].b_j(O_{t+1}) \tag{2.57}$$

The idea of the Viterbi algorithm is to keep the track of the maximized argument in Equation 2.57 for each t and j through the array $\psi_t(j)$. The complete procedure is as follows:

- Initialization:

$$\delta_1(i) = \pi_i b_i(O_1),\ 1 \leq i \leq N$$

$$\psi_1(i) = 0$$

- Recursion:

$$\delta_t(j) = max\ _{1 \leq i \leq N} \left[\delta_{t-1}(i) a_{ij}\right] b_j(O_t),\ 2 \leq t \leq T,\ 1 \leq j \leq N$$

$$\psi_t(j) = argmax\ _{1 \leq i \leq N} \left[\delta_{t-1}(i) a_{ij}\right],\ 2 \leq t \leq T,\ 1 \leq j \leq N$$

- Termination:

$$P^* = max\ _{1 \leq i \leq N} \left[\delta_T(i)\right]$$

$$q_T^* = argmax\ _{1 \leq i \leq N} \left[\delta_T(i)\right]$$

- Backtracking:

$$q_t^* = \psi_{t+1}(q_{t+1}^*),\ t = T-1, T-2, ..., 1$$

It should be noted that steps of the Viterbi algorithm are very similar to the forward algorithm. However, except the marginalization step in Equation 2.48, a maximization (recursion step) is performed.

Solution to Problem 3

The problem of finding HMM parameters to maximize the probability of observations in the training data set given the model is the most difficult problem among the the above three. Since no direct solution exists to analytically estimate the parameters, iterative procedures such as Baum-Welch algorithm have been developed for this purpose. These algorithms are equivalent to the solution of the problem under the EM framework. In this section, EM derivations to estimate HMM parameters are provided [40].

According to the definition in Equation 2.25, let us define the Q-function as follows:

$$Q(\lambda|\lambda^{old}) = \sum_{Q} log\ p(O, Q|\lambda) p(Q|O, \lambda^{old}) \tag{2.58}$$

In order to be consistent with Baum's auxiliary function, knowing that $p(O, Q|\lambda^{old}) = p(Q|O, \lambda^{old})p(O|\lambda^{old})$, Equation 2.58 can be written as [39]

$$Q(\lambda|\lambda^{old}) = \sum_{Q} log\ p(O, Q|\lambda) p(O, Q|\lambda^{old}) \tag{2.59}$$

Using the chain rule of probability, and following the expressions after Equation 2.45, the likelihood $p(O, Q|\lambda)$ can be written as

$$p(O, Q|\lambda) = \pi_{q_1} b_{q_1}(O_1) \prod_{t=2}^{T} a_{q_{t-1}q_t} b_{q_t}(O_t) \tag{2.60}$$

According to Equation 2.60, it is straightforward to show that the Q-function in Equation 2.59 becomes as follows [39]:

$$Q(\lambda|\lambda^{old}) = \sum_{Q} (log\ \pi_{q_1})\ p(O, Q|\lambda^{old}) + \sum_{Q} \left(\sum_{t=2}^{T} log\ a_{q_{t-1}q_t} \right)\ p(O, Q|\lambda^{old}) \tag{2.61}$$

$$+ \sum_{Q} \left(\sum_{t=1}^{T} log\ b_{q_t}(O_t) \right)\ p(O, Q|\lambda^{old})$$

The parameters which we wish to optimize in Equation 2.61 are in three separate terms of the sum. Therefore, each term can be optimizaed independently.

The first term in Equation 2.61 can be simplified as follows:

$$\sum_{Q} (log\ \pi_{q_1})\ p(O, Q|\lambda^{old}) = \sum_{i=1}^{N} log\ (\pi_i)\ p(O, q_1 = i|\lambda^{old}) \tag{2.62}$$

For the constraint $\sum_{j=1}^{N} \pi_i = 1$ to hold, a Lagrange multiplier should be introduced, and π_i is obtained as follows [39]:

$$\pi_i = \frac{p(O, q_1 = i|\lambda^{old})}{p(O|\lambda^{old})} \tag{2.63}$$

Similarly, the second term in Equation 2.61 can be simplified as

$$\sum_{Q} \left(\sum_{t=2}^{T} log\ a_{q_{t-1}q_t} \right)\ p(O, Q|\lambda^{old}) = \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{t=2}^{T} log\ (a_{ij})\ p(O, q_{t-1} = i, q_t = j|\lambda^{old}) \tag{2.64}$$

Again, the Lagrange multiplier should be introduced to satisfy the constraint $\sum_{j=1}^{N} a_{ij} = 1$, and the final result is as follows:

$$a_{ij} = \frac{\sum_{t=2}^{T} p(O, q_{t-1} = i, q_t = j|\lambda^{old})}{\sum_{t=2}^{T} p(O, q_{t-1} = i|\lambda^{old})} \tag{2.65}$$

Finally, the last term of Equation 2.61, can be written as

$$\sum_{Q}\left(\sum_{t=1}^{T} log\ b_{q_t}(O_t)\right)\ p(O,Q|\lambda^{old}) = \sum_{j=1}^{N}\sum_{t=1}^{T} log\ b_j(O_t)\ p(O,q_t=j|\lambda)) \tag{2.66}$$

For this problem, the Lagrange multiplier should be introduced to satisfy $\sum_{l=1}^{M} b_i(l) = 1$. Only the observations which are equal to ν_k will contribute to the k^{th} probability value. Therefore,

$$b_j(k) = \frac{\sum_{t=1}^{T} p(O,q_t=j|\lambda)\delta(O_t,k)}{\sum_{t=1}^{T} p(O,q_t=j|\lambda)} \tag{2.67}$$

where $\delta(O_t,k) = 1\ if\ O_t = k$, and 0 *otherwise*.

In order to write the update Equations 2.63, 2.65 and 2.67 with the same notations as the Baum Welch algorithm, let us define the probability of being in state S_i at time t, and S_j at time $t+1$, given the model and observations as follows [39]:

$$\xi_t(i,j) = p(q_t = S_i, q_{t+1} = S_j|O,\lambda^{old}) \tag{2.68}$$

Using the Bayes and chain rules, the expression in Equation 2.68 can be written as

$$\xi_t(i,j) = \frac{p(q_t = S_i, q_{t+1} = S_j, O|\lambda^{old})}{p(O|\lambda^{old})} = \frac{\alpha_t(i)a_{ij}b_j(O_{t+1})\beta_{t+1}(j)}{\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_t(i)a_{ij}b_j(O_{t+1})\beta_{t+1}(j)} \tag{2.69}$$

where all the terms in Equation 2.69 have been defined previously.

From the definition of $\gamma_t(i)$ in Equation 2.53 and $\xi_t(i,j)$ in Equation 2.69, the update formulas can be written as follows [39]:

$$\pi_i = \gamma_1(i) \tag{2.70}$$

$$a_{ij} = \frac{\sum_{t=1}^{T-1}\xi_t(i,j)}{\sum_{t=1}^{T-1}\gamma_t(i)} \tag{2.71}$$

$$b_j(k) = \frac{\sum_{t=1}^{T}\gamma_t(j)\delta(O_t,k)}{\sum_{t=1}^{T}\gamma_t(j)} \tag{2.72}$$

The inference procedure in Equations 2.53 and 2.69, which uses the complete observation sequence $O_{1:T}$ to infer the distribution of each hidden state q_t through the forward-backward algorithm, is usually called *smoothing*. One might only use observations up to the current sample time to do such inference. This second problem

is known as a *filtering* problem in literature. A particular form of filtering for HMM applications, which is known as Hamilton's filtering algorithm, will be introduced and used in next chapters [44].

There are many other interesting problems which have been previously addressed in HMM literature. Continuous observation densities, autoregressive HMMs, state duration densities, and implementation issues such as scaling are some of the examples [40]. Some of these problems will be addressed in next chapters.

Chapter 3

Operating Condition Diagnosis Based on HMM with Adaptive Transition Probabilities in Presence of Missing Observations

In this chapter a new approach for modeling and monitoring of multivariate processes in presence of faulty and missing observations is introduced. It is assumed that operating modes of the process can transit to each other following a Markov chain model. Transition probabilities of the Markov chain are time varying as a function of a scheduling variable. Therefore, the transition probabilities will be able to vary adaptively according to different operating modes. In order to handle the problem of missing observations and unknown operating regimes, the Expectation Maximization (EM) algorithm is used to estimate the parameters. The proposed method is tested on two simulations and one industrial case studies. The industrial case study is the abnormal operating condition diagnosis in the primary separation vessel of oil sands processes. In comparison to the conventional methods, the proposed method shows superior performance in detection of different operating conditions of the process.

3.1 Introduction

Regime switching systems have been of a great interest in the field of economics since 1990. It starts with Hamilton's study in 1988 where a non-linear filter was developed to learn about the changes in regime and find the maximum likelihood estimation of the parameters [46]. Later, he developed an EM algorithm framework to find the

A version of this chapter has been published in N. Sammaknejad, B. Huang, W. Xiong, A. Fatehi, F. Xu and A. Espejo (2015). Operating Condition Diagnosis Based on HMM with Adaptive Transition Probabilities in Presence of Missing Observations. AIChE Journal 61(2), 477–491 [45].

maximum likelihood estimation of the process parameters with discrete shifts [44]. The shifts were modeled as the outcome of a Markov process. The goal of this study was to find the main changes of the asset prices from observable events.

Since then, applications of switching Markov regimes have been widely used in the field of economics. Bollen et al. studied applications of the regime-switching models to analyze the dynamic behavior of foreign exchange rates. They found that prices do not obey a Markov regime switching behavior [47]. Ang et al. performed a similar study in interest rates of United States, Germany and the United Kingdom and concluded that regime switching models have better forecasts in comparison to single-regime models [48]. Pelletier introduced a regime switching dynamic correlation model for the variance between different time series. It is shown in their empirical case study that the developed model has a better performance in comparison to the previous studies [49]. In the same year, Mount et al. showed that a stochastic regime switching system is able to model the behavior of wholesale electricity prices and the price spikes [50].

The idea of considering time-varying transition probabilities in regime switching systems was initially proposed by Diebold et al. [51]. They developed an EM algorithm framework for parameter estimation in cases where the transition probabilities are function of underlying economic fundamentals. However, there are several limitations in their study. First, it is assumed that the scheduling variable (economic fundamentals) can only accept some limited discrete values. In other words, the transition behavior of the scheduling variable between different operating modes, which is an important factor in industrial processes, is not considered. Second, they solve the problem considering only two possible hidden modes for the process. Finally, their proposed optimization procedure for the non-linear terms is very dependent to the initial values. Filardo et al. made a clear picture of the advantages in considering time varying transition probabilities over fixed transition probabilities afterwards [52]. Later, Otranto considered a specific multi-chain Markov switching model where the transition probabilities are dependent to the regime of other variables. This approach was successful in predicting the regime of analyzed variables [53].

Although hidden Markov models and regime switching systems have been widely studied in the field of economics, their applications in the field of system identification and fault detection are sparse. Discrete time Markov jump linear systems and sudden failures have been previously reviewed in literature [54]. Jin et al. developed a solution strategy for identification of switched Markov autoregressive exogenous systems under the EM framework [55]. Ghasemi et al. proposed a parameter estimation method for a condition monitoring equipment with a certain failure rate structure.

Hidden states (modes) were assumed to transit following a hidden Markov model and observations were assumed to be imprecise. They used a maximum likelihood estimation framework to obtain the parameters [25]. Wong et al. proposed to use hidden Markov models as a generalization to the mixture of Gaussian approach in order to model the faults due to sensor malfunctions. They showed that these types of faults can be detected more appropriately using a HMM-based model [26]. Jiang et al. proposed a new method for fault detection of gear transition system. They modeled the system behavior as a three state continuous time Markov process. Parameter estimation was based on the EM algorithm framework [23]. The detailed proof of their mathematical derivations is available in their recent article [56].

There has also been a great effort in handling the problem of missing data in recent years. Deng et al. studied identification of non-linear parameter varying systems with missing output data using particle filter. The model is appropriate for the processes which work in multiple operating conditions [57]. Different approaches to handle the problem of missing data using the EM algorithm is discussed in detail in literature [29]. Multivariate process monitoring methods have also been broadly reviewed in literature. Keshavarz et al. compared the application of Bayesian and EM methods in multivariate change point detection [58].

In this chapter, we propose a new modeling and monitoring strategy for multivariate processes which follow a Markov regime switching behavior with time varying transition probabilities in the presence of missing and faulty observations. Since the scheduling variable is usually a good indication to the current operating mode of the process, transition probabilities are considered to be time-varying as a function of the scheduling variable. Therefore, transitions of the process between different operating modes are taken into account by defining the transition probabilities as distributions which are function of the underlying scheduling variable. In comparison to conventional HMMs, this structure shows a far better performance for the processes which have an asymmetric time-varying transition behavior between different operating modes, i.e., when some of the modes are far from the majority and the scheduling variable provides more flexibility in the modeling and filtering steps. Furthermore, a certain structure is considered for the operating modes in the transition probability matrix as the operating modes can transit to each other only in a logical manner, e.g., normal, abnormal and then the faulty mode. This structure will reduce the computational cost and provide an appropriate framework for the industrial processes with continuous transitions from the normal to faulty modes.

Since industrial data are usually subject to missing observations and unknown operating regimes, the problem is solved under the EM algorithm framework. In the

maximization step (M-step) of the EM algorithm, the non-linear interior point local optimization algorithm is adopted to find the optimal value of some of the unknown parameters numerically. Since local numerical non-linear optimization methods are usually sensitive to the initial guess, the initial values of the local optimization problem for the first iteration of the EM algorithm are obtained from an optimization based on the first few generations of the Genetic Algorithm (GA). Other initial values for the unknown parameters in the EM algorithm are obtained by assuming that the observations follow a mixture of multivariate Gaussian distributions.

After the parameter estimation step, Hamilton's filter is applied to infer the hidden operating mode of the process for the test data set [44]. The accuracy of the algorithm is tested on both simulation and industrial case studies and compared to conventional HMMs. The industrial case study is the abnormal operating condition diagnosis in the primary separation vessel which is an important early separation step in oil sands processes. The method shows satisfactory predictions in recognition of different operating conditions of the process.

The remainder of this chapter is organized as follows: Section 3.2 is the problem statement where the model structure and unknown parameters are introduced. Section 3.3 reviews the steps of the EM algorithm for parameter estimation. Section 3.4 is the application of Hamilton's filter to infer the hidden operating mode of the process. Section 3.5 includes the results of the simulation and industrial case studies and Section 3.6 concludes this chapter.

3.2 Problem Statement

The data set for the process variables are considered as follows:

$$Y = \begin{pmatrix} y_{11} & y_{12} & \cdots & y_{1N} \\ y_{21} & y_{22} & \cdots & y_{2N} \\ \vdots & \vdots & \vdots & \vdots \\ y_{P1} & y_{P2} & \cdots & y_{PN} \end{pmatrix} = (Y_1, Y_2, ..., Y_N)$$

where P is the number of process variables and N is the number of sample times.

The observable data (C_{obs}) include $Y_O = \{Y_{t_1}, ..., Y_{t_\alpha}\}$, and $H = \{H_1, ..., H_N\}$, where H is the scheduling variable. The missing data (C_{miss}) include $Y_M = \{Y_{m_1}, ..., Y_{m_\beta}\}$, and the hidden operating modes of the process at different sample times, i.e., $I = \{I_1, ..., I_N\}$. This hidden operating mode corresponds to the operating condition of the process, e.g., normal, abnormal and fault. Unlike abnormal modes, when the process reaches a faulty mode, the likelihood of returning to a normal operation significantly reduces.

One could easily compare this terminology with the corresponding terms in a state-space model. The introduced operating mode (operating condition) in this chapter, corresponds to the states of a state-space model. Similar to a state-space model, Observations (Y), which correspond to the process outputs, are functions of the underlying states (Equation 3.1). However, the data set is only partially available (Y_O). The transition probability matrix (Equation 3.5) corresponds to the state, or system matrix (A matrix) in the state-space representation.

It is obvious that the union of Y_O and Y_M is Y. The missing data (Y_M) might come from different sources such as computer disconnections, sensor failures and data collection errors [59]. Three main types of missing data have been introduced in literature [60]: 1) Missing not at random (MNAR) 2) Missing at random (MAR) 3) Missing completely at random (MCAR). In the MNAR case, the probability of missingness depends on the missing data. In the MAR case, the probability of missingness does not depend on the missing data, but depends on the observed data. In the MCAR case, the probability of missingness is independent of both the missing and observed data. The missing data in the simulation case studies of this chapter are assumed to be completely missed at random.

Observations at the k^{th} time step are assumed to follow a multivariate normal distribution with mean vector μ_i and covariance matrix Σ_i given the operating mode i, i.e.,

$$f(Y_k|I_k = i) \sim N_P(\mu_i, \Sigma_i) \tag{3.1}$$

where

$$N_P(\mu_i, \Sigma_i) = (2\pi)^{-P/2}|\Sigma_i|^{-1/2}exp(-\frac{1}{2}(Y_k - \mu_i)^T\Sigma_i^{-1}(Y_k - \mu_i))$$

where $1 \leq k \leq N$, $1 \leq i \leq M$, and M is the number of available hidden operating modes in the process.

π_i is the initial state (operating mode) distribution of the Markov chain ($\pi_i = P(I_1 = i)$) and the time varying transition probabilities are defined as a function of the scheduling variable at the previous time step, i.e.,

$$\alpha(k)_{ij} = P(I_k = j \mid I_{k-1} = i, H_{k-1}) \tag{3.2}$$

When $i = j$, $\alpha(k)_{ii}$ follows the distribution in Equation (3.3). Otherwise, it follows the distribution in Equation (3.4).

$$\alpha(k)_{ii} = \frac{2\gamma_{ii}exp(\frac{-(H_{k-1} - H_i)^2}{2\sigma_{H_i}^2})}{1 + exp(\frac{-(H_{k-1} - H_i)^2}{2\sigma_{H_i}^2})} \tag{3.3}$$

$$\alpha(k)_{ij,i\neq j} = \gamma_{ij}(1 - \alpha(k)_{ii}) \tag{3.4}$$

where $1 \leq k \leq N$, $1 \leq i, j \leq M$, σ_{H_i} is an indicator for the validity of the scheduling variable in operating mode i, H_i is the mean value of the scheduling variable in operating mode i, and γ_{ij}'s provide more flexibility in estimation of the distribution for the transition probability $\alpha(k)_{ij}$.

Having such distributions in Equations 3.3 and 3.4, continuous transitions of the scheduling variable, and consequently the process, between different operating modes are taken into account. One should note that in the case when $H_{k-1} = H_i$, $\alpha(k)_{ij}$ turns to a constant value and the transition probability matrix behaves as a conventional Markov chain model. When H_{k-1} starts to deviate from H_i, the distribution in Equation 3.3 decreases, while the distributions in Equation 3.4 start to grow, i.e., the probability to stay is the same mode decreases and, as a result, the probability of transition to other modes increases. This framework provides an appropriate structure for modeling and monitoring of the processes with time-varying asymmetric transitions between different operating modes, i.e., when some of the operating modes are far from the majority and the scheduling variable can provide more flexibility in the modeling and filtering steps. Industrial examples of such cases are provided in the case studies. Another advantage of considering such distributions is the appearance of the linear constraints on $\gamma_{ij,i\neq j}$'s in Equations 3.6, 3.7 and 3.8. Having such linear constraints, makes the optimization problem to find $\gamma_{ij,i\neq j}$ analytically tractable using Lagrange multipliers (Equations 3.27, 3.28).

Moreover, the $M \times M$ transition probability matrix is assumed to follow the structure in Equation (3.5).

$$\begin{bmatrix}
\alpha(k)_{11} & \cdots & \alpha(k)_{1Q} & 0 & \cdots & 0 \\
\vdots & \vdots & \vdots & \vdots & \cdots & \vdots \\
\alpha(k)_{P1} & \cdots & \alpha(k)_{PQ} & 0 & \cdots & 0 \\
\alpha(k)_{(P+1)1} & \cdots & \cdots & \cdots & \cdots & \alpha(k)_{(P+1)M} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
\alpha(k)_{Q1} & \cdots & \cdots & \cdots & \cdots & \alpha(k)_{QM} \\
0 & \cdots & 0 & \alpha(k)_{(Q+1)(Q+1)} & \cdots & \alpha(k)_{(Q+1)(M)} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
0 & \cdots & 0 & \alpha(k)_{(M)(Q+1)} & \cdots & \alpha(k)_{(M)(M)}
\end{bmatrix} \tag{3.5}$$

This structure imposes a logical transition between different operating modes of the process, i.e., the process cannot enter the faulty modes right after leaving the normal modes, and the abnormal modes are some intermediate modes which have the capability to transit to both normal and faulty modes. Also, from the faulty

modes the process can just transit to the abnormal modes. This transition behavior is graphically presented in Figure 3.1. This structure makes the model more appropriate for a wide class of process industry applications where the process continuously changes between different operating modes rather than sudden discrete jumps from normal to faulty. Furthermore, it reduces the number of required parameters and the computational time in both parameter estimation and on-line operating mode recognition (filtering) steps.

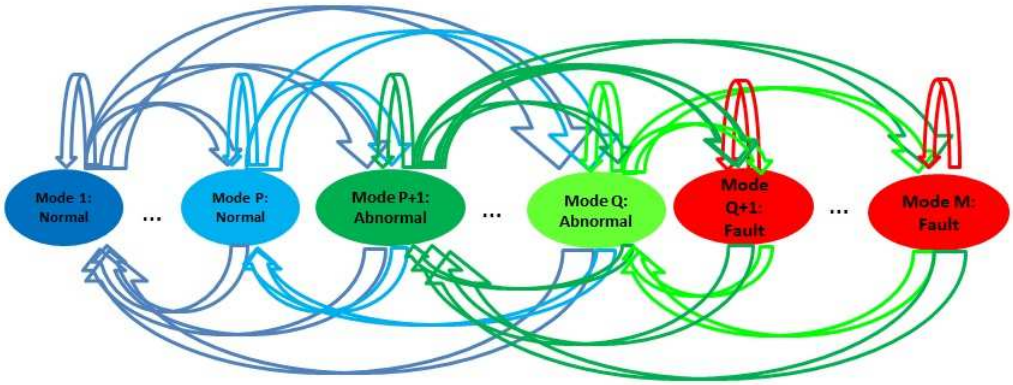

Figure 3.1: The diagram of operating mode transitions used for fault detection purpose of this chapter (normal modes: 1 to P, abnormal modes: $P+1$ to Q, faulty modes: $Q+1$ to M)

According to the structure introduced in Equation (3.5), there are three types of constraints on the parameters of the different operating modes.

Constraints for the normal operating modes:

$$if\ 1 \leq i \leq P,\ 1 \leq j \leq Q:\ 0 \leq \gamma_{ij} \leq 1,\ \sum_{j=1, j\neq i}^{j=Q} \gamma_{ij} = 1$$

$$if\ 1 \leq i \leq P,\ Q+1 \leq j \leq M:\ \gamma_{ij} = \alpha_{ij} = 0 \tag{3.6}$$

Constraints for the abnormal operating modes:

$$if\ P+1 \leq i \leq Q,\ 1 \leq j \leq M:\ 0 \leq \gamma_{ij} \leq 1,\ \sum_{j=1, j\neq i}^{j=M} \gamma_{ij} = 1 \tag{3.7}$$

Constraints for the faulty operating modes:

$$if\ Q+1 \leq i \leq M,\ 1 \leq j \leq Q:\ \gamma_{ij} = \alpha_{ij} = 0$$

$$if\ Q+1 \leq i \leq M,\ Q+1 \leq j \leq M:\ 0 \leq \gamma_{ij} \leq 1,\ \sum_{j=Q+1, j\neq i}^{j=M} \gamma_{ij} = 1 \tag{3.8}$$

These constraints are used later in the derivations of the EM algorithm.

In summary, the unknown parameters to be estimated from the EM algorithm are the mean vectors and covariance matrices of the different modes (μ_i and Σ_i), parameters of the transition probabilities (γ_{ii}'s and γ_{ij}'s) and the validity of the scheduling variable in different operating modes (σ_{H_i}). The mean value of the scheduling variable at different operating modes (H_i's) are assumed to be known from the historical data (find more information in Section 3.3.2).

3.3 Parameter Estimation Based on the Expectation Maximization Algorithm

EM algorithm finds the maximum likelihood estimation of the unknown parameters by iteratively switching between the expectation (E) and maximization (M) steps [28].

In the E-step of the EM algorithm, the Q-function, which is the conditional expectation of the complete data, is calculated:

$$Q(\theta \mid \theta^{old}) = E_{C_{miss}|(\theta^{old}, C_{obs})}\{log\ f(C_{obs}, C_{miss} \mid \theta)\} \tag{3.9}$$

where θ^{old} is the vector of parameters for the previous iteration, C_{miss} is the missing data-set and C_{obs} is the observed data-set.

In the M-step, the set of parameters that maximizes the Q-function are calculated:

$$\theta^{new} = argmax_{\theta} Q(\theta \mid \theta^{old}) \tag{3.10}$$

This procedure is iteratively repeated until some stopping criterion is satisfied.

3.3.1 Expectation Step

In the expectation step, the expected value of the complete-data log-likelihood function is calculated. Observed and missing data and the unknown parameters have been previously introduced in the problem statement section.

$$Q(\theta \mid \theta^{old}) = E_{C_{miss}|(\theta^{old}, C_{obs})}\{log\ f(C_{obs}, C_{miss} \mid \theta)\} \tag{3.11}$$

$$= E_{I_{1:N}, Y_M|(\theta^{old}, C_{obs})}\{log\ f(Y_{1:N}, H_{1:N}, I_{1:N} \mid \theta)\}$$

Using the chain rule, the probability density function in Equation (3.11) can be decomposed as follows:

$$f(Y_{1:N}, H_{1:N}, I_{1:N} \mid \theta) \tag{3.12}$$

$$= f(Y_{1:N} \mid H_{1:N}, I_{1:N}, \theta) P(I_{1:N} \mid H_{1:N}, \theta) P(H_{1:N} \mid \theta)$$

Each term of Equation (3.12) is explained in Equations (3.13) to (3.15) respectively.

$$f(Y_{1:N} \mid H_{1:N}, I_{1:N}, \theta) = \prod_{k=1}^{N} f(Y_k \mid Y_{k-1}, ..., Y_1, H_{1:N}, I_{1:N}, \theta) \tag{3.13}$$

$$= \prod_{k=1}^{N} f(Y_k | I_k, \theta)$$

In Equation (3.13), we have used the fact that given the model identity I, the conditional distribution of Y is independent of the scheduling variable H. Furthermore, Y_k follows the multivariate normal distribution in Equation (3.1) given the hidden operating mode at time k. Also,

$$P(I_{1:N} \mid H_{1:N}, \theta) = \prod_{k=1}^{N} P(I_k \mid I_{k-1}, ..., I_1, H_{1:N}, \theta) \tag{3.14}$$

$$= P(I_1) \prod_{k=2}^{N} P(I_k \mid I_{k-1}, H_{k-1}, \theta)$$

Equation 3.14 is derived based on the Markov property of the model.

H_i is independent of θ, and therefore, the last term in Equation (3.12) can be considered as a constant in the Q function, i.e.,

$$P(H_{1:N} \mid \theta) = Const \tag{3.15}$$

Following Equations (3.12) to (3.15), the Q-function in Equation (3.11) can be written as

$$Q(\theta \mid \theta^{old}) = E_{I_{1:N}, Y_M | (\theta^{old}, C_{obs})} \{ \sum_{k=1}^{N} log f(Y_k | I_k, \theta)) \tag{3.16}$$

$$+ \sum_{k=2}^{N} log P(I_k \mid I_{k-1}, H_{k-1}, \theta)) + log P(I_1) + log\ (Const) \}$$

In the first step, the expected value in Equation (3.16) is calculated with respect to the hidden operating modes (I_k), i.e.,

$$Q(\theta \mid \theta^{old}) = \tag{3.17}$$

$$E_{Y_M|(\theta^{old},C_{obs},I)} \sum_{I_1} ... \sum_{I_N} P(I_1, ..., I_N \mid \theta^{old}, C_{obs})$$

$$\{\sum_{k=1}^{N} logf(Y_k|I_k, \theta) + \sum_{k=2}^{N} logP(I_k \mid I_{k-1}, H_{k-1}, \theta) + logP(I_1) + log\ (Const)\}$$

Equation (3.17) can be further simplified as

$$Q(\theta \mid \theta^{old}) = \tag{3.18}$$

$$E_{Y_M|(\theta^{old},C_{obs},I)}\{\sum_{i=1}^{M}\sum_{k=1}^{N} P(I_k = i \mid \theta^{old}, C_{obs}) logf(Y_k \mid I_k = i, \theta)$$

$$+ \sum_{i=1}^{M}\sum_{j=1}^{M}\sum_{k=1}^{N} P(I_k = j, I_{k-1} = i \mid \theta^{old}, C_{obs}) log\alpha_{ij}(k)$$

$$+ \sum_{i=1}^{M} P(I_1 = i \mid \theta^{old}, C_{obs}) log\pi_i + log\ Const\}$$

In the next step, expectation is calculated with respect to missing observations (Y_M), i.e,

$$Q(\theta \mid \theta^{old}) = \tag{3.19}$$

$$\int_{Y_m} \sum_{i=1}^{M}\sum_{k=1}^{N} P(I_k = i \mid \theta^{old}, C_{obs}) logf(Y_k \mid I_k = i, \theta) \times P(Y_{m_1:m_\beta} \mid \theta^{old}, C_{obs}, I) dY_{m_1:m_\beta}$$

$$+ \int_{Y_m} \sum_{i=1}^{M}\sum_{j=1}^{M}\sum_{k=1}^{N} P(I_k = j, I_{k-1} = i \mid \theta^{old}, C_{obs}) log\alpha_{ij}(k) \times P(Y_{m_1:m_\beta} \mid \theta^{old}, C_{obs}, I) dY_{m_1:m_\beta}$$

$$+ \int_{Y_m} \sum_{i=1}^{M} P(I_1 = i \mid \theta^{old}, C_{obs}) log\pi_i \times P(Y_{m_1:m_\beta} \mid \theta^{old}, C_{obs}, I) dY_{m_1:m_\beta} + log\ Const$$

Since the integration is with respect to the missing observations, Equation (3.19) can be simplified as,

$$Q(\theta \mid \theta^{old}) = \tag{3.20}$$

$$\sum_{i=1}^{M} \sum_{k=t_1}^{t_\alpha} logf(Y_k \mid I_k = i, \theta) P(I_k = i \mid \theta^{old}, C_{obs})$$

$$+ \sum_{i=1}^{M} \sum_{k=m_1}^{m_\beta} P(I_k = i \mid \theta^{old}, C_{obs}) \times \int P(Y_k \mid \theta^{old}, C_{obs}, I_k = i) logf(Y_k \mid I_k = i, \theta) dY_k$$

$$+ \sum_{i=1}^{M} \sum_{j=1}^{M} \sum_{k=1}^{N} P(I_k = j, I_{k-1} = i \mid \theta^{old}, C_{obs}) log\alpha_{ij}(k) + \sum_{i=1}^{M} P(I_1 = i \mid \theta^{old}, C_{obs}) log\pi_i$$

$$+log\ Const$$

Assuming that Y_k follows the multivariate Gaussian distribution in Equation (3.1) given the hidden operating mode at time k, and using the properties of the expected value of the quadratic form, the integral term in Equation (3.20) can be derived as

$$\int P(Y_k \mid \theta^{old}, C_{obs}, I_k = i) logf(Y_k \mid I_k = i, \mu_i, \Sigma_i) dY_k \tag{3.21}$$

$$= -\frac{1}{2} log((2\pi)^P |\Sigma_i|) - \frac{1}{2} (tr(\Sigma_i^{-1} \Sigma_i^{old}) + (\mu_i^{old} - \mu_i)^T \Sigma_i^{-1} (\mu_i^{old} - \mu_i))$$

Details of the derivations in Equations 3.21, 3.23 and 3.24 are available in Appendix A.

Finally, the Q-function is written as

$$Q(\theta \mid \theta^{old}) = \tag{3.22}$$

$$\sum_{i=1}^{M} \sum_{t_1}^{t_\alpha} logf(Y_k \mid I_k = i, \theta) P(I_k = i \mid \theta^{old}, C_{obs})$$

$$+ \sum_{i=1}^{M} \sum_{k=m_1}^{m_\beta} P(I_k = i \mid \theta^{old}, C_{obs})$$

$$\times(-\frac{1}{2}log((2\pi)^P|\Sigma_i|) - \frac{1}{2}(tr(\Sigma_i^{-1}\Sigma_i^{old}) + (\mu_i^{old} - \mu_i)^T\Sigma_i^{-1}(\mu_i^{old} - \mu_i)))$$

$$+\sum_{i=1}^{M}\sum_{j=1}^{M}\sum_{k=1}^{N} P(I_k = j, I_{k-1} = i \mid \theta^{old}, C_{obs})log\alpha_{ij}(k)$$

$$+\sum_{i=1}^{M} P(I_1 = i \mid \theta^{old}, C_{obs})log\pi_i + log\ Const$$

3.3.2 Maximization Step

In the maximization step, derivatives of the Q-function are taken with respect to the unknown parameters and then set to zero. For the parameters without an analytical solution, the optimal value of the parameters are found following a numerical optimization procedure.

In order to find the optimal mean vector of the different operating modes, derivatives of the first two terms in Equation (3.22) are taken with respect to μ_i and then set to zero. Using the derivative properties of the vectors, the final result for the mean vector of each mode is obtained as in Equation (3.23).

$$\mu_i^{new} = \frac{\sum_{k=t_1}^{t_\alpha} Y_k P(I_k = i \mid \theta^{old}, C_{obs}) + \sum_{k=m_1}^{m_\beta} \mu_i^{old} P(I_k = i \mid \theta^{old}, C_{obs})}{\sum_{k=1}^{N} P(I_k = i \mid \theta^{old}, C_{obs})} \tag{3.23}$$

Derivatives with respect to the covariance matrix of each mode are taken and set to zero in a similar manner. The final result is presented in Equation (3.24).

$$(\Sigma_i)^{new} = \frac{\sum_{k=t_1}^{t_\alpha}(Y_k - \mu_i^{new})(Y_k - \mu_i^{new})^T P(I_k = i \mid \theta^{old}, C_{obs})}{\sum_{k=1}^{N} P(I_k = i \mid \theta^{old}, C_{obs})} \tag{3.24}$$

$$+\frac{\sum_{k=m_1}^{m_\beta}((\Sigma_i^{old}) + (\mu_i^{old} - \mu_i^{new})(\mu_i^{old} - \mu_i^{new})^T)P(I_k = i \mid \theta^{old}, C_{obs})}{\sum_{k=1}^{N} P(I_k = i \mid \theta^{old}, C_{obs})}$$

The optimization problem to find π_i is constrained by $\sum_{i=1}^{M} \pi_i = 1$ and as a results the Lagrange multiplier λ is introduced:

$$\pi_i^{new} = argmax_{\pi_i}\{\sum_{i=1}^{M} P(I_1 = i|\theta^{old}, C_{obs})log\pi_i + \lambda(\sum_{i=1}^{M}\pi_i - 1)\} \tag{3.25}$$

Taking the derivative of Equation (3.25) with respect to the Lagrange multiplier and π_i, and solving the set of linear equations, the expression in Equation (3.26) is obtained.

$$\pi_i^{new} = P(I_1 = i \mid \theta^{old}, C_{obs}) \tag{3.26}$$

Similarly, the optimization problem to find $\gamma_{ij,i\neq j}$ is constrained by $\sum_{j=1}^{M}\gamma_{ij,i\neq j}=1$. Therefore, Lagrange multiplier λ' is introduced.

$$\gamma_{ij,i\neq j}^{new}=argmax_{\gamma_{ij,i\neq j}}\{\sum_{i=1}^{M}\sum_{j\neq i=1}^{M}\sum_{k=2}^{N}P(I_k=j,I_{k-1}=i|\theta^{old},C_{obs})\times log(\alpha_{ij}(k)) \tag{3.27}$$

$$+\lambda'(\sum_{j\neq i=1}^{M}\gamma_{ij}-1)\}$$

where $\alpha_{ij}(k)$ is introduced as a function of γ_{ij} in Equations (3.3) and (3.4). Taking the derivative of Equation (3.27) with respect to $\gamma_{ij,i\neq j}$ and the Lagrange multiplier and then solving the set of linear equations, the following result is obtained:

$$(\gamma_{ij,i\neq j})^{new}=\frac{\sum_{k=2}^{N}P(I_k=j,I_{k-1}=i\mid\theta^{old},C_{obs})}{\sum_{j\neq i=1}^{M}\sum_{k=2}^{N}P(I_k=j,I_{k-1}=i\mid\theta^{old},C_{obs})} \tag{3.28}$$

Due to the existence of the exponential function in transition probabilities ($\alpha_{ii}(k)$), the unknown parameters σ_{H_i} and γ_{ii} cannot be obtained analytically when maximizing the cost function in Equation (3.29).

$$(\sigma_{H_i},\gamma_{ii})_{1\leq i\leq M}^{new}=argmax_{\sigma_{H_i},\gamma_{ii}}\sum_{i=1}^{M}\sum_{j=1}^{M}\sum_{k=2}^{N}P(I_t=j,I_{t-1}=i|\theta^{old},C_{obs})\times log(\alpha_{ij}(k)) \tag{3.29}$$

$$S.t.\ 0\leq\gamma_{ii}\leq 1,\ \sigma_{H_{min}}\leq\sigma_{H_i}\leq\sigma_{H_{max}},\gamma_{ij}\neq 0$$

Local non-linear optimization methods to find the optimal value of the unknowns (σ_{H_i} and γ_{ii}) are very sensitive to the initial values and it is possible that the optimization problem converges to a local optimum rather than a global. On the other hand, some of the moderate optimization techniques, like Genetic Algorithm (GA), do not require initial values, and if some certain criteria such as parallel searching, efficient interactions between different search trajectories and intelligent steps with appropriate step sizes are considered, the algorithm will normally reach a global solution, or a better local one. However, these methods are time consuming if used for global multivariate optimization of large data-bases in the presence of missing observations [61, 62].

In this chapter, we will use the following procedure in order to find the optimal value of the unknowns without analytical solutions in the Maximization step. Similar

procedures have been previously introduced in literature [63, 64]. Using this procedure, an intelligent random sampling for initialization of some of the parameters in the EM algorithm is used. Unlike previous initialization techniques, which provide completely random initial values and select the one solution with the largest likelihood [34], GA will provide the initial values based on the population's fitness and a target sampling rate. Consequently, the low performance initial values will be generated with very small probabilities [62], and it will be more likely to have the appropriate initial values for the EM algorithm.

1. At the first iteration of the EM algorithm, start the non-linear optimization problem in Equation (3.29) with only a few generations of the Genetic Algorithm.

2. Continue the optimization of the function in Equation (3.29) with results of the GA as the initial values for the local interior point non-linear constrained optimization algorithm (more details are available in the references [65, 66]).

3. Having the optimal values from the previous step, continue the maximization step following Equations (3.23), (3.24), (3.26) and (3.28).

4. Save the calculated optimal values as θ^{old} for the next iteration which starts from step 2.

The initial values for the mean vectors and covariance matrices (μ_i and Σ_i) of the different modes in the EM algorithm can be obtained from an initial solution based on a mixture of multivariate Gaussian distributions assumption for the observations (Y_O). The initial values for $\gamma_{ij,i\neq j}$ can be selected to be equal to 0.5, assuming equal probability for all the transitions. In the cases where operating modes of the scheduling variable are unclear, the mean values of the scheduling variable at each operating mode ($H_i's$) and the initial values for the validity of the scheduling variable at each operating mode (σ_{H_i}) can be obtained assuming that the scheduling variable follows a mixture of Gaussian distributions.

The optimization problem will iterate between the E and M steps until the convergence criterion is satisfied. The convergence criterion in this framework is set to be less than the absolute value of the likelihood change in two successive iterations.

Furthermore, $P(I_k = j, I_{k-1} = i \mid \theta^{old}, C_{obs})$ and $P(I_k = i \mid \theta^{old}, C_{obs})$ are required to complete the maximization step in Equations (3.23) to (3.29). These terms are calculated as follows:

$$if\ Y_k\ is\ observed,$$

$$P(I_k = j, I_{k-1} = i \mid \theta^{old}, C_{obs}) = P(I_k = j, I_{k-1} = i \mid Y_{t_1}, ..., Y_{t_\alpha}, \theta^{old}, H_1, ..., H_N) =$$

$$\frac{f(Y_{t_1}, ..., Y_{t_\alpha} \mid I_k = j, I_{k-1} = i, \theta^{old}, H_1, ..., H_N) \times P(I_k = j, I_{k-1} = i \mid \theta^{old}, H_1, ..., H_N)}{\sum_{i=1}^{M} \sum_{j=1}^{M} f(Y_{t_1}, ..., Y_{t_\alpha} \mid I_k = j, I_{k-1} = i, \theta^{old}, H_1, ..., H_N) \times P(I_k = j, I_{k-1} = i \mid \theta^{old}, H_1, ..., H_N)}$$

$$= \frac{f(Y_k \mid I_k = j, \theta^{old}) P(I_k = j \mid I_{k-1} = i, H_{k-1}, \theta^{old}) \times P(I_{k-1} = i \mid \theta^{old}, H_1, ..., H_{k-2})}{\sum_{i=1}^{M} \sum_{j=1}^{M} f(Y_k \mid I_k = j, \theta^{old}) P(I_k = j \mid I_{k-1} = i, H_{k-1}, \theta^{old}) \times P(I_{k-1} = i \mid \theta^{old}, H_1, ..., H_{k-2})} \tag{3.30}$$

$$if\ Y_k\ is\ missing,$$

$$P(I_k = j, I_{k-1} = i \mid \theta^{old}, C_{obs}) =$$

$$P(I_k = j, I_{k-1} = i \mid \theta^{old}, H_1, ..., H_N) =$$

$$P(I_k = j \mid I_{k-1} = i, \theta^{old}, H_{k-1}) P(I_{k-1} = i \mid \theta^{old}, H_1, ..., H_{k-2}) \tag{3.31}$$

where in Equations (3.30) and (3.31), $P(I_k = j \mid I_{k-1} = i, \theta^{old}, H_{k-1}) = \alpha_{ij}(k)^{old}$, $f(Y_k \mid I_k = j, \theta^{old})$ follows the multivariate normal distribution in Equation (3.1) with mean vector and covariance matrix μ_i^{old} and Σ_i^{old} obtained from the previous iteration, and $P(I_{k-1} = i \mid \theta^{old}, H_1, ..., H_{k-2})$ is obtained through discrete-valued state propagation of Markov chain starting from the initial value of $P(I_1 = i \mid \theta^{old}, C_{obs}) = \pi_i^{old}$.

Finally, $P(I_k = i \mid \theta^{old}, C_{obs})$ can be obtained from summation of $P(I_k = i, I_{k-1} = j \mid \theta^{old}, C_{obs})$ over all the possible modes for I_{k-1}, i.e.,

$$P(I_k = i \mid \theta^{old}, C_{obs}) = \sum_{j=1}^{M} P(I_k = i, I_{k-1} = j \mid \theta^{old}, C_{obs}) \tag{3.32}$$

3.4 Operating Mode Recognition

On-line operating mode (state) recognition is needed for fault detection. Through an on-line application, probability of the hidden process modes given the observations ($P(I_k \mid Y_k, ..., Y_1, H_k, ..., H_1)$) can be calculated using Hamilton's filtering strategy [44] as follows:

1. Calculate the joint probability of the modes I_k and I_{k-1} given the information up to time $k-1$:

$$P(I_k, I_{k-1} \mid Y_{k-1}, ..., Y_1, H_{k-1}, ..., H_1) = \tag{3.33}$$

$$P(I_k \mid I_{k-1}, H_{k-1}) P(I_{k-1} \mid Y_{k-1}, ..., Y_1, H_{k-1}, ..., H_1)$$

where $P(I_k \mid I_{k-1}, H_{k-1}) = \alpha_{ij}(k)$, and $P(I_{k-1} \mid Y_{k-1}, ..., Y_1, H_{k-1}, ..., H_1)$ is the previous output of the filter.

2. Update the probability of the modes I_k, I_{k-1} using the new observations at time k:

$$P(I_k, I_{k-1} \mid Y_k, ..., Y_1, H_k, ..., H_1) = \frac{P(Y_k, I_k, I_{k-1} \mid Y_{k-1}, ..., Y_1, H_k, ..., H_1)}{P(Y_k \mid Y_{k-1}, ..., Y_1, H_k, ..., H_1)} \tag{3.34}$$

where

$$P(Y_k, I_k, I_{k-1} \mid Y_{k-1}, ..., Y_1, H_k, ..., H_1)$$

$$= P(Y_k \mid I_k, I_{k-1}, Y_{k-1}, ..., Y_1, H_k, ..., H_1) \times P(I_k, I_{k-1} \mid Y_{k-1}, ..., Y_1, H_k, ..., H_1)$$

$$= f(Y_k \mid I_k) P(I_k, I_{k-1} \mid Y_{k-1}, ..., Y_1, H_{k-1}, ..., H_1)$$

where $f(Y_k \mid I_k)$ follows the multivariate normal distribution in Equation (3.1) and $P(I_k, I_{k-1} \mid Y_{k-1}, ..., Y_1, H_{k-1}, ..., H_1)$ is known from Equation (3.33). The denominator in Equation (3.34) can be calculated as

$$P(Y_k \mid Y_{k-1}, ..., Y_1, H_k, ..., H_1)$$

$$= \sum_{I_k=1}^{M} \sum_{I_{k-1}=1}^{M} P(Y_k, I_k, I_{k-1} \mid Y_{k-1}, ..., Y_1, H_k, ..., H_1)$$

3. The output of the filter will be,

$$P(I_k \mid Y_k, ..., Y_1, H_k, ..., H_1) = \sum_{I_{k-1}=1}^{M} P(I_k, I_{k-1} \mid Y_k, ..., Y_1, H_k, ..., H_1) \tag{3.35}$$

3.5 Results and discussion

As previously stated in the introduction and problem statement sections, the advantage of using the proposed structure of this chapter over conventional HMMs can be very well demonstrated when a process has an asymmetric time varying transition behavior between different operating modes, i.e., when some of the operating modes are far from the majority and the scheduling variable helps in the modeling and filtering steps by providing more flexibility. Examples of such situations are demonstrated in the case studies of this section.

3.5.1 A Numerical Case Study

In this simulation case study, we consider a system that operates in four operating modes which are normal (mode 1), abnormal 1 (mode 2), abnormal 2 (mode3) and fault (mode 4). Observations of each mode are assumed to follow different multivariate normal distributions. The abnormal modes (modes 2 and 3) are assumed to be close to each other and far from the normal and faulty modes as in Table 3.1. Operating modes of the process follow the structure proposed in Figure 3.1. In order to test the validity of the algorithm for missing observations, some observations (10 %) are randomly missed in the simulation at various sampling instants. The missing data are assumed to be completely missing at random. The scheduling variable is assumed to linearly transit between different operating conditions. Parameters used for the simulation are presented in Table 3.1. Parameters γ_{13}, γ_{14}, γ_{24}, γ_{34}, γ_{41} and γ_{43} can be further obtained from Equations (3.6), (3.7) and (3.8) respectively.

Table 3.1: System parameters to generate the simulation data

$\pi_0 = [0.25, 0.25, 0.25, 0.25]$
$\gamma_{11} = 0.98, \gamma_{12} = 0.5, \gamma_{21} = 0.5, \gamma_{22} = 0.95, \gamma_{23} = 0.3$
$\gamma_{31} = 0.2, \gamma_{32} = 0.6, \gamma_{33} = 0.89, \gamma_{42} = 0.4, \gamma_{44} = 0.92$
$\mu_1 = [5\ 3], \mu_2 = [10\ 8], \mu_3 = [11\ 9], \mu_4 = [18\ 16]$
$\Sigma_1 = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 2.5 \end{pmatrix}, \Sigma_2 = \begin{pmatrix} 3 & 0.75 \\ 0.75 & 4.5 \end{pmatrix}$
$\Sigma_3 = \begin{pmatrix} 4 & 0.4 \\ 0.4 & 5.5 \end{pmatrix}, \Sigma_4 = \begin{pmatrix} 2 & 0.6 \\ 0.6 & 0.5 \end{pmatrix}$
$\sigma_{H_1} = 5, \sigma_{H_2} = 4, \sigma_{H_3} = 4.5, \sigma_{H_4} = 3$

Following the expectation-maximization procedure introduced in section 3, the final estimated value of the parameters from a training data-set including 8000 data

are obtained as presented in Table 3.2. As it is clear from this table, the estimated parameters are close to true parameters of the process. Due to the nature of the industrial process which will be later used in this chapter, with small sampling rates and large number of operating modes, using such large training data-sets are more appropriate for the proposed method to provide robust process identification results. However, size of the training data set might vary for other industrial applications according to their sampling rate and number of operating modes.

Table 3.2: Estimated parameters from the EM algorithm

$$\pi_0 = [0.25, 0.25, 0.25, 0.25]$$
$$\gamma_{11} = 0.9912, \gamma_{12} = 0.6633, \gamma_{21} = 0.6756, \gamma_{22} = 0.9775, \gamma_{23} = 0.2882$$
$$\gamma_{31} = 0.3037, \gamma_{32} = 0.5379, \gamma_{33} = 0.8861, \gamma_{42} = 0.3470, \gamma_{44} = 0.9397$$
$$\mu_1 = [5.1991\ 3.2050], \mu_2 = [9.8667\ 7.8267]$$
$$\mu_3 = [10.8162\ 8.7122], \mu_4 = [17.0019\ 14.9031]$$
$$\Sigma_1 = \begin{pmatrix} 1.2727 & 0.7264 \\ 0.7264 & 3.0525 \end{pmatrix}, \Sigma_2 = \begin{pmatrix} 3.4465 & 1.0722 \\ 1.0722 & 4.7376 \end{pmatrix}$$
$$\Sigma_3 = \begin{pmatrix} 4.7419 & 0.6518 \\ 0.6518 & 5.6666 \end{pmatrix}, \Sigma_4 = \begin{pmatrix} 2.6854 & 0.6330 \\ 0.6330 & 0.6890 \end{pmatrix}$$
$$\sigma_{H_1} = 5.2613, \sigma_{H_2} = 4.1880, \sigma_{H_3} = 4.3707, \sigma_{H_4} = 3.4833$$

In order to compare the performance of the proposed method in operating condition diagnosis with conventional HMMs [25, 26], first, another data-set including 8000 data is generated from the same model in Table 3.1 for training purposes. Since conventional hidden Markov models cannot deal with missing observations, the complete data set is assumed to be observable and the performance is only compared in the adaptive property of the new technique rather than handling of the missing data. Next, a validation data-set is generated from the same model, which is presented in Figure 3.2. Results of the filtering procedure to find the probability of the hidden operating modes given observations based on the proposed method of this chapter (Section 3.4) are presented in Figure 3.3. Based on these probabilities, true and the estimated operating modes of the process are presented in Figure 3.4.

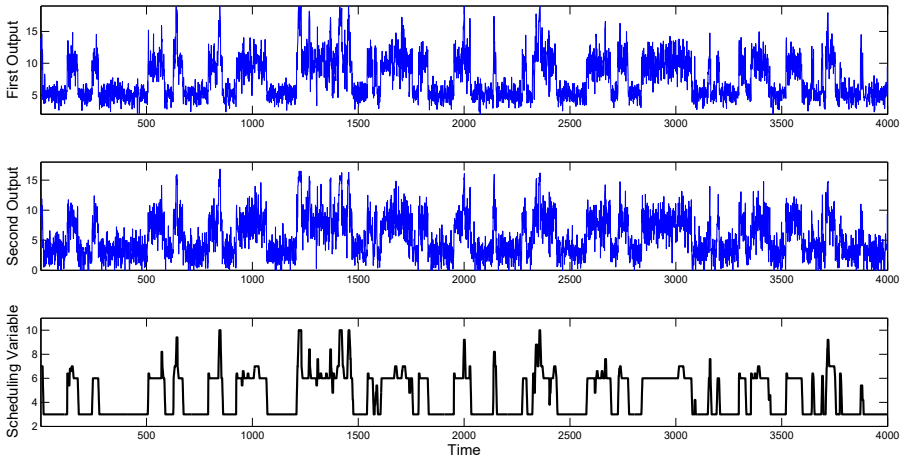

Figure 3.2: Validation data-set

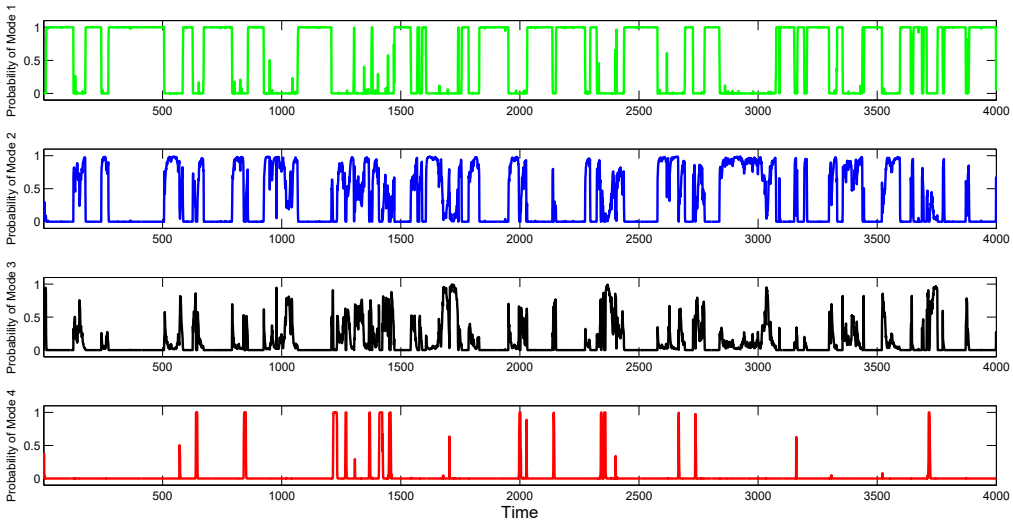

Figure 3.3: Probability of the hidden operating modes for the validation data-set using the proposed method of this chapter

As presented in Figure 3.4, except for some very fast changes in the dynamics of the process which cause some false alarms (time instants around 1700 and 2700 for example), the method is generally able to detect the true operating mode of the process.

Results of the probability calculation and operating mode recognition based on conventional multivariate HMMs are presented in Figures 3.5 and 3.6.

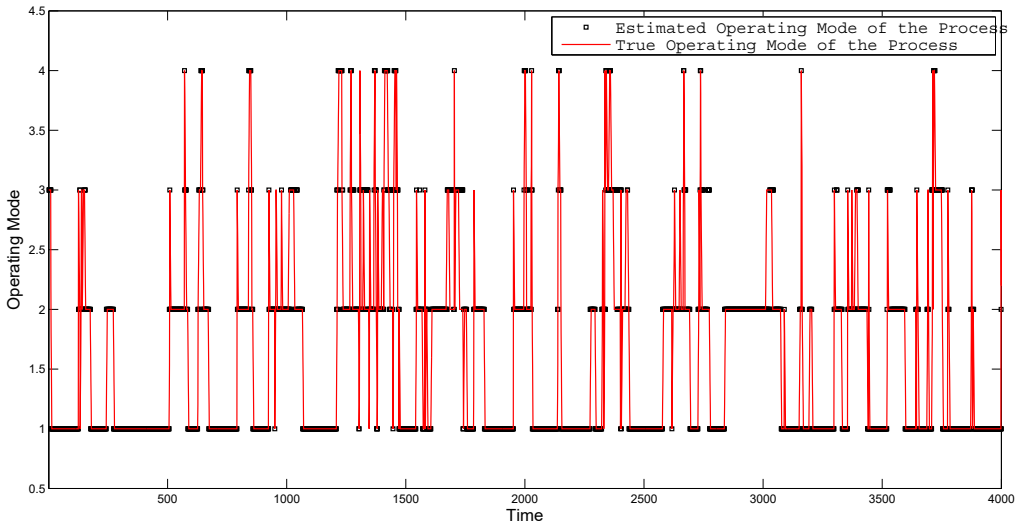

Figure 3.4: True and estimated hidden operating modes of the process for the validation data-set using the proposed method of this chapter

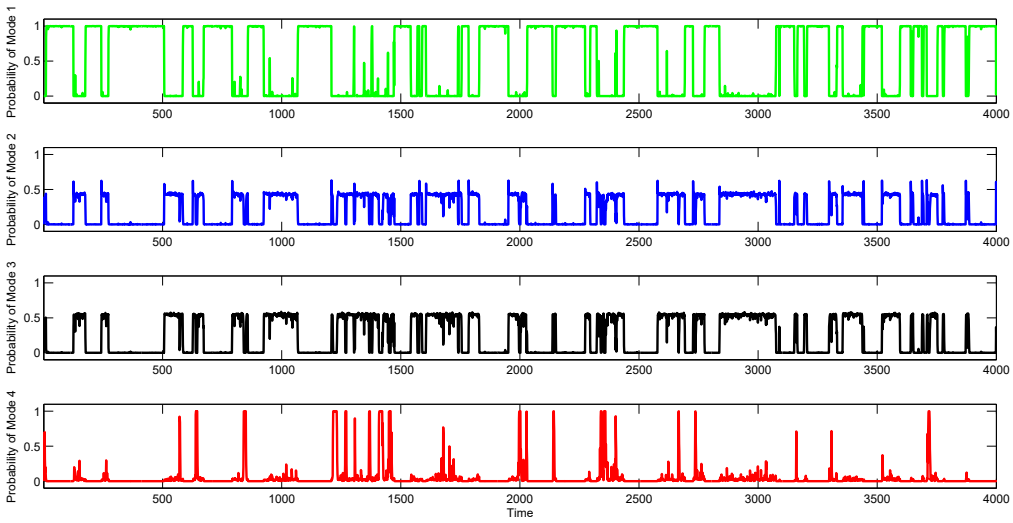

Figure 3.5: Probability of the hidden operating mode for the validation data-set based on a conventional multivariate hidden Markov model

As presented in Figure 3.5, applying conventional HMMs, the developed model is unable to distinguish between the two close abnormal operating modes (modes 2 and 3), i.e., probability of the observations given these modes are close to each other and close to 0.5. Consequently, in Figure 3.6, one could see that based on the probabilities in Figure 3.5, at several time instants, operating mode 2 is incorrectly categorized as

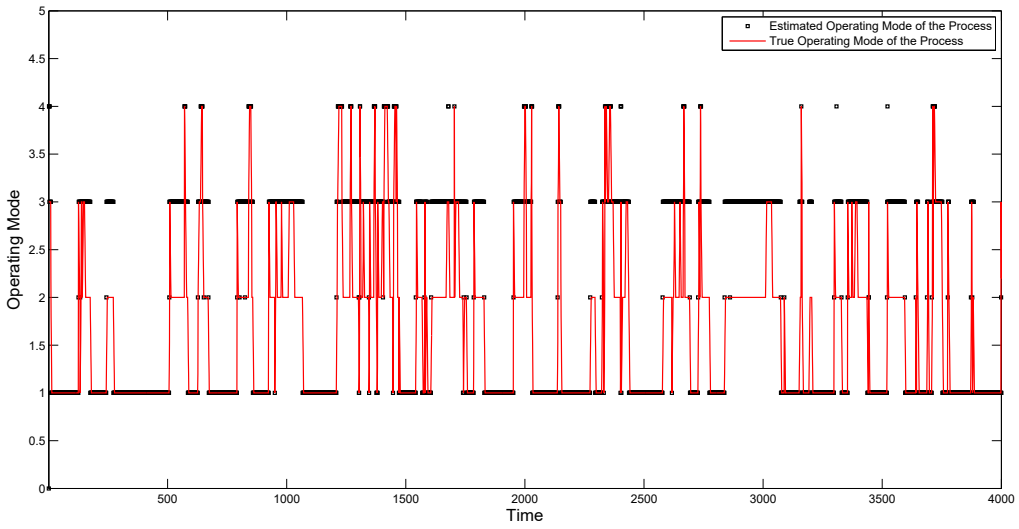

Figure 3.6: True and estimated hidden operating modes of the process for the validation data-set based on a conventional multivariate hidden Markov model

3.

This example illustrates one of the cases where applying time-varying transition probabilities based on the distributions introduced in Equations (3.3) and (3.4) can provide more accurate predictions for the true hidden operating mode of the process. As this example shows, existence of an accurate scheduling variable can provide more flexibility to model and monitor the process transitions between different operating modes.

3.5.2 A Simulation Study

In this example, the proposed method is tested on the two CSTRs in series introduced by Henson et al. [67]. The irreversible exothermic first order reaction $A \longrightarrow B$ occurs in the two reactors in series. The feed enters the first reactor with flow rate q_f and temperature and concentration C_{Af} and T_f respectively. The product of the first reactor is then feed to the second reactor. A parallel flow (q_c) is used as the coolant. The process is illustrated in Figure 3.7.

The process works in open loop condition. Concentration of the product C_{A2} is the important output variable for control purposes. In the steady state condition C_{A2} is around $0.05(\frac{mol}{L})$.

Temperature of the first reactor (T_1), which can provide a pre-indication to the

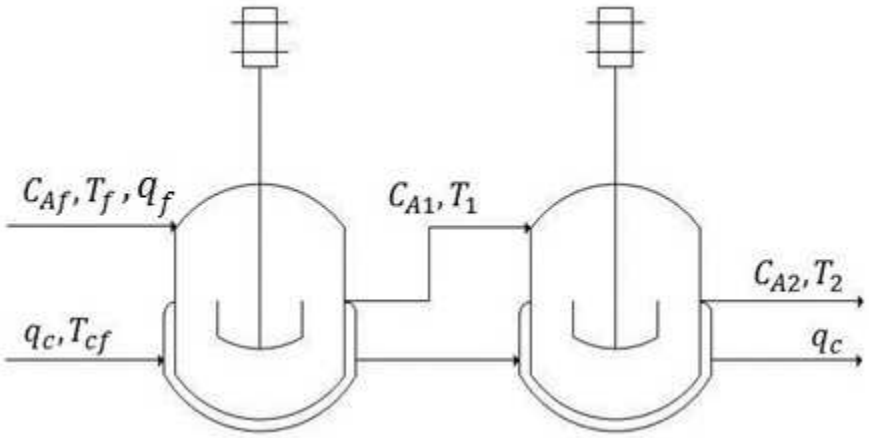

Figure 3.7: Two CSTRs in series [67]

operating condition of the process before receiving the final product, is selected as the scheduling variable. It is assumed that there is no measurement for the coolant flow-rate ($q_c(\frac{L}{min})$) and coolant flow-rate is selected as the disturbance with variance 10.

The main cause of the changes in operating condition of the process is the sudden changes in the feed flow-rate (q_f). In this example, the feed flow-rate is assumed to vary between 3 operating modes following the Markov switching model given in Equation 3.5. Normal (mode 1) and abnormal (mode 2) operations occur when the feed flow-rate is around its steady-state value, i.e., $q_{f(Mode_1)} = 105.4(\frac{L}{min})$ and $q_{f(Mode_2)} = 112.6(\frac{L}{min})$. The faulty mode occurs when feed flow-rate suddenly increases ($q_{f(Mode_3)} = 134.3(\frac{L}{min})$) and the coolant flow-rate is not enough to maintain a constant process temperature. In such situations, the output temperature (T_2) suddenly increases. This is followed by a very low product concentration (C_{A2}). Therefore, these two key variables (T_2 and C_{A2}) are selected as indicators of the operating condition of the process. An example of the normal, abnormal and faulty operating conditions of the process is illustrated in Figure 3.8. The output temperature is in Kelvin unit. Parameters of the transition probability matrix, which cause the switching behavior in the feed flow-rate, are selected as $\gamma_{11} = 0.95$, $\gamma_{21} = 0.7$, $\gamma_{22} = 0.93$, $\gamma_{33} = 0.97$, $\sigma_{H1} = 15$, $\sigma_{H2} = 13$ and $\sigma_{H3} = 10$. Parameters γ_{12}, γ_{23}, and γ_{32} can be further obtained from Equations (3.6), (3.7) and (3.8) respectively.

The scheduling variable and the disturbance to the process are presented in Figure 3.9. Applying a moving average filter, the scheduling variable is filtered to provide an

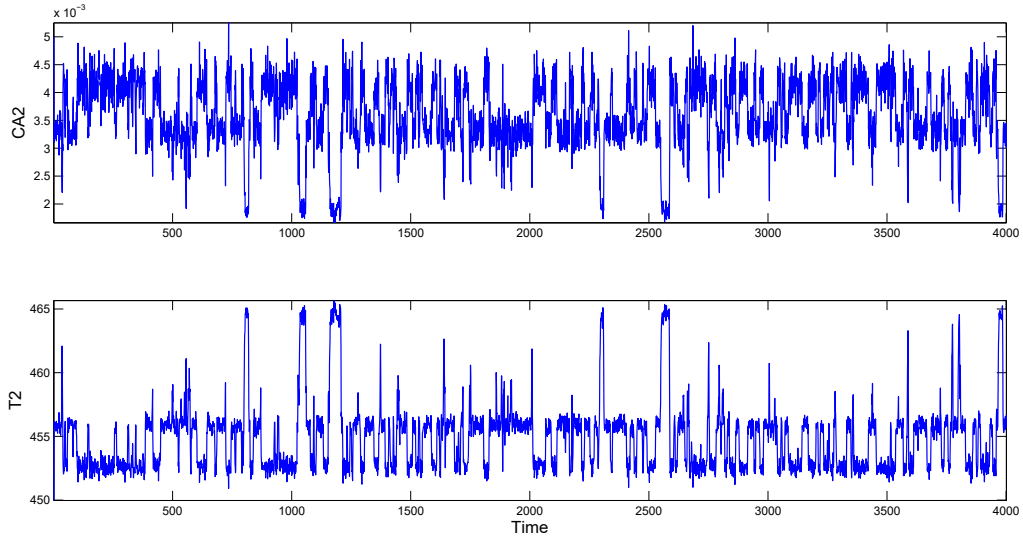

Figure 3.8: Different operating modes for the process variables (validation data-set)

overal indication for the transitions of the process between different operating modes. According to the previous discussion on the feed flow rate, in this example, the faulty mode is selected far from the normal and abnormal operations of the process. Similar to the previous example, this asymmetric mode transition provides a condition to more clearly demonstrate the advantage of the proposed method of this chapter over conventional techniques.

The historical data set including 8000 data points, which are different but from the same model as Figure 3.8, is used for training purposes. Similar to previous example, 10 % of the training data are assumed to be randomly missing. In industrial applications, such missing information might be due to sudden shifts in process status. Results of the parameter estimation based on expectation maximization algorithm introduced in section 3 are presented in Table 3.3.

In order to compare the performance of the proposed method in this chapter and conventional HMMs, both methods are tested on the data in Figures 3.8 and 3.9. The training data-set to train conventional HMMs is the same as the historical data-set that is used for parameter estimation in Table 3.3. However, here, the complete data-set is assumed to be observable since conventional HMMs cannot deal with the missing observations. Therefore, the comparison is made only in the time-varying property of the new model.

Using the proposed method of this chapter, probability of the different operating

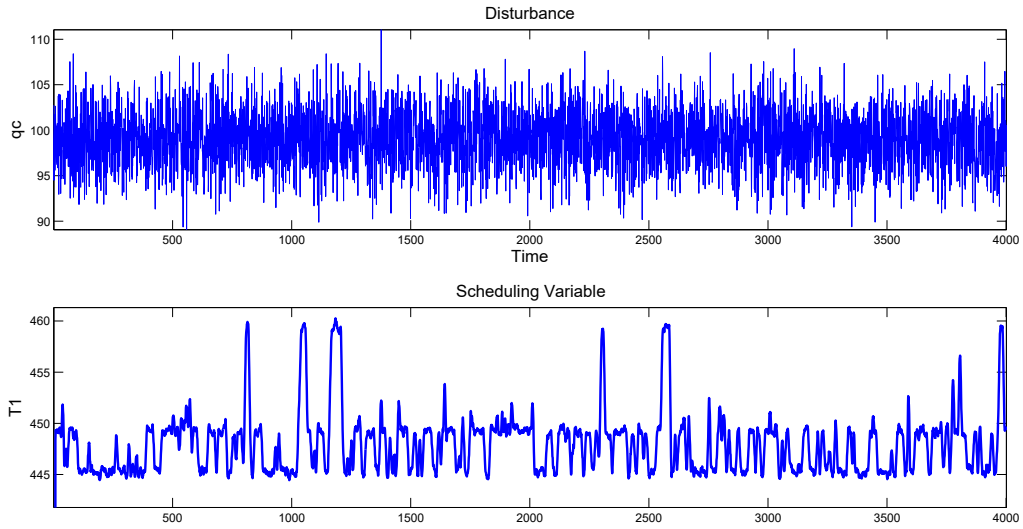

Figure 3.9: The scheduling variable and disturbance to the process (validation dataset)

Table 3.3: Estimated parameters for the CSTRs in series using the EM algorithm

$$\pi_0 = [0.3333, 0.3333, 0.3333]$$
$$\gamma_{11} = 0.9472, \gamma_{21} = 0.6301, \gamma_{22} = 0.9142, \gamma_{33} = 0.9968,$$
$$\mu_1 = [0.0041\ 452.7567], \mu_2 = [0.0033\ 455.8058], mu_3 = [0.0023\ 461.5816]$$
$$\Sigma_1 = \begin{pmatrix} 9.2697 \times 10^{-8} & -2.0526 \times 10^{-4} \\ -2.0526 \times 10^{-4} & 0.5908 \end{pmatrix}$$
$$\Sigma_2 = \begin{pmatrix} 7.1064 \times 10^{-8} & -2.1382 \times 10^{-4} \\ -2.1382 \times 10^{-4} & 0.7967 \end{pmatrix}$$
$$\Sigma_3 = \begin{pmatrix} 3.0473 \times 10^{-7} & -1.8 \times 10^{-3} \\ -1.8 \times 10^{-3} & 11.3670 \end{pmatrix}$$
$$\sigma_{H_1} = 0.9089, \sigma_{H_2} = 2.7259, \sigma_{H_3} = 9.8989$$

modes given new observations are presented in Figure 3.10. Based on these probabilities, true and estimated operating conditions of the process are demonstrated in Figure 3.11.

As presented in Figure 3.11, similar to the previous case study, other than some very fast changes in the process variables, the proposed method of this chapter is able to detect the different operating modes of the process.

Results of the filtering and operating condition diagnosis based on conventional HMMs are presented in Figures 3.12 and 3.13 respectively.

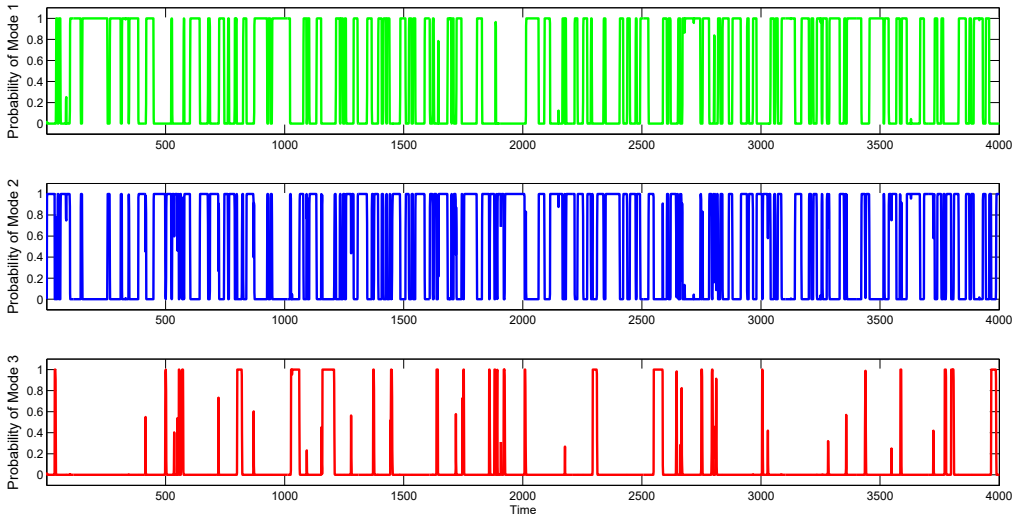

Figure 3.10: Probability of the hidden operating modes for the CSTRs in series based on the proposed method of this chapter and the validation data-set in Figure 3.8

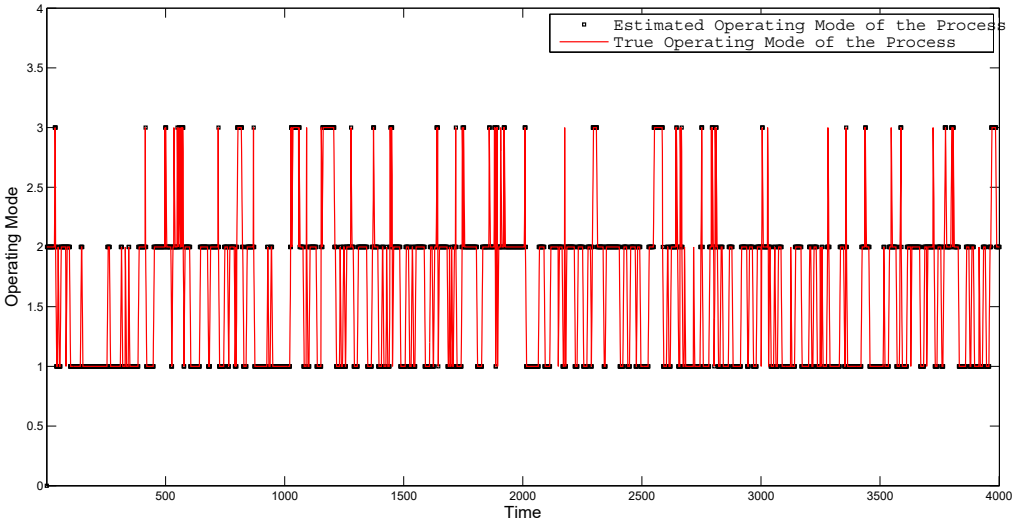

Figure 3.11: True and estimated operating modes of the process for the CSTRs in series based on the proposed method of this chapter and the validation data-set in Figure 3.8

Comparing the results in Figures 3.11 and 3.13, one could easily observe that the conventional method provides several false alarms in detection of the faulty mode. This is another good example to show the merit of the proposed method of this chapter for the cases where the process operates among asymmetric operating modes and adaptive transition probabilities provide more flexibility for overall monitoring of

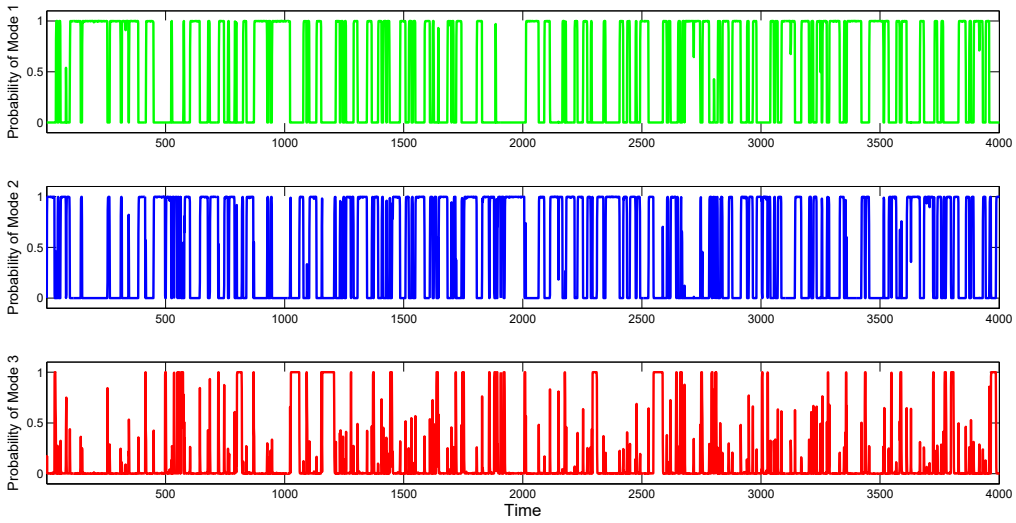

Figure 3.12: Probability of the hidden operating modes for the CSTRs in series using conventional HMMs and the validation data-set in Figure 3.8

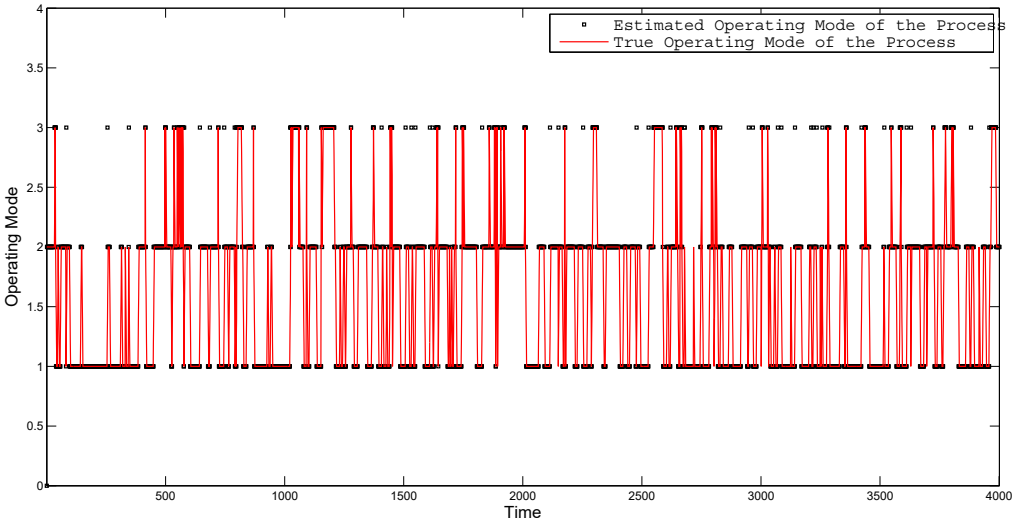

Figure 3.13: True and estimated operating modes of the process for the CSTRs in series using conventional HMMs and the validation data-set in Figure 3.8

the process. In this example, existence of the scheduling variable helps the model to adapt to the new modes. Therefore, the new framework eventually enables the model to better detect the faulty mode which behaves far from the normal and abnormal modes.

3.5.3 An Industrial Case Study

This industrial case study will be presented in Chapter 6, and compared to the methods in other chapters of the thesis.

3.6 Conclusion

In this chapter a novel approach for modeling and monitoring of the time-varying multivariate regime switching systems subject to fault and missing observations is introduced. Due to the existence of the missing observations and unknown operating modes, the EM algorithm is applied to find the unknown parameters. Also, an optimization procedure is introduced to reduce the sensitivity of the EM algorithm to its initial values.

The proposed method is tested on two simulations and one industrial case studies and shows a superior performance in detecting the different operating modes of the process in comparison to the conventional methods. In general, application of time varying transition probabilities, as introduced in this chapter, provides better classifications for different operating modes of the process. This improvement can be more clearly observed in processes with asymmetric time varying transitions between different operating modes.

Chapter 4

Robust Diagnosis of Operating Mode Based on Time Varying Hidden Markov Models

In this chapter, the time varying HMM structure of previous chapter is further improved to deal with negative effect of irregular measurements such as outliers. Consequently, this chapter proposes a robust process monitoring and diagnosis strategy based on HMMs with time varying transition probabilities. We model observations around different process operating modes by different multivariate Student t-distributions to describe different likelihoods of outliers. Time varying transition probabilities assist the model to adapt to new operating conditions. The quality of data in each mode, which is usually affected by the percentage of outliers, is treated by assigning the appropriate degree of freedom for the multivariate t distribution adaptively. The method is compared with other available recent techniques in literature using simulation and experimental studies and shows a superior performance.

4.1 Introduction

With advent of modern measurement and data storage techniques, data driven process monitoring methods have become popular. Recently, Yin et al. performed two review studies to compare the available data based process monitoring methods [12, 13]. Partial Least Squares (PLS) and Principal Component Analysis (PCA) are two of the well known methods in this area with frequent applications in industrial processes [68, 69]. Dynamic approaches to such data classification methods have become of a

A version of this chapter has been published in N. Sammaknejad, B. Huang, Y. Lu (2015). Robust Diagnosis of Operating Mode Based on Time Varying Hidden Markov Models. IEEE Transactions on Industrial Electronics. DOI: 10.1109/TIE.2015.2478743.

great interest in last decade. Hidden Markov Models (HMMs) provide an appropriate mathematical tool to handle such problems. Bruckner et al. proposed a statistical approach based on HMMs to monitor sensor data [70]. Different layers of their hierarchical model structure correspond to different components of the real process. In a new study, an adaptive framework to process monitoring based on HMMs and symbolic episode representation is proposed [71]. Jiang et al. proposed a new method for monitoring of the gear shaft system. They modeled the process using a three state homogeneous Markov process [23]. Jager et al. developed a combination of dimension reduction based on PCA and HMMs to monitor laser welding processes [72]. They showed that, in such cases, HMMs are able to more accurately model the temporal behavior of the observations. In the previous chapter, a general multivariate framework for process monitoring based on HMMs with time varying transition probabilities is proposed. Consideration of time varying transition probabilities provides an appropriate structure to model processes with time dependent shifts among different working conditions [45].

Similar process monitoring techniques have been applied to real industrial processes. In recent studies, Soualhi et al., Boukra et al. and Gritli et al. suggested novel fault detection methods for induction motors [73, 74, 75]. Fault detection in vehicle motors and steering systems are other examples of such applications in real life processes [76, 77]. Application of all these process monitoring strategies for real industrial case studies might be significantly affected by data quality. Outliers, which are usually caused by sensor malfunctions, human errors in data collection and experiment conduction, and unusual process disturbances, can cause a biased parameter estimation [78]. Conventional outlier removal techniques, which are based on certain thresholds obtained from normal process operation data, might cause loss of information. Previous studies show that parameter estimation for such problems under the assumption that process observations follow Gaussian, or mixture Gaussian distributions might result in inaccurate estimations [79]. Jin et al. proposed to use a contaminated Gaussian distribution to reduce the negative effect of the outliers [80]. The idea is to assume that the noise term follows a mixture of Gaussian distributions. However, a fixed variance is considered for the Gaussian component corresponding to the outliers. A more general approach to deal with the robustness issue is to assume that observations follow a t distribution [81]. Small values of the degree of freedom in a t distribution will provide heavier tails which significantly downweight the effect of outliers during the identification process [29]. In a very recent study, Lu et al. proposed a general framework for robust identification of nonlinear processes [82]. In their study, the noise term is assumed to follow a t distribution and the degree

of freedom is estimated according to the quality of the data. Other applications of the Expectation Maximization (EM) algorithm for fault diagnosis in the presence of missing observations have been recently addressed in literature [83].

In this chapter, we provide a robust approach for process operation mode diagnosis using HMMs with a time dependent structure for the transition probability matrix. Time varying structure of the model, as proposed in the previous chapter, provides a condition to more appropriately model process transitions. However, effect of the data quality on the distribution of observations in each operating mode has not been considered. Here, robustness is considered by assigning the appropriate degree of freedom to the multivariate t distribution for the observations in each operating mode. Consequently, in industrial studies where it is expected to have lower quality data when operating near the faulty modes, assigning the appropriate degree of freedom according to the data quality will assist the model to diagnose process operation conditions more robustly. In comparison to the previous studies on robust data modeling using HMMs ([84, 85, 86]), application of time varying transition probabilities will further assist the model to adapt to new operating conditions. This provides a more general framework for monitoring of the industrial processes with a time varying behavior with respect to both operating mode and data quality. EM algorithm is applied to solve the problem. The developed strategy is tested on simulation and experimental examples and all demonstrate a superior performance over the existing techniques.

The remaining sections of this chapter are arranged as follows: In Section 4.2, problem formulation based on the t distribution in the presence of time varying transition probabilities is introduced. Section 4.3 provides a robust iterative procedure to estimate the unknown parameters under the EM framework. Section 4.4 explains the filtering procedure to diagnose the operation mode during on-line applications. In Section 4.5, the proposed method is tested on simulation and experimental examples, and compared to the available recent techniques from literature. Section 4.6 is the conclusion.

4.2 Problem Statement

The observed data set for the parameter estimation purpose of this chapter is given as

$$Y = \begin{pmatrix} y_{11} & y_{12} & \cdots & y_{1N} \\ y_{21} & y_{22} & \cdots & y_{2N} \\ \vdots & \vdots & \vdots & \vdots \\ y_{P1} & y_{P2} & \cdots & y_{PN} \end{pmatrix} = (Y_1, Y_2, ..., Y_N) \tag{4.1}$$

where P indicates the number of process variables and N is the size of the data set.

To account for outliers which typically have unusual large or small values, each observation vector Y_k (χ in Equation 4.2) in the data set in Equation 4.1 is considered to follow a multivariate t distribution as follows:

$$t_P(\chi|\mu, \Sigma, \nu) = \frac{\Gamma(\frac{\nu+P}{2})|\Sigma|^{-1/2}}{(\pi\nu)^{P/2}\Gamma(\frac{\nu}{2})\{1 + \frac{\delta(\chi|\mu,\Sigma)}{\nu}\}^{(\nu+P)/2}} \tag{4.2}$$

where χ is the vector of observations, μ is the location (mean related variable), Σ is the positive definite inner product matrix (covariance matrix related variable) and ν is the degree of freedom in each operating mode [29]. The Γ function is given as $\Gamma(\alpha) = \int_0^\infty z^{\alpha-1}e^{-z}dz$ and δ is the squared Mahalanobis distance between χ and μ with the covariance matrix Σ, that is, $\delta(\chi|\mu, \Sigma) = (\chi - \mu)\Sigma^{-1}(\chi - \mu)$ [29, 82].

In this study, the process is considered to operate in various modes. These operating modes might be considered as different working conditions of the process, for example, normal, abnormal and faulty. The set of discrete modes is presented as $I_k \in \{I_1, ..., I_N\}$ where $I_k = i$, $1 \leq i \leq M$, and M is a positive integer corresponding to the number of available modes.

In general, one could consider a multivariate t distribution as a multivariate normal distribution with varying weighted covariance matrix [29]. Having said this, and according to the definition of the operating mode mentioned in the previous paragraph, distribution of the observation vector at time k given the current operating mode can be expressed as,

$$f(Y_k|I_k = i, R_k) \sim N_P(\mu_i, \Sigma_i/R_k) \tag{4.3}$$

$$N_P(\mu_i, \Sigma_i/R_k) = (2\pi)^{-P/2}|\Sigma_i/R_k|^{-1/2}exp(-\frac{1}{2}(Y_k - \mu_i)^T(\Sigma_i/R_k)^{-1}(Y_k - \mu_i))$$

where μ_i and Σ_i are the mean vector and covariance matrix related variables for mode i, and R_k is the assigned scalar weight for the observation vector Y_k.

According to the definition of the t distribution, the random variable corresponding to the weight R_k should follow a *gamma* distribution [29]. In this study, according to Equation 4.4, weights are considered to follow different *gamma* distributions for different modes. This provides a condition for the t distribution to adapt to data quality in various modes.

$$g(R_k|I_k = i) \sim gamma(\frac{1}{2}\nu_i, \frac{1}{2}\nu_i) \tag{4.4}$$

where ν_i is the degree of freedom in each mode. The density function in Equation 4.4 is a special case of the general form of the *gamma* density function in Equation 4.5 [29].

$$gamma(\alpha, \beta) = \frac{\beta^\alpha r^{\alpha-1} e^{-\beta r}}{\Gamma(\alpha)}, r > 0, \alpha > 0, \beta > 0 \tag{4.5}$$

By integrating out the weight R_k from the joint density function of the observations and weight formed from Equations 4.3 and 4.4, the general form of the multivariate t distribution in Equation 4.2 is obtained [29]. Also, it can be proved that the multivariate t distribution in Equation 4.2 becomes marginally Gaussian as the degree of freedom tends to infinity [29]. However, in comparison to the Gaussian case, the current form of the distribution in Equation 4.2 provides a condition to downweight the effect of observations with large Mahalanobis distances (outliers) [29]. This will be further discussed in the next section.

In order to model the transition behavior of the process between different operating modes, a procedure similar to the previous chapter is developed for this problem. In this structure, a Markov chain models the transitions of the process among various modes. But, transition probabilities are time dependent, according to the variations in the scheduling variable, as follows [45]:

$$\alpha(k)_{ij} = P(I_k = j \mid I_{k-1} = i, H_{k-1}) \tag{4.6}$$

where $\alpha(k)_{ij}$ is the time varying transition probability which is dependent on the scheduling variable H_{k-1}.

This structure provides more flexibility for modelling of the processes with time varying shifts among various modes [45]. The following structure is considered for the transition probabilities [45]:

$$\alpha(k)_{ii} = \frac{2\xi_i exp(\frac{-(H_{k-1} - H_i)^2}{2\sigma^2_{H_i}})}{1 + exp(\frac{-(H_{k-1} - H_i)^2}{2\sigma^2_{H_i}})} \tag{4.7}$$

$$\alpha(k)_{ij,i\neq j} = \gamma_{ij}(1 - \alpha(k)_{ii}) \tag{4.8}$$

where H_i is the mean value and σ_{H_i} is the validity of the scheduling variable in mode i. The procedure to find H_i through the historical data has been explained in the previous chapter. γ_{ij} provides more flexibility for estimation of the transition probability $\alpha(k)_{ij,i\neq j}$. Note that, according to Equations 4.7 and 4.8, when H_{k-1} is close to H_i, the probability of remaining in the current operating mode increases. Otherwise, $\alpha(k)_{ii}$ reduces, which results in an increase in the switching probability $\alpha(k)_{ij}$.

According to the fact that some industrial processes, e.g., many chemical processes, have infrequent transitions between operating modes, ξ_i, which is the corresponding term for $\gamma_{ij,i=j}$, usually takes values close to one due to tendency to remain in the current operating condition [45]. In the first example of this article, $\xi_i = 1 - \varepsilon_i$ $(0 \leq \varepsilon_i \leq 0.1)$ will however be treated as a tuning parameter. Note that this will not significantly affect the generality of the algorithm since $\alpha(k)_{ii}$ can still take all the values from greater than 0 to ξ_i according to Equation 4.7. Having this simplification, the computation complexity of the general form of the non-linear optimization can be greatly reduced.

Owing to the existence of hidden variables such as the operating mode, the EM algorithm for maximum likelihood estimation is adopted. The observed data set (C_{obs}) include $Y = \{Y_1, ..., Y_N\}$ and the scheduling variable $H = \{H_1, ..., H_N\}$. The missing data set or hidden variables (C_{miss}) include the discrete operating modes $I = \{I_1, ..., I_N\}$, and the weight factors $R = \{R_1, ..., R_N\}$ corresponding to each mode. The unknown parameters (θ) to be estimated include the mean vectors, covariance matrices and degrees of freedom for the multivariate t distribution for each operating mode (μ_i, Σ_i and ν_i), as well as γ_{ij}, σ_{H_i} and ξ_i (unless considered as a tuning parameter) in the time-varying transition probabilities ($1 \leq i, j \leq M$, M has been defined previously).

Graphical illustration of the proposed model in this chapter is presented in Figure 4.1. This figure presents the graphical representation of the model at two consecutive sample times k and $k + 1$. As explained in Equations 4.3 and 4.4, in each operating mode I_k, the model assigns the appropriate scalar weight R_k to downweight the effect of outliers. Furthermore, in the presence of an appropriate scheduling variable H_{k-1}, which is able to indicate the true operating mode with an acceptable degree of uncertainty, following Equations 4.6 to 4.8, the time varying Markov chain structure enables the model to adapt to new operating modes. Combination of the robustness and time varying properties provides a general framework for diagnosis of the current operating mode I_k in industrial applications.

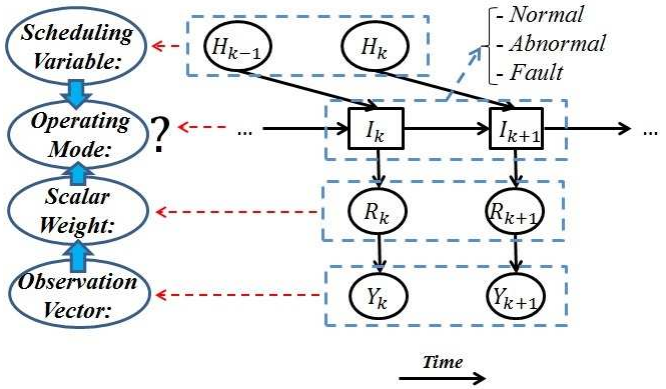

Figure 4.1: Graphical illustration of the proposed model in this chapter

The ultimate goal of this chapter is to diagnose the current operating mode of the process given the observations up to the current time. A filtering algorithm is used for this purpose. The algorithm is tested on simulation and experimental data.

4.3 Robust Parameter Estimation

In this section, the procedure for robust parameter estimation following the EM algorithm framework is explained. The EM algorithm maximizes the expected value of the complete-data log likelihood function, also known as the Q-function, iteratively. Iterations are repeated until some convergence criterion is satisfied [28]. This convergence criterion is selected to be less than a given likelihood change (absolute value) in two successive iterations.

In the context of generalized EM algorithm, it has been proved that if the value of Q-function in each iteration is greater than or equal to the previous iteration, the optimization is convergent and will converge to at least a local optimum [29]. The proposed procedure of this articles satisfies such a need. With appropriate initialization techniques, results of this iterative procedure can converge to the true values [34]. The initialization technique of this article is similar to the procedure introduced in the previous chapter where the initial values are obtained based on an initial solution assuming that observations follow a mixture of Gaussian distributions.

The expectation step starts from the following definition of the Q-function:

$$Q(\theta \mid \theta^{old}) = E_{C_{miss}|(\theta^{old}, C_{obs})}\{log\ P(C_{obs}, C_{miss} \mid \theta)\} \tag{4.9}$$

where different terms of this equation have been described previously. The superscript '*old*' here refers to the parameters estimated in the previous iteration.

Equation 4.9 can be further written as,

$$Q(\theta \mid \theta^{old}) = \tag{4.10}$$

$$E_{I_{1:N},R_{1:N}|(\theta^{old},C_{obs})}\{log\ P(Y_{1:N}, H_{1:N}, I_{1:N}, R_{1:N} \mid \theta)\}$$

The chain rule of probability is used to decompose the density function in Equation 4.10 as follows:

$$P(Y_{1:N}, H_{1:N}, I_{1:N}, R_{1:N} \mid \theta) = \tag{4.11}$$

$$P(Y_{1:N} \mid H_{1:N}, I_{1:N}, R_{1:N}, \theta)P(R_{1:N} \mid H_{1:N}, I_{1:N}, \theta)\times$$

$$P(I_{1:N} \mid H_{1:N}, \theta)P(H_{1:N} \mid \theta)$$

Each term of Equation 4.11 can be further simplified:

$$P(Y_{1:N} \mid H_{1:N}, I_{1:N}, R_{1:N}, \theta) = \tag{4.12}$$

$$\prod_{k=1}^{N} P(Y_k \mid Y_{k-1}, ..., Y_1, H_{1:N}, I_{1:N}, R_{1:N}, \theta) = \prod_{k=1}^{N} P(Y_k|I_k, R_k, \theta)$$

This equation is obtained based on the assumption that each observation vector at time k, i.e., Y_k, follows the conditional distribution shown in Equation 4.3. Also, given the mode identity (I_k) and weight (R_k), the scheduling variable does not provide any further information on the conditional distribution of Y_k .

The second term in Equation 4.11 can be decomposed in a similar manner,

$$P(R_{1:N} \mid H_{1:N}, I_{1:N}, \theta) = \prod_{k=1}^{N} P(R_k \mid R_{k-1:1}, H_{1:N}, I_{1:N}, \theta) \tag{4.13}$$

$$= \prod_{k=1}^{N} P(R_k \mid I_k, \theta)$$

where Equation 4.13 is obtained based on the assumption that data quality is only dependent on the operating mode of the process; for example, more abnormal data might be generated while the process is getting closer to the faulty modes.

The third term in Equation 4.11 can be written as,

$$P(I_{1:N} \mid H_{1:N}, \theta) = \prod_{k=1}^{N} P(I_k \mid I_{k-1}, ..., I_1, H_{1:N}, \theta) = \tag{4.14}$$

$$P(I_1)\prod_{k=2}^{N} P(I_k \mid I_{k-1}, H_{k-1}, \theta)$$

where this equation is obtained based on the Markov property, and the assumptions used to develop the distributions in Equation 4.6 to 4.8.

The scheduling variables are independent of the unknown parameters (θ), and consequently, the last term in Equation 4.11 can be considered as a constant value, that is,

$$P(H_{1:N} \mid \theta) = Const \tag{4.15}$$

Having Equation 4.9 to 4.15, and using the properties of the *log* operator, the Q-function can be written as,

$$Q(\theta \mid \theta^{old}) = E_{I_{1:N},R_{1:N}|(\theta^{old},C_{obs})}\{\sum_{k=1}^{N} logP(Y_k|I_k,R_k,\theta)+ \tag{4.16}$$

$$\sum_{k=1}^{N} logP(R_k|I_k,\theta) + \sum_{k=2}^{N} logP(I_k \mid I_{k-1}, H_{k-1}, \theta) + logP(I_1) + log\ (Const)\}$$

First, the expected value with respect to the hidden operating modes is obtained, that is,

$$Q(\theta \mid \theta^{old}) = E_{R_{1:N}|(\theta^{old},C_{obs},I)} \tag{4.17}$$

$$\{\sum_{i=1}^{M}\sum_{k=1}^{N} P(I_k = i|\theta^{old}, C_{obs})logP(Y_k|I_k = i, R_k, \theta) + \sum_{i=1}^{M}\sum_{k=1}^{N} P(I_k = i|\theta^{old}, C_{obs})logP(R_k|I_k = i, \theta)$$

$$+\sum_{i=1}^{M}\sum_{j=1}^{M}\sum_{k=2}^{N} P(I_k = j, I_{k-1} = i|\theta^{old}, C_{obs})log\alpha_{ij}(k) + \sum_{i=1}^{M} P(I_1 = i|\theta^{old}, C_{obs})log\pi_i + log\ (Const)\}$$

where π_i is the initial state distribution of the Markov chain model.

In order to find the expected value with respect to the weights $R_{1:N}$, first, distributions of $logP(Y_k|I_k, R_k, \theta)$ and $logP(R_k|I_k, \theta)$ are obtained based on Equations 4.3 and 4.4 as follows:

$$logP(Y_k|I_k, R_k, \theta) = \tag{4.18}$$

$$-\frac{P}{2}log(2\pi) - \frac{1}{2}log|\Sigma_{I_k}| + \frac{P}{2}logR_k - \frac{1}{2}R_k(Y_k - \mu_{I_k})^T\Sigma_{I_k}^{-1}(Y_k - \mu_{I_k})$$

$$logP(R_k|I_k,\theta) = \tag{4.19}$$

$$-log\Gamma(\frac{1}{2}\nu_{I_k}) + \frac{1}{2}\nu_{I_k}log(\frac{1}{2}\nu_{I_k}) + \frac{1}{2}\nu_{I_k}(logR_k - R_k) - logR_k$$

It can be observed that Equation 4.18 and 4.19 are linear with respect to the hidden variables R_k and $logR_k$. Therefore, in order to find the expected value in (4.17), it is only required to find $E(R_k|I_k = i, Y_k, H_k, \theta^{old})$ and $E(logR_k|I_k = i, Y_k, H_k, \theta^{old})$ [29]. Since the wight R_k given each operating mode follows the *gamma* distribution in Equation 4.4, and the observation vector Y_k given the operating mode and the weight follows the normal distribution in Equation 4.3, knowing the fact that *gamma* distribution is a conjugate prior for Gaussian likelihood, it can be concluded that $R_k|I_k = i, Y_k, H_k, \theta^{old}$ should also follow a gamma distribution with parameters given as [29]

$$R_k|I_k = i, Y_k, H_k, \theta^{old} \sim gamma(\frac{\nu_i^{old} + P}{2}, \frac{\nu_i^{old} + \delta(Y_k|\mu_i^{old}, \Sigma_i^{old})}{2}) \tag{4.20}$$

Therefore, according to the properties of the *gamma* distribution, $E(R_k|I_k = i, Y_k, H_k, \theta^{old})$ and $E(logR_k|I_k = i, Y_k, H_k, \theta^{old})$ are obtained as follows [29]:

$$E(R_k|I_k = i, Y_k, H_k, \theta^{old}) = \frac{\nu_i^{old} + P}{\nu_i^{old} + \delta(Y_k|\mu_i^{old}, \Sigma_i^{old})} = r_{ik}^{old} \tag{4.21}$$

$$E(logR_k|I_k = i, Y_k, H_k, \theta^{old}) = log(r_{ik}^{old}) + \{\psi(\frac{\nu_i^{old} + P}{2}) - log(\frac{\nu_i^{old} + P}{2})\} \tag{4.22}$$

where $\psi(\nu)$ in Equation 4.22 is the *Digamma* function defined as [29]

$$\psi(\nu) = \frac{\frac{\partial\Gamma(\nu)}{\partial\nu}}{\Gamma(\nu)} \tag{4.23}$$

In order to further simplify Equation 4.17, using the Bayesian formulation, the posterior distributions of $P(I_k = j, I_{k-1} = i|\theta^{old}, C_{obs})$ and $P(I_k = i|\theta^{old}, C_{obs})$ can be obtained as follows:

$$P(I_k = j, I_{k-1} = i \mid \theta^{old}, C_{obs}) =$$

$$P(I_k = j, I_{k-1} = i \mid Y_1, ..., Y_N, \theta^{old}, H_1, ..., H_N) =$$

$$\frac{P(Y_k \mid \mu_j^{old}, \Sigma_j^{old}, \nu_j^{old}) \times \alpha(k)_{ij}^{old} \times P(I_{k-1} = i \mid \theta^{old}, H_1, ..., H_{k-2})}{\sum_{i=1}^{M} \sum_{j=1}^{M} P(Y_k \mid \mu_j^{old}, \Sigma_j^{old}, \nu_j^{old}) \times \alpha(k)_{ij}^{old} \times P(I_{k-1} = i \mid \theta^{old}, H_1, ..., H_{k-2})} = \tau_{ijk}^{old} \tag{4.24}$$

The posterior distribution of $P(I_k = i|\theta^{old}, C_{obs})$ can be obtained through the marginalization of Equation 4.24 over all the possible modes for I_{k-1}, that is,

$$P(I_k = i \mid \theta^{old}, C_{obs}) = \tag{4.25}$$

$$\sum_{j=1}^{M} P(I_k = i, I_{k-1} = j \mid \theta^{old}, C_{obs}) = \tau_{ik}^{old}$$

Having Equations 4.18 to 4.25, Equation 4.17 can be written as

$$Q(\theta \mid \theta^{old}) = \sum_{i=1}^{M} \sum_{k=1}^{N} \tau_{ik}^{old} Q_1(\mu_i, \Sigma_i) + \sum_{i=1}^{M} \sum_{k=1}^{N} \tau_{ik}^{old} Q_2(\nu_i) \tag{4.26}$$

$$+ \sum_{i=1}^{M} \sum_{j=1}^{M} \sum_{k=2}^{N} \tau_{ijk}^{old} Q_3(\gamma_{ij}, \sigma_{H_i}) + \sum_{i=1}^{M} \tau_{i1}^{old} Q_4(\pi_i) + log\ (Const)$$

where in this equation,

$$Q_1(\mu_i, \Sigma_i) = -\frac{P}{2} log(2\pi) - \frac{1}{2} log|\Sigma_i|$$

$$+\frac{P}{2}(log(r_{ik}^{old}) + \{\psi(\frac{\nu_i^{old} + P}{2}) - log(\frac{\nu_i^{old} + P}{2})\}) - \frac{1}{2} r_{ik}^{old} (Y_k - \mu_i)^T \Sigma_i^{-1} (Y_k - \mu_i)$$

$$Q_2(\nu_i) = -log\Gamma(\frac{1}{2}\nu_i) + \frac{1}{2}\nu_i log(\frac{1}{2}\nu_i)$$

$$+(\frac{1}{2}\nu_i - 1)(log(r_{ik}^{old}) + \{\psi(\frac{\nu_i^{old} + P}{2}) - log(\frac{\nu_i^{old} + P}{2})\}) - \frac{1}{2}\nu_i r_{ik}^{old}$$

$$Q_3(\gamma_{ij}, \sigma_{H_i}) = log\alpha_{ij}(k)$$

$$Q_4(\pi_i) = log\pi_i$$

In order to update the parameters under the EM framework, the Q-function, as obtained in Equation 4.26, should be maximized. This is achieved by taking derivative of Equation 4.26 with respect to the unknown parameters, and then setting it to zero. Therefore, the mean vector (μ_i) should be updated as follows:

$$\frac{\partial Q(\theta \mid \theta^{old})}{\partial \mu_i} = \frac{\partial \sum_{i=1}^{M} \sum_{k=1}^{N} \tau_{ik}^{old} Q_1(\mu_i, \Sigma_i)}{\partial \mu_i} = 0$$

$$\Rightarrow \mu_i^{new} = \frac{\sum_{k=1}^{N} \tau_{ik}^{old} r_{ik}^{old} Y_k}{\sum_{k=1}^{N} \tau_{ik}^{old} r_{ik}^{old}} \tag{4.27}$$

The covariance matrix Σ_i should be updated in a similar manner, that is,

$$\frac{\partial Q(\theta \mid \theta^{old})}{\partial \Sigma_i} = \frac{\partial \sum_{i=1}^{M} \sum_{k=1}^{N} \tau_{ik}^{old} Q_1(\mu_i, \Sigma_i)}{\partial \Sigma_i} = 0$$

$$\Rightarrow \Sigma_i^{new} = \frac{\sum_{k=1}^{N} \tau_{ik}^{old} r_{ik}^{old} (Y_k - \mu_i^{new})(Y_k - \mu_i^{new})^T}{\sum_{k=1}^{N} \tau_{ik}^{old}} \tag{4.28}$$

Maximization of the Q-function to find π_i is constrained by $\sum_{i=1}^{M} \pi_i = 1$ according to the properties of the Markov chain model. Consequently, the Lagrange multiplier η should be introduced:

$$\pi_i^{new} = argmax_{\pi_i} \{ \sum_{i=1}^{M} P(I_1 = i \mid \theta^{old}, C_{obs}) log \pi_i + \eta (\sum_{i=1}^{M} \pi_i - 1) \} \tag{4.29}$$

Taking the derivative with respect to the Lagrange multiplier (η) and π_i a set of linear equations will be obtained. As the solution of this set of equations, the following expression for the initial state distribution of the Markov chain can be obtained:

$$\pi_i^{new} = P(I_1 = i \mid \theta^{old}, C_{obs}) \tag{4.30}$$

Similarly, for the Markov chain property of the transition probabilities to hold ($\sum_{j=1}^{M} \alpha_{ij}(k) = 1$), the constraint $\sum_{j=1}^{M} \gamma_{ij,i \neq j} = 1$ should be satisfied. Thus, for maximization of the Q-function with respect to $\gamma_{ij,i \neq j}$, the Lagrange multiplier η' should be introduced:

$$\gamma_{ij,i \neq j}^{new} = argmax_{\gamma_{ij,i \neq j}} \{ \sum_{i=1}^{M} \sum_{j \neq i=1}^{M} \sum_{k=2}^{N} \tau_{ijk}^{old} \times log(\alpha_{ij}(k)) + \eta' (\sum_{j \neq i=1}^{M} \gamma_{ij} - 1) \} \tag{4.31}$$

Taking the derivative with respect to $\gamma_{ij,i \neq j}$ and η', as the solution of a linear set of equations, $\gamma_{ij,i \neq j}$ is obtained as follows:

$$(\gamma_{ij,i \neq j})^{new} = \frac{\sum_{k=2}^{N} \tau_{ijk}^{old}}{\sum_{j \neq i=1}^{M} \sum_{k=2}^{N} \tau_{ijk}^{old}} \tag{4.32}$$

Due to the appearance of the exponential function in $\alpha_{ij}(k)$, no closed form solution exists for σ_{H_i}. Thus, a non-linear constraint optimization problem should be solved to obtain this parameter as follows:

$$(\sigma_{H_i})_{1\leq i\leq M}^{new} = argmax_{\sigma_{H_i}} \sum_{i=1}^{M}\sum_{j=1}^{M}\sum_{k=2}^{N} \tau_{ijk}^{old} \times log(\alpha_{ij}(k)) \tag{4.33}$$

$$S.t.\ \sigma_{H_{min}} \leq \sigma_{H_i} \leq \sigma_{H_{max}}$$

As explained in the problem statement, for processes with fast transitions between operating modes, ξ_i in the definition of $\alpha(k)_{ii}$ should be considered as an unknown parameter (not a tuning parameter), and estimated through the iterations of the EM algorithm. In this case, the more general form of Equation 4.33, which is already introduced in the previous chapter should be used. However, as in the second example of this chapter, appropriate methods for the initialization of the nonlinear optimization should be considered and computation is more involved [45].

In this study, based on the quality of data in each mode, different degrees of freedom are automatically assigned to the t distribution of the corresponding mode. This will provide more flexibility in the modelling since, normally, it is expected to have more outlier data while the process gets closer to the abnormal or faulty modes. The optimal value of the degree of freedom for each mode (ν_i) is updated by taking the derivative of the Q-function with respect to ν_i, and then, setting it to zero, that is,

$$\frac{\partial Q(\theta \mid \theta^{old})}{\partial \nu_i} = \frac{\partial \sum_{i=1}^{M}\sum_{k=1}^{N} \tau_{ik}^{old} Q_2(\nu_i)}{\partial \nu_i} = 0 \tag{4.34}$$

Having $Q_2(\nu_i)$ as introduced in Equation 4.26, and based on Equation 4.34, it is not difficult to show that the following expression should hold [82]:

$$-\psi(\nu_i/2) + log(\nu_i/2) + 1 + \{\psi(\frac{\nu_i^{old}+P}{2}) - log(\frac{\nu_i^{old}+P}{2})\} \tag{4.35}$$

$$+\frac{1}{\sum_{k=1}^{N}\tau_{ik}^{old}} \sum_{k=1}^{N} \tau_{ik}^{old}(logr_{ik}^{old} - r_{ik}^{old}) = 0$$

where ψ is the *Digamma* function introduced in Equation 4.23. Solution of this nonlinear equation using appropriate numerical techniques will provide the appropriate degree of freedom according to the data quality.

Having outliers in the training data set will usually result in lower degrees of freedom. During the parameter estimation, effect of the degree of freedom assigned

to each operating mode will appear in r_{ik}^{old} in Equation 4.21. Accordingly, due to the larger Mahalanobis distance ($\delta(Y_k|\mu,\Sigma) = (Y_k - \mu)\Sigma^{-1}(Y_k - \mu)$), and based on Equation 4.21, r_{ik}^{old} decreases, and therefore, based on Equations 4.27 and 4.28, the estimated parameters will be less affected by outliers [82].

4.4 Operating Condition Diagnosis

The proposed process monitoring strategy introduced in this article consists of two steps:

1. Parameter estimation based on the solutions provided in Equations 4.27 to 4.35.

2. Operating condition diagnosis based on the estimated parameters ($\hat{\Theta}$) and Hamilton's filtering algorithm [44].

The purpose of the second step is to find the probability of current operating mode at time k given observations and the estimated parameters, that is, $P(I_k \mid Y_k, ..., Y_1, H_k, ..., H_1, \hat{\Theta})$. Solutions to this filtering problem in the presence of the scheduling variable is available in Section 3.4 of the previous chapter [45]. In this study, the same algorithm is adopted. The difference is the updating step, where in this study, distribution of observations given the operating mode at time k, that is, $P(Y_k|I_k)$, follows multivariate Student t as expressed in Equation 4.2, with parameters obtained by following the parameter estimation procedure in the earlier section. One could consider this filtering algorithm as an optimal estimator to infer discrete HMM modes. There are other similar optimal algorithms for such a problem in literature [40]. The Viterbi algorithm is a well known example of such methods [40]. In comparison to the proposed method of this article, the Viterbi algorithm finds an optimal sequence of states. Consequently, dynamic programming and additional computational cost are inevitable.

4.5 Results and Discussion

Unlike the previous studies, in this chapter, it is assumed that data follow various multivariate t distributions for different operating modes. Therefore, in the following examples, using a time varying Markov chain structure, the comparison is made between the Gaussian ([45]) and t distribution. In all the examples, 2/3 of the data is used for training purposes, and the rest for validation. Due to the limited space, only cross-validation results are presented. In both examples, the estimated operating

mode is obtained as the solution of the filtering problem introduced in Section 4.4, that is, the mode with the largest probability of observations up to current time is selected as the current operating condition.

4.5.1 Tennessee Eastman Process

In this example, monitoring of the overall temperature in the Tennessee Eastman (TE) process is considered. The simulation example is selected from the study on decentralized control of the product rate in the TE process [87]. A schematic of the process is presented in Figure 4.2 ([87]).

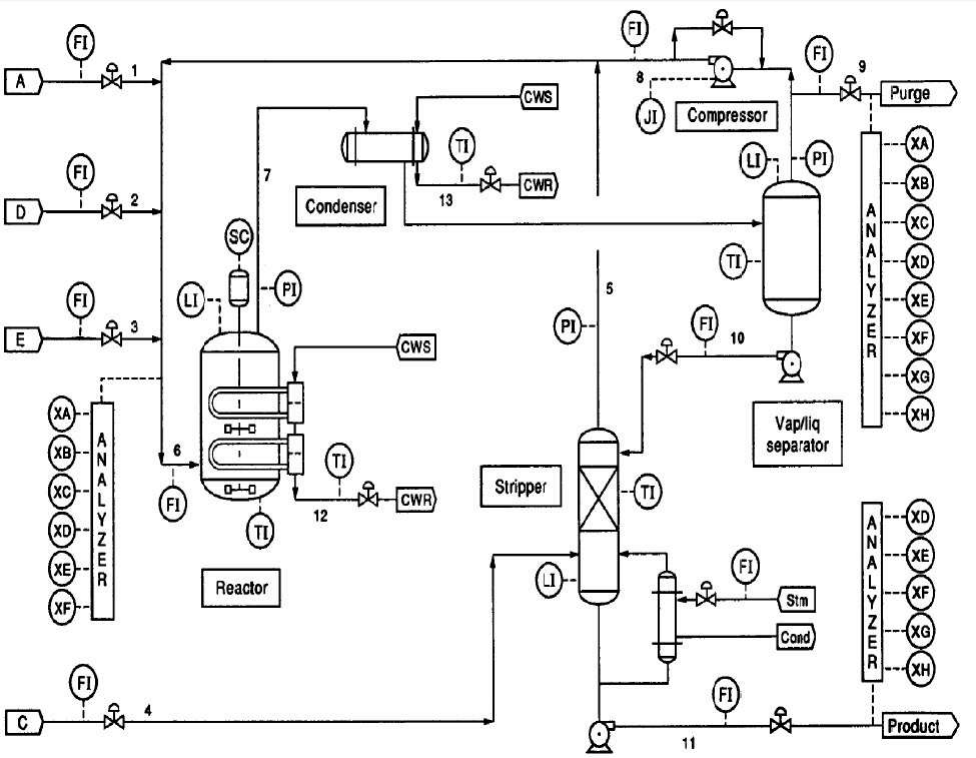

Figure 4.2: TE Process [87]

TE process is composed of four main components: 1- Exothermic two-phase reactor, 2- Flash separator, 3- Reboiled Stripper and 4- Recycle compressor. Six measurements should be maintained at the desired set points including, 1- Production rate, 2- Mole % G in product, 3- Reactor pressure, 4- Reactor liquid level, 5- Separator liquid level, and 6- Stripper liquid level [87].

In this example, it is assumed that some disturbances occur in the reactor temperature which is located in one of the earliest stages of the process. The effect of this disturbance is then propagated through the top stream of the reactor to the flash

separator, and then, the reboiled stripper. This causes three operating modes including mode 1 (low temperature), mode 2 (average temperature) and mode 3 (high temperature). The temperatures of the two latter components are measured through some noisy signals in the presence of outliers. Separator coolant temperature, which can provide some intermediate information between these two components, is selected as the scheduling variable. 10 % of observations in all operating modes are outliers, that is, 10 % of the generated data are replaced with ouliers, randomly selected from a uniform distribution beyond three standard deviations of the mean value of the variable. For the scheduling variable, the noise is smoothed from measurements. Therefore, the scheduling variable provides an indirect indication (with uncertainty) of the process operating mode. Cross-validation data and the scheduling variable are presented in Figure 4.3 and Figure 4.4 respectively. The total run time of this experiment is 440 hours, i.e., each 10,000 time steps in Figure 4.3 to Figure 4.9 correspond to $\frac{10,000}{40,000} \times 440 = 100$ hours.

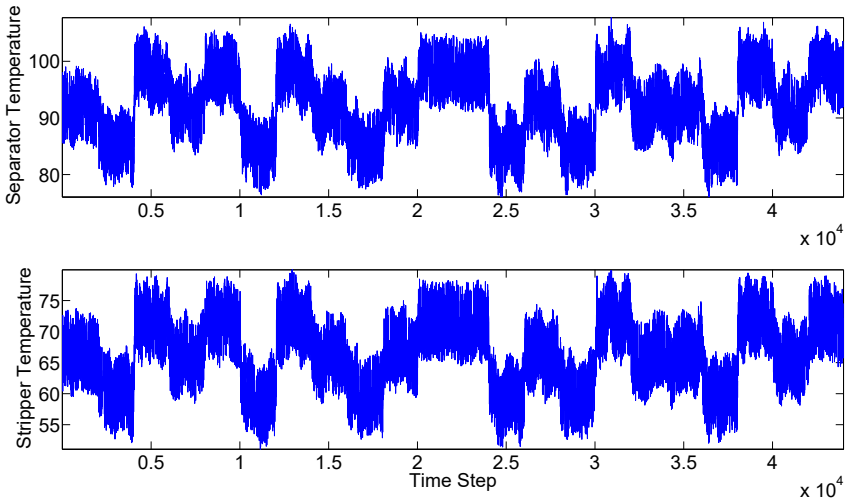

Figure 4.3: Measured observations for monitoring of the TE process

Results of the filtering problem to find the probability of mode identities up to the current time, assuming that observations follow different multivariate Gaussian distributions, are presented in Figure 4.5. Estimated and true operating modes of the process based on these probabilities are presented in Figure 4.6.

Based on these results, it is clear that a large number of false alarms have occurred in mode 2, which has a similar behavior as both modes 1 and 3, in many time intervals. It appears that parameter estimation is biased by the data of the faulty mode.

Results of the filtering for estimating modes based on the t distribution are pre-

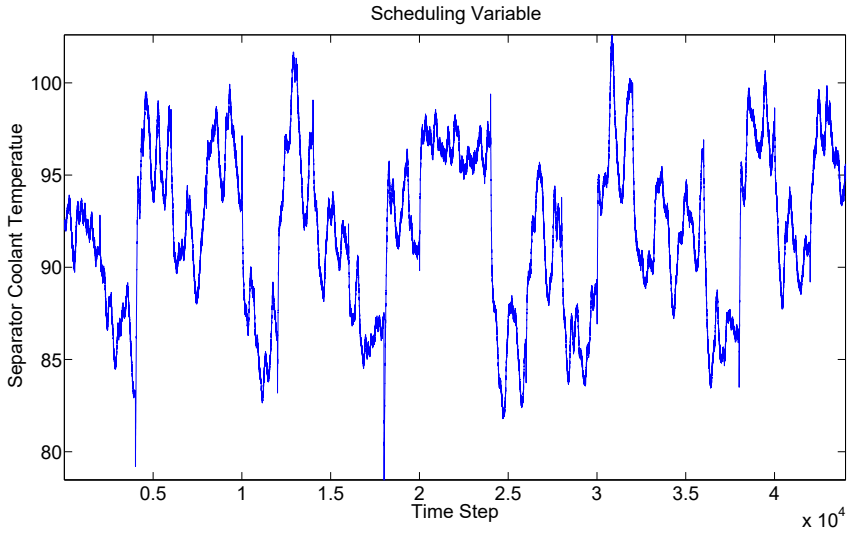

Figure 4.4: Scheduling variable for monitoring of the TE process

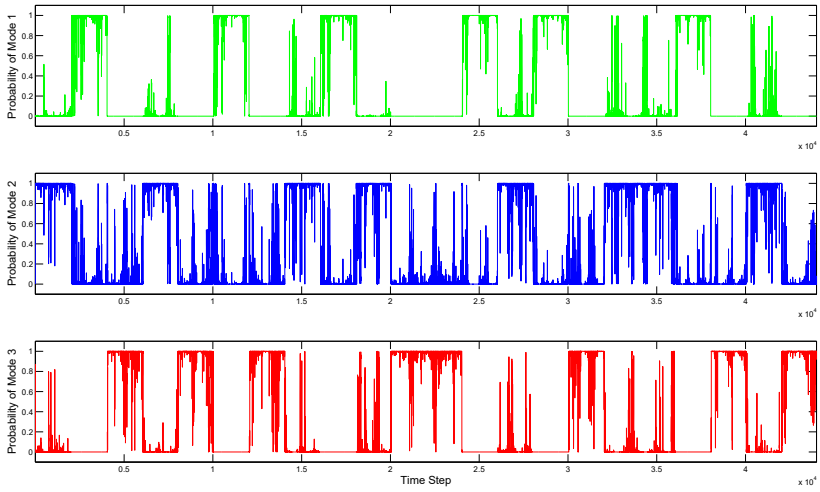

Figure 4.5: Probability of the observations in Figure 4.3 given each operating mode based on the Gaussian distribution assumption

sented in Figures 4.7 and 4.8 respectively. According to the results, it is obvious that the number of mis-classifications, specifically in the intermediate mode, have been greatly reduced. In general, the percentage of false alarms has been decreased from 4.1613 % to 4.0567 %. The effectiveness of the method will be more clearly demonstrated in the next example when there are more outliers in the data. This clearly demonstrates the advantage in considering the t distribution with adaptive degrees of freedom according to the data quality.

In order to show the importance of the time varying transition probabilities as

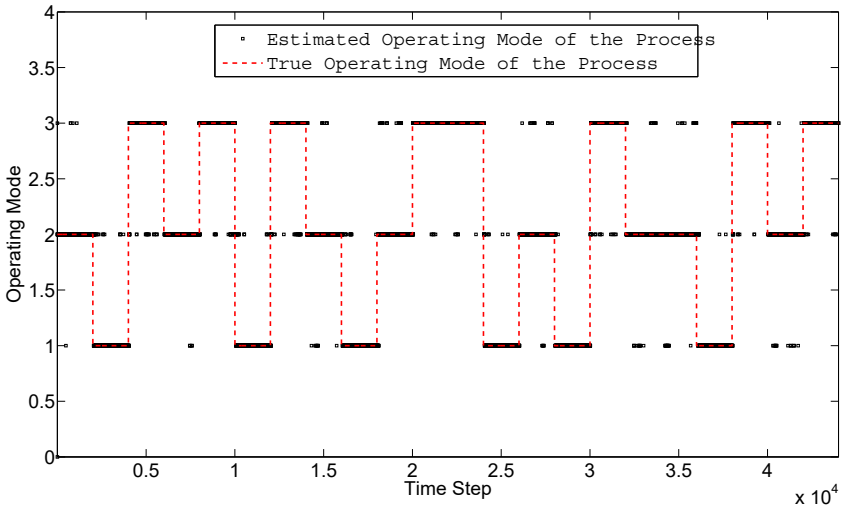

Figure 4.6: Estimated and true operating modes of the process based on the observations in Figure 4.3 and the Gaussian distribution assumption

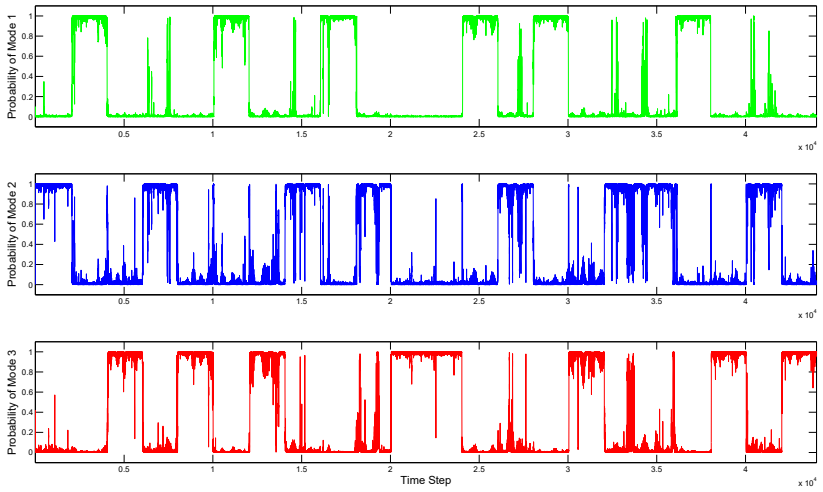

Figure 4.7: Probability of the observations in Figure 4.3 given each operating mode based on the t distribution assumption

proposed in Equations 4.6 to 4.8, variations of the baseline transition probabilities a_{11}, a_{22} and a_{33} are presented in Figure 4.9. Comparing Figure 4.3 and Figure 4.9, it is clear that when approaching a new operating mode i, according to Equation 4.7, probability of staying in that mode (a_{ii}) increases, and based on Equation 4.8, probability of transiting to other modes (a_{ij}) will decrease. Consequently, this flexible structure appropriately considers the time varying nature of industrial processes.

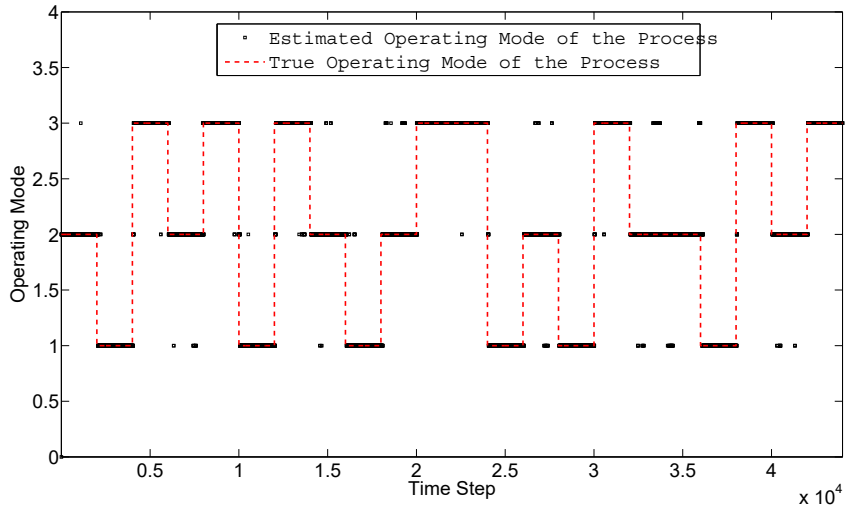

Figure 4.8: Estimated and true operating modes of the process based on the observations in Fig. 4.3 and t distribution assumption

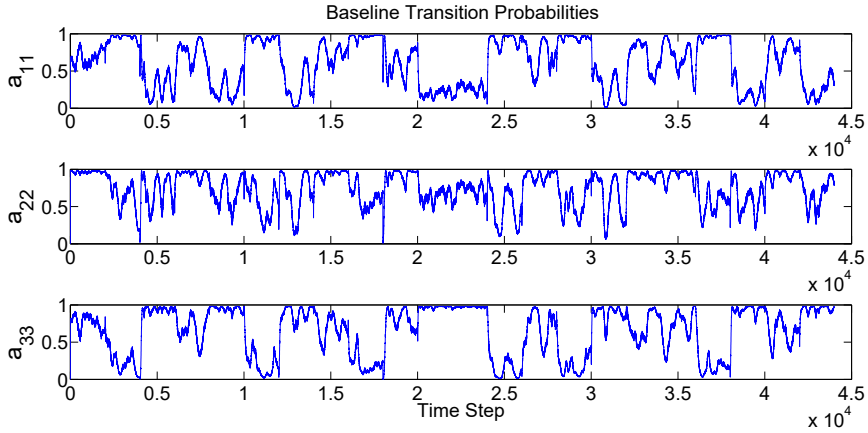

Figure 4.9: Variations of the baseline transition probabilities for the presented results in Fig. 4.3

To observe the effect of outliers on the estimation of the degree of freedom, the percentage of outliers in various operating modes is gradually increased and the corresponding degree of freedom is estimated automatically. Results are presented in Figure 4.10.

These results confirm a fact which was introduced in literature as the "breakdown" point [82]. As expected, by increasing the percentage of outliers in different operating modes, the degree of freedom decreases to downweight the effect of outliers, that is, in Equation 4.21, for the data with a larger Mahalanobis distance, the decrease in the degree of freedom (ν_i^{old}) will result in a decrease in the weight r_{ik}^{old}. Consequently,

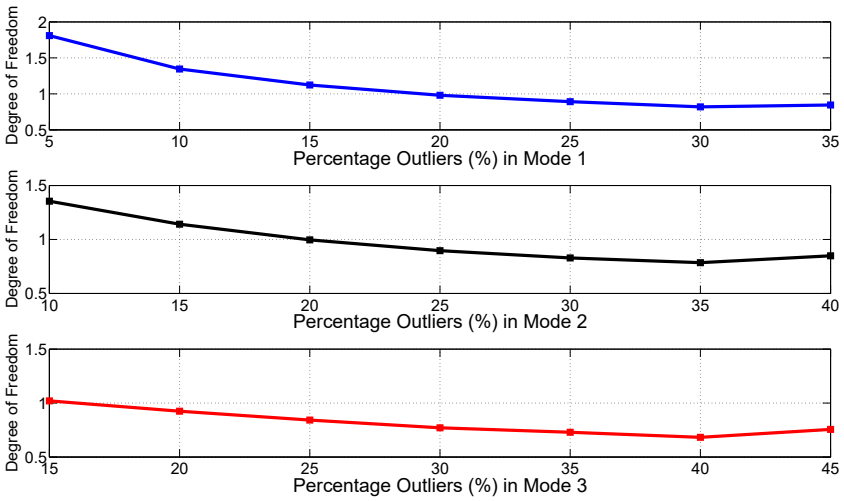

Figure 4.10: Variations of the degree of freedom while increasing the percentage of outliers in the TE process

according to Equations 4.27 and 4.28, the outliers will have less effect on parameter estimation. However, further increase in the percentage of outliers will cause the t distribution to have very heavier tails, and finally, at some point, the t distribution starts to behave as a uniform distribution in order to adapt to very poor quality data. At the breakdown point, which occurs right before this extreme case, the degree of freedom will no longer show the decreasing correlation with the percentage of outliers, and unless Equation 4.35 is solved with appropriate constraints with respect to ν_i (greater than zero), the solution might not be numerically tractable anymore. Our experience in this study shows that the exact breakdown point might vary for different processes according to the mean and variance of the data and the type of outliers.

4.5.2 Experimental Evaluation

In this section, the proposed process monitoring strategy is applied on a real life experiment. The balls in tubes experiment, which is a well known experimental set up in literature [88], is used here. The schematic of the experiment is presented in Figure 4.11. In this set up, a constant air volume is provided for the process through the main fan. The air is blown to four tubes through the lower manifold. Four balls are located in four tubes whose heights are measured through ultrasonic sensors. Each tube has a separate DC fan which controls the height of the ball through a separate PID controller. The remaining air in the process is disposed to environment through the upper and lower manifold outputs.

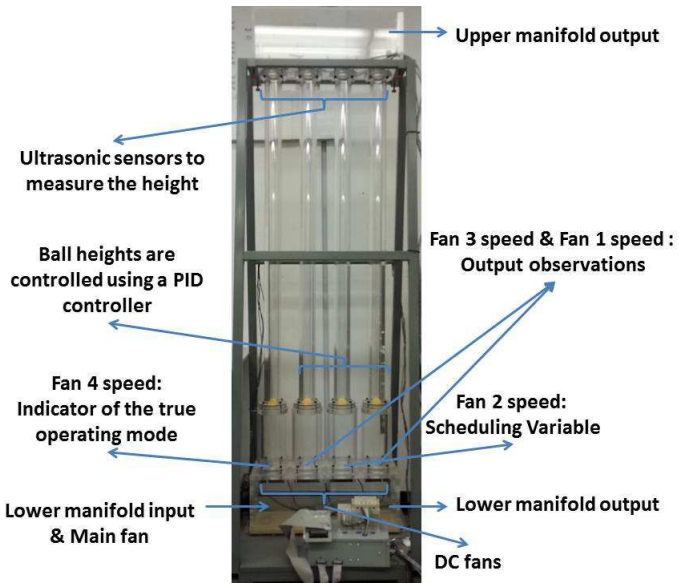

Figure 4.11: Balls in tubes experiment

In order to constantly maintain the ball heights at the same level, the PID controllers impose a lower limit constraint on the available air flow in the lower manifold. Since no measurement device exists to measure the available air flow in the lower manifold, the goal of this process monitoring experiment is to infer the mode of the available air flow for the controllers through the current measurements. Mode 1 is the normal operation of the process where sufficient air is available for the controllers, and mode 2 is the abnormal operation where the air flow suddenly decreases due to the disturbances in the process.

Fan 4 speed, which is the only fan not in closed loop along with its corresponding ball, will change between a maximum and minimum value thus, providing less, or more air (modes 2 and 1 respectively) to the other 3 fans. Therefore, fan 4 speed gives an indication (with uncertainty) of the true operating mode of the process. Fans 1 to 3 are in closed loop with the controlled ball heights. When fan 4 speed is set to its maximum value, less air will be provided for fans 1 to 3. Ideally, speed of fans 1 to 3 should increase accordingly to maintain the ball heights at the previous level in the presence of less air. However, fan 3, which is located right after fan 4, shows a reverse behavior. This is due to the sudden effect of fan 4 on fan 3. Fans 1 and 2, which are located farther, behave more normally as expected. Figure 4.12 illustrates the cross-validation data. The sample time of this experiment is one second.

According to the previous discussion, the speeds of fans 1 and 3 are the observations for the mode diagnosis purposes. Fan speed 2, which provides some intermediate information between process observations, is chosen as the scheduling variable. Noise

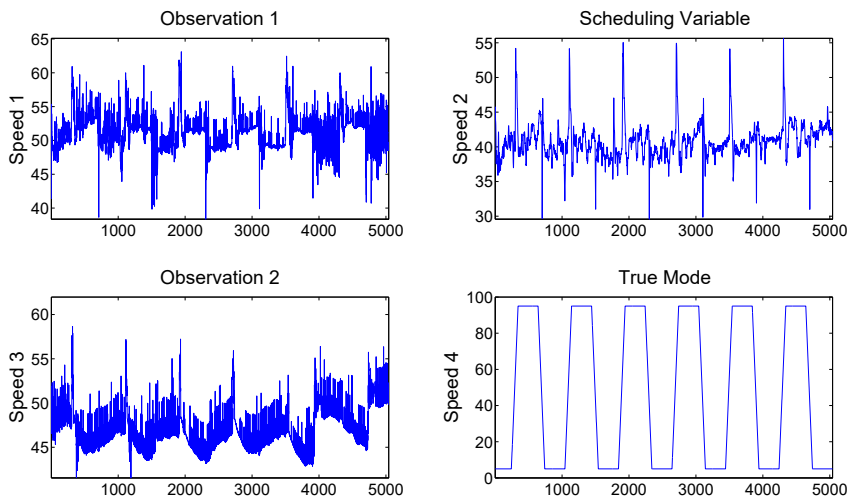

Figure 4.12: Changes in the fan speeds 1 to 3 after the change in the fan speed 4

and outlier data are filtered from the scheduling variable to provide a better indication to the operating condition of the process. It should be noted that poor selection of the scheduling variable will result in a larger local mode validity (σ_{H_i}) in Equation 4.7. Consequently, $\alpha(k)_{ii} \simeq \xi_i$, and, if ξ_i is considered as an unknown during the parameter estimation, the negative effect of the poor scheduling variable will be automatically considered in the estimation of ξ_i. Results of the filtering and operating condition diagnosis based on the Gaussian distribution assumption are presented in Figure 4.13 and Figure 4.14 respectively.

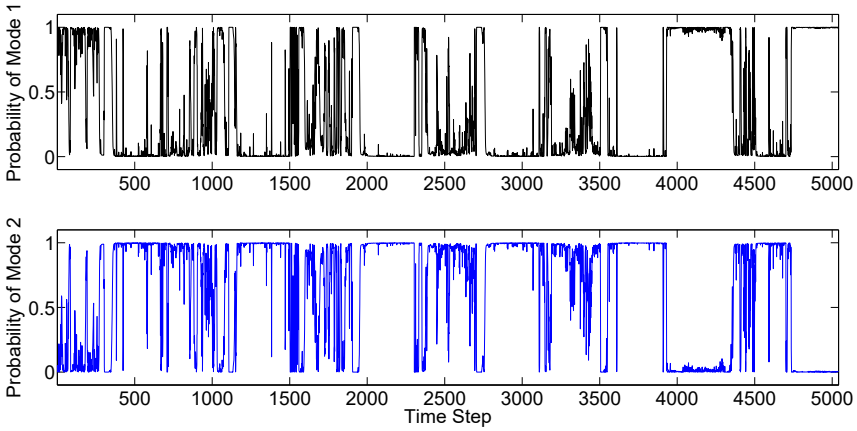

Figure 4.13: Probability of the observations in Figure 4.12 given each operating mode based on the Gaussian distribution assumption

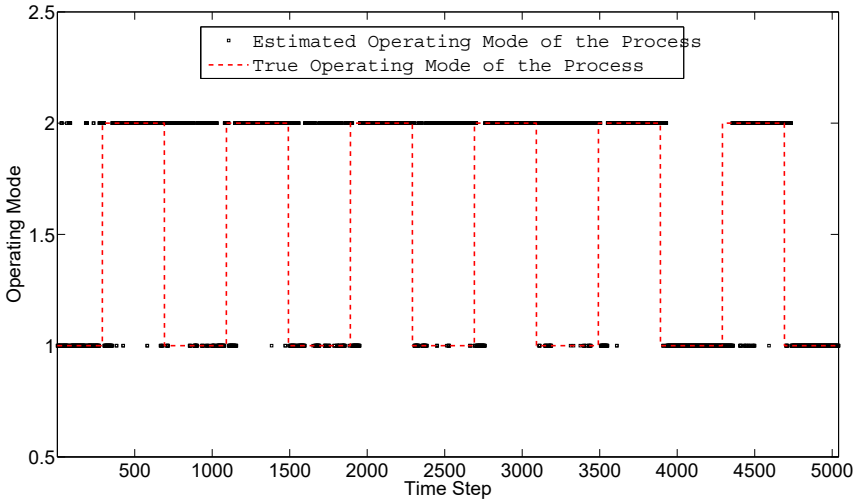

Figure 4.14: Estimated and true operating modes of the process based on the observations in Figure 4.12 and the Gaussian distribution assumption

According to the results, it is clear that the Gaussian assumption fails to provide good estimations of the true operating mode of the process in many instants. This is a good example to show the significant merit of the t over Gaussian distribution. Since a lot of low quality data are generated due to the poor measurements of the ultrasonic sensors, distributions with heavier tails are required to describe the observations. t distribution better suits for such conditions. Results of the probability calculation and mode diagnosis based on the t distribution assumption are presented in Figure 4.15 and Figure 4.16.

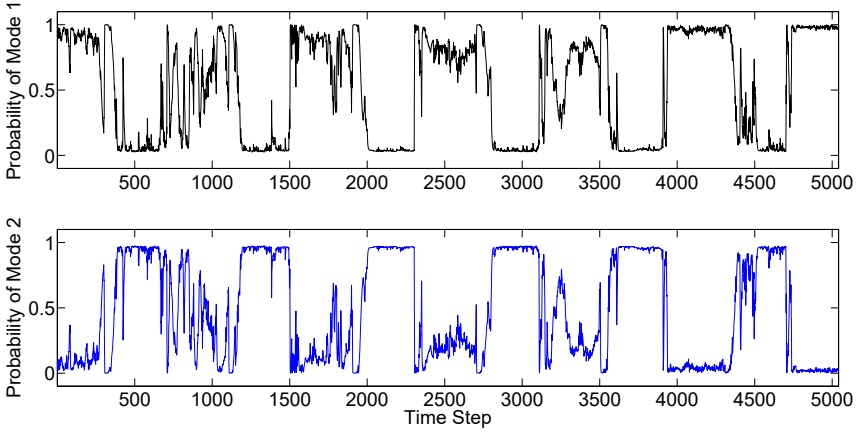

Figure 4.15: Probability of the observations in Figure 4.12 given each operating mode based on the t distribution assumption

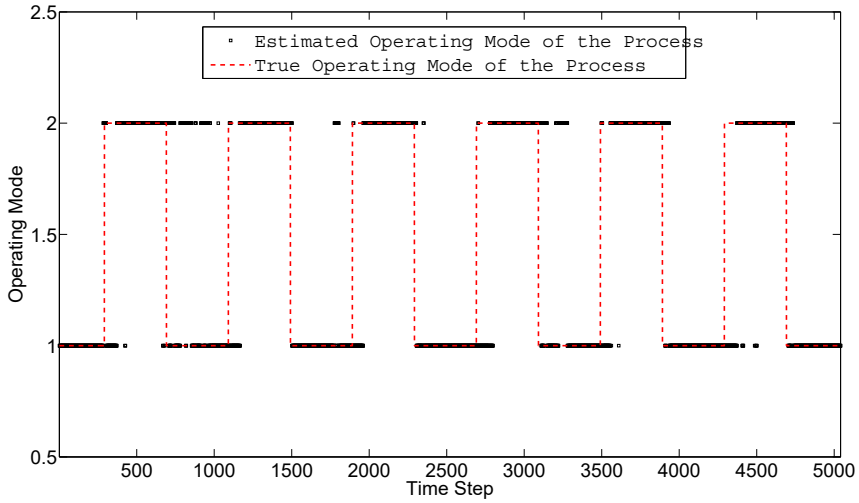

Figure 4.16: Estimated and true operating modes of the process based on the observations in Figure 4.12 and t distribution assumption

These results indicate a good potential of the proposed identification and diagnostic method in the presence of outliers. Using a multivariate t distribution other than the Gaussian, percentage of false alarms has been decreased almost by half from 33.4457 % to 16.3856 %. According to the observation data set (Figure 4.12), it is clear that more outliers have appeared in the case of more air flow (mode 1). As illustrated in Figure 4.16, consideration of the t distribution for the observations, which results in a lower degree of freedom for mode 1 ($\nu_1 = 4.2652$, and $\nu_2 = 10.7467$), provides a more accurate diagnosis in general.

4.6 Conclusion

In this chapter, a novel approach for robust mode diagnosis of time varying processes is proposed. Application of adaptive transition probabilities provides more flexibility to consider the time varying behaviour of the process, while adopting the t distribution for observations makes the model adaptive to the data quality. In the presence of low quality data, which is more likely to occur in the abnormal and faulty modes, t distribution provides heavier tails and better classifications by automatically adopting the appropriate degree of freedom resulting in a more robust diagnosis.

The proposed method is tested on both simulation and experimental examples, and in the presence of low quality data, demonstrates a better performance over other techniques from literature.

Chapter 5

Adaptive Monitoring of the Process Operation Based on Symbolic Episode Representation and Hidden Markov Models

Unlike Chapters 3 and 4 where continuous signals are used for process monitoring purposes, in this chapter, a combination of Qualitative Trend Analysis (QTA) to discretize the observations, and HMMs, is used for data classification. Accordingly, our focus will be on key features of the signals rather than the details. First, continuous time signals are converted to discrete observations using the method of triangular representation. Since there is a large difference in the means and variances of the durations and magnitudes of the triangles at different operating modes, adaptive fuzzy membership functions are applied for discretization. The expectation maximization (EM) algorithm is used to obtain parameters of the different modes for the durations and magnitudes assuming that states transit to each other according to a Markov chain model. Applying Hamilton's filter, probability of each state given new duration and magnitude is calculated to weight the membership functions of each mode previously obtained from a fuzzy C-means clustering. After adaptive discretization step, having discrete observations available, the combinatorial method for training hidden Markov models (HMMs) with multiple observations is used for overall classi-

A version of this chapter has been published in N. Sammaknejad, B. Huang, A. Fatehi, Y. Miao, F. Xu and A. Espejo (2014). Adaptive Monitoring of the Process Operation Based on Symbolic Episode Representation and Hidden Markov Models with Application Toward an Oil Sand Primary Separation. Computers and Chemical Engineering 71(4), 281–297 [71].

Section 5.7 of this chapter has been published in N. Sammaknejad, B. Huang (2014). Process Monitoring Based on Symbolic Episode Representation and Hidden Markov Models - A Moving Window Approach. Proceedings of the 5^{th} International Symposium on Advanced Control of Industrial Processes (ADCONIP). Hiroshima, Japan [95].

fication of the process. Furthermore, a search algorithm is proposed to find the more informative observations of a window of recent data. Application of the method is studied on both simulation and industrial case studies. The industrial case study is the detection of normal and abnormal process conditions in the primary separation vessel (PSV) of an oil sand industry. The method shows an overall good performance in detecting normal and risky operating conditions.

5.1 Introduction

In order to achieve safe production and decrease manufacturing cost, fault diagnosis along with process monitoring is becoming increasingly important. There are three main approaches for process monitoring: the knowledge based approach, the model based approach and the data driven approach. The knowledge based approach is based on qualitative models. The model based approach is based on analytical models which are complex for large systems. The data driven framework is appropriate for large multivariate processes [89]. Most of the approaches based on qualitative models use pattern recognition techniques to extract features from historical process data, e.g., signal directed graphs, fault trees, fuzzy systems, neural networks or qualitative trend analysis [90].

Wong et al. introduced a strategy for detection of abnormal trends using important process features and qualitative information from a signal [19]. They use the method of triangular representation, initially developed by Cheung, Stephanopoulos and Bakshi ([91, 92, 93]) in order to discretize the continuous time observation sequences using appropriate fuzzy membership functions and rules. In their study, first, the high frequency noise is removed using wavelet analysis. Second, continuous time observations are converted to discrete numbers using the method of triangular representation and appropriate fuzzy membership functions and rules. In the overall decision making step, each variable is classified based on its corresponding HMM and the overall classification is based on a Back Propagation Neural Network (BPNN) which uses the generated probabilities of each HMM as the input [20].

The main disadvantage of the method of triangular representation is the loss of information when providing symbolic observations. Fixed fuzzy membership functions might provide imprecise classifications for the modes with smaller means and variances as the membership functions are biased by the modes with larger means and variances.

The proposed adaptive fuzzification method in this chapter will provide more accurate discrete observations considering different modes for the durations and mag-

nitudes of the triangles. Applying the EM algorithm, we will first divide the historical data of durations and magnitudes to different modes assuming that states can transit to each other at probabilities that obey Markov property. Fuzzy membership functions for each mode are obtained following a Fuzzy C-Means (FCM) clustering approach. When a new observation for the magnitude and duration is available, the probability of the observation given each mode is calculated using Hamiltons filter [46]. These probabilities are then used as weights to combine means and variances of the membership functions of the different modes. Finally, using the adaptive membership functions at each time step and the method of triangular representation, the discrete observations are generated. Having the discrete observations available, a multivariate HMM approach is adopted for overall classification of the process [94].

It is shown that using a multivariate scheme to train HMMs for discrete observations provides better results in comparison to the BPNN approach [95]. The multivariate scheme reduces the computational time, considers the interactions between different inputs and reduces the number of false alarms. Combination of adaptive fuzzification to discretize the continuous observations and multivariate HMM modeling shows a good performance in detection of normal and faulty operating conditions in both simulation and industrial case studies. The industrial case study is selected as abnormal operating condition diagnosis in the primary separation vessel (PSV) of an oil sand industry.

In Section 7, an optimal search algorithm is proposed to find the more informative observations of a recent window of data. The algorithm is developed for the case of fixed fuzzy membership functions. Similar algorithms can be considered for the case of adaptive fuzzification in future studies.

Figure 5.1 is a summary of the proposed process monitoring strategy in this chapter where state recognition, adaptive fuzzification and multivariate HMM modeling are added to the previous studies. The procedure of state recognition and adaptive fuzzification is presented in Figure 5.2. The multivariate HMM modeling step is adopted from literature [94].

The remainder of this chapter is organized as follows: Sections 5.2 and 5.3 are a review on the data pre-processing based on wavelet analysis and the method of triangular representation. Section 5.4 reviews state recognition applying the EM algorithm and Hamiltons filter. In Section 5.5, the procedure of adaptive fuzzification is explained. Section 5.6 briefly reviews the multivariate scheme adopted here to train HMMs for multiple observation sequences. Section 5.7 is the proposed moving window approach to find the more informative observations of a window. Section 5.8 is the simulation case study. Section 5.9 is the industrial case study, and Section 5.10

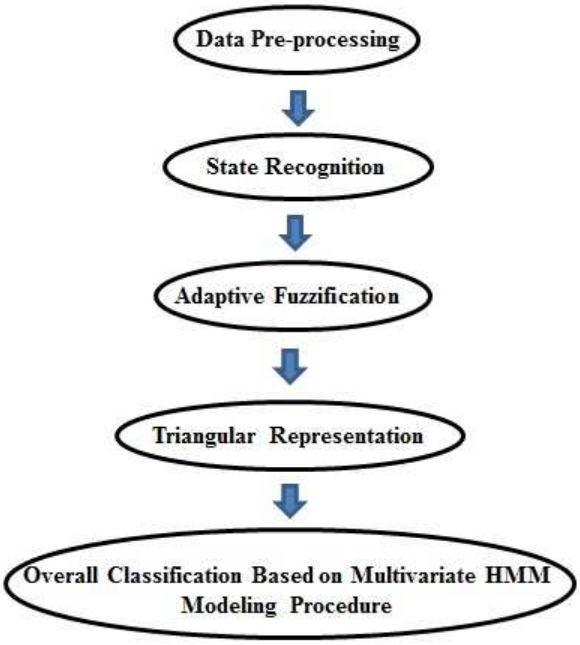

Figure 5.1: The proposed process monitoring approach in this study

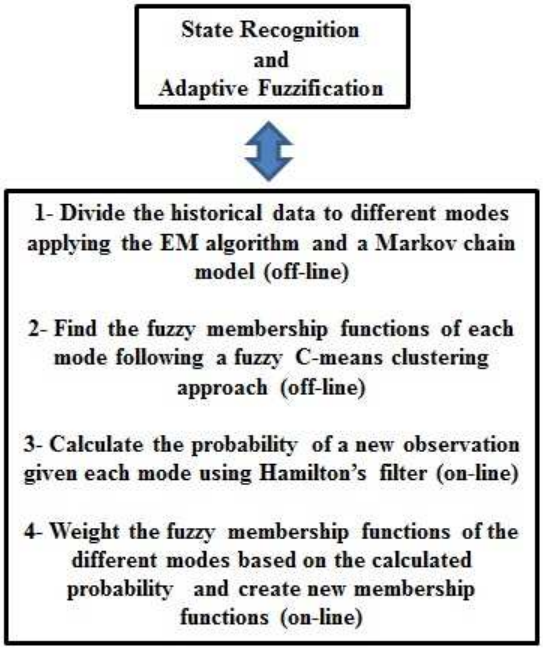

Figure 5.2: The procedure of adaptive fuzzification and state recognition

concludes the chapter.

5.2 Data Preprocessing

The first step in representation and classification of a signal is data filtering. Typically, a filter removes nuisance information over a range of frequencies, which is determined by filter parameter. Several methods are available in order to filter a signal, e.g., moving average, Gaussian filter, Fourier transform or wavelet analysis. Compared with moving average, Gaussian filtering and Fourier analysis, the wavelet analysis possesses excellent time-frequency properties since it uses a time-scale region. Therefore, using wavelet analysis, we can get a multi-scale description of trends and features, which enables us to analyze the data efficiently. Using wavelet analysis, the original signal emerges as two signals, the low frequency part of the signal which is called approximation and the high frequency part of the signal which is called the detail. This process is called decomposition in wavelet and can be performed iteratively [96].

5.3 Triangular Representation

Cheung and Stephanopoulos treat the problem of trend representation graphically by using the simple idea that at the extrema and inflection points, the first and second derivatives are zero respectively. Thus, an episode is described as any part of a signal or process trend with a constant sign of the first and second derivative. This leads to the set of triangles and lines that are defined using seven letters of the alphabet ([91, 92, 93]).

Some of the advantages of this feature extraction method are as follows:

1. it converts a signal into a symbolic sequence, which captures the most important qualitative and quantitative information contained in the signal.

2. Compared with filtered process data, the symbolic form of observations is appropriate as the input to a classifying system such as hidden Markov models. In situations where model-based approaches cannot help in fault detection, i.e., the complexity of the system does not allow one to derive a model for the normal operation of the system, the proposed pattern recognition method can greatly help in reducing the complexity of the problem.

3. Since this approach tries to capture the trends rather than exact quantities of data, it is necessary to remove high frequency noises before feature extraction. Therefore, this approach is less sensitive to noise.

The main disadvantage of the method of triangular representation is the loss of information through the discretization step.

The procedure for triangular representation of a trend is as follows [93]:

After smoothing a signal for extraction of basic trends, according to its extrema and inflection points, the signal will be divided into episodes. As illustrated in Figure 5.3, an episode consists of an extremum and a neighbored inflection point, making it a triangle. Each triangle is made of vertices found from first or second order zero crossing where the sign of the first and second derivatives are remained constant in this segment.

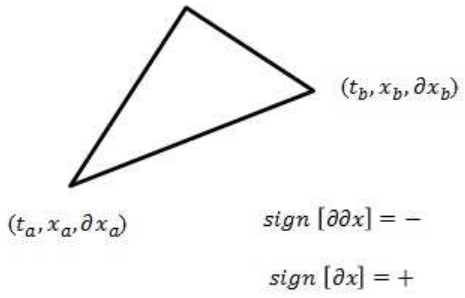

Figure 5.3: A sample episode for describing process trends [19]

Based on the definition of an episode, seven kinds of triangles named as A, B, C, D, E, F and G can be defined as presented in Figure 5.4. Types E-G are three kinds of line in a smoothed trend and can be substituted with other types of triangles. Therefore, the triangular representation method in this chapter is simplified to only contain four types of triangle: A-D.

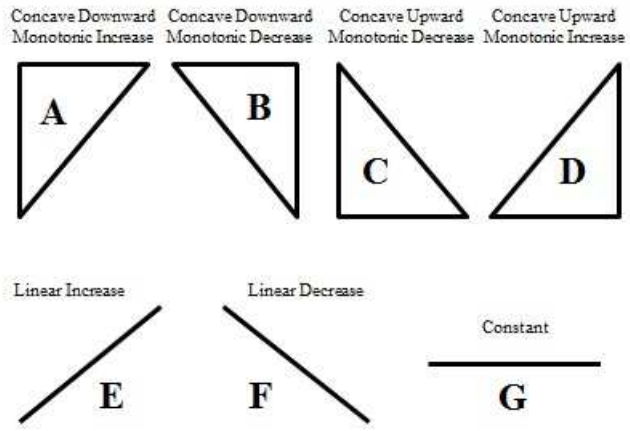

Figure 5.4: Seven types of triangles [19]

Fuzzy classification is the process of grouping elements into a fuzzy set using appropriate fuzzy rules and membership functions [97]. As presented in Figure 5.5, the quantitative values of magnitude and duration of a triangular episode can be transferred into completely symbolic variables. Having three membership functions of large, medium and small for magnitude, and three membership functions of long, middle and short for duration of every type of triangle in Figure 5.4, there are nine possible outcomes using fuzzy classification as depicted in Figure 5.5. This ends to $4 \times 9 = 36$ discrete observations in total. For example, lmA stands for a large magnitude, middle duration, type "A" triangle.

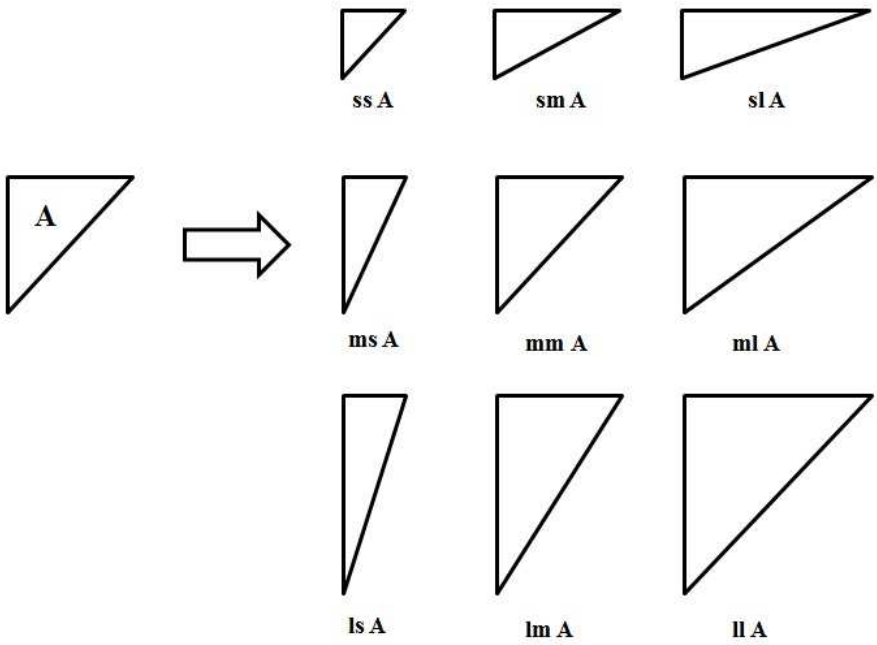

Figure 5.5: 'A' Triangle transformed to 9 sub-types using appropriate fuzzy rules and membership functions [19]

5.4 State Recognition

A drawback of using fixed membership functions for fuzzy classification of all the durations and magnitudes (m and d) is that they are dominated with modes (states) with larger means and variances. Therefore, they provide imprecise classifications for the modes (states) with average means and variances since most of the observations will be incorrectly categorized as small. In this chapter, the fuzzy membership functions for symbolic representation are adapted according to the state of the durations and magnitudes of the signal. States are assumed to transit to each other following a

Markov chain model and observations follow the Gaussian distributions in Equation 5.5.

The probabilistic framework constructed for this state recognition is explained in this section. Suppose that the duration (d_t) and magnitude (m_t) for the new observation is defined as:

$$d_t = x_t - x_{t-1} \tag{5.1}$$

$$m_t = y_t - y_{t-1}$$

where x corresponds to variations of the signal in time direction and y is the value of the signal.

The probability of each state given the new observation is calculated using Hamiltons filter [46]. The input of the filter is the conditional probability $P(I_{t-1}|m_{t-1}, d_{t-1}, ..., m_0, d_0)$ and has the output $P(I_t|m_t, d_t, ..., m_0, d_0)$, where I_t is an indicator for the mode (state) of the observations at time t. The general form of Hamilton's filter is modified for this specific problem assuming that observations (m_t and d_t) follow normal distributions and are independent of each other given the states (to be described mathematically in Equation 5.5).

The output of the filter is obtained as:

$$P(I_t|m_t, d_t, ..., m_0, d_0) = \sum_{I_{t-1}=1}^{M} P(I_t, I_{t-1}|m_t, d_t, ..., m_0, d_0) \tag{5.2}$$

where M is the number of states.

The joint probability of the states I_t, I_{t-1} given the information up to time $t-1$ is calculated as:

$$P(I_t, I_{t-1}|m_{t-1}, d_{t-1}, ..., m_0, d_0) = P(I_t|I_{t-1})P(I_{t-1}|m_{t-1}, d_{t-1}, ..., m_0, d_0) \tag{5.3}$$

$P(I_t|I_{t-1})$ is the Markov model transition probability (α_{ij}) and $P(I_{t-1}|m_{t-1}, d_{t-1}, ..., m_0, d_0)$ is known from the input to the filter. This joint probability of the states I_t and I_{t-1} is further updated using the new observations at time t:

$$P(I_t, I_{t-1}|m_t, d_t, ..., m_0, d_0) = \frac{P(m_t, d_t, I_t, I_{t-1}|m_{t-1}, d_{t-1}, ..., m_0, d_0)}{P(m_t, d_t|m_{t-1}, d_{t-1}, ..., m_0, d_0)} \tag{5.4}$$

where

$$P(m_t, d_t, I_t, I_{t-1}|m_{t-1}, d_{t-1}, ..., m_0, d_0) = f(m_t, d_t|I_t, I_{t-1}, m_{t-1}, d_{t-1}, ..., m_0, d_0)$$

$$\times P(I_t, I_{t-1}|m_{t-1}, d_{t-1}, ..., m_0, d_0)$$

where $P(I_t, I_{t-1}|m_{t-1}, d_{t-1}, ..., m_0, d_0)$ is obtained previously from Equation 5.3, and $f(m_t, d_t|I_t, I_{t-1}, m_{t-1}, d_{t-1}, ..., m_0, d_0) = f(m_t, d_t|I_t)$ is calculated assuming that observations (m_t and d_t) follow the normal distributions in Equation 5.5 and are independent of each other given the states, and finally,

$$P(m_t, d_t|m_{t-1}, d_{t-1}, ..., m_0, d_0) = \sum_{I_t=1}^{M} \sum_{I_{t-1}=1}^{M} P(m_t, d_t, I_t, I_{t-1}|m_{t-1}, d_{t-1}, ..., m_0, d_0)$$

The above procedure is based on the probability of the observation and Markov model of state transitions. The parameters of these probability functions $(\mu_m^i, \sigma_m^i, \mu_d^i, \sigma_d^i, \alpha_{ij}, \pi_i)$ can be calculated using the Expectation Maximization (EM) algorithm. Details are presented in the following subsection.

5.4.1 Parameter Estimation Based on the EM Algorithm

The unknown parameters could be estimated as the solution of a maximum likelihood estimation problem. However, due to a large number of unknowns and modes, regular optimization algorithms could not provide solutions directly. The EM algorithm is an appropriate alternative for such situations and provides closed form solutions for the unknown parameters. Parameter estimation for regime switching systems under the EM framework has been of interest since 1990 [44, 51, 55]. In this chapter we will derive a mathematical frame-work for regime switching systems when multiple independent observations are available.

It is assumed that observations of the durations and magnitudes ($\{m_t\}_{t=1}^N$and $\{d_t\}_{t=1}^N$) are independent of each other and identically distributed (i.i.d.) given the states (Equation 5.5):

$$(m_t|I_t = i; \mu_m^i, \sigma_m^i) \sim N(\mu_m^i, (\sigma_m^i)^2),\ i = 1, ..., M \tag{5.5}$$

$$(d_t|I_t = i; \mu_d^i, \sigma_d^i) \sim N(\mu_d^i, (\sigma_d^i)^2),\ i = 1, ..., M$$

where $\mu_{d,m}^i$ and $\sigma_{d,m}^i$ are means and variance of the different states.

Later, means and variances of the fuzzy membership functions in each mode ($\sigma_{d^i,m^i}^{sml,\ ave,\ lrg}$ and $\mu_{d^i,m^i}^{sml,\ ave,\ lrg}$) will be separately calculated based on the observations of the mode and the FCM approach. It will be further discussed in the next section.

In order to have probabilistic transitions between different states of the system,

it is assumed that they follow a Markov chain with transition probability (α_{ij}) and initial state probability (π_i) given as

$$\alpha_{ij} = P(I_t = j | I_{t-1} = i),\ i, j = 1, ..., M \ and \ t = 1, ..., N \tag{5.6}$$

$$\pi_i = P(I_1 = i),\ i = 1, ..., M$$

The EM algorithm finds the unknown parameters ($\mu_m^i, \sigma_m^i, \mu_d^i, \sigma_d^i, \alpha_{ij}, \pi_i$) by iterating between the E (expectation) and M (maximization) steps [55]. In the E-step of the EM algorithm, the conditional expectation of the complete data (known as the Q-function) is calculated:

$$Q(\Theta|\Theta^{old}) = E_{I|\Theta^{old}, C_{obs}}\{log\ f(C_{obs}, I|\Theta)\} \tag{5.7}$$

where Θ^{old} is the vector of parameters for the previous iteration, I is the unknown (hidden) state, C_{obs} is the vector of observations ($\{m_t, d_t\}_{t=1}^N$), f is the probability distribution function and Θ is the set of unknown parameters.

In the M-step, the set of parameters that maximizes the Q-function, are calculated:

$$\Theta^{new} = max_\Theta\ Q(\Theta|\Theta^{old}) \tag{5.8}$$

For the problem of this chapter, the E-step can be formulated as,

$$Q(\Theta|\Theta^{old}) = E_{I|(\Theta^{old}, C_{obs})}\{log\ f(m_N, ..., m_1, d_N, ..., d_1, I_N, ..., I_1|\Theta)\} \tag{5.9}$$

$$= E_{I|(\Theta^{old}, C_{obs})}\{log \prod_{t=1}^{N} f(m_t, d_t, I_t | m_{t-1}, ..., m_1, d_{t-1}, ..., d_1, I_{t-1}, ..., I_1, \Theta)\}$$

$$= E_{I|(\Theta^{old}, C_{obs})}\{log \prod_{t=1}^{N} f(m_t|I_t, \Theta) f(d_t|I_t, \Theta) P(I_t|I_{t-1}, \Theta)\}$$

$$= E_{I|(\Theta^{old}, C_{obs})}\{log(P_{I_1})\} + E_{I|(\Theta^{old}, C_{obs})}\{\sum_{t=1}^{N} [logf(m_t|I_t, \Theta) + logf(d_t|I_t, \Theta)]\}$$

$$+E_{I|(\Theta^{old}, C_{obs})}\{logP(I_t|I_{t-1}, \Theta)\}$$

$$= \sum_{i=1}^{M} P(I_1 = i|\Theta^{old}, C_{obs}) log\pi_i$$

$$+\sum_{i=1}^{M}\sum_{t=1}^{N} P(I_t = i|\Theta^{old}, C_{obs}) \left[logf(m_t|I_t = i, \mu_m^{(i)}, \sigma_m^{(i)}) + logf(d_t = i|I_t, \mu_d^{(i)}, \sigma_d^{(i)}) \right]$$

$$\sum_{i=1}^{M}\sum_{j=1}^{M}\sum_{t=2}^{N} P(I_t = j, I_{t-1} = i|\Theta^{old}, C_{obs}) log\alpha_{ij}$$

In these derivations we have used $f(m_t|m_{t-1}, ..., m_1, d_t, ..., d_1, I_t, I_{t-1}, ..., I_1, \Theta) = f(m_t|I_t, \Theta)$,
$f(d_t|m_{t-1}, ..., m_1, d_{t-1}, ..., d_1, I_t, I_{t-1}, ..., I_1, \Theta) = f(d_t|I_t, \Theta)$ and
$P(I_t|m_{t-1}, ..., m_1, d_{t-1}, ..., d_1, I_{t-1}, ..., I_1, \Theta) = P(I_t|I_{t-1}, \Theta)$ according to the Markov property and, implicitly, the assumption that observations (m_t and d_t) are independent of each other given the hidden state I_t.

In the M-step, derivatives of the Q-function are taken with respect to the unknown parameters:

$$\frac{\partial \sum_{i=1}^{M}\sum_{t=1}^{N} P(I_t = i|\Theta^{old}, C_{obs}) \left[log(\frac{1}{\sigma_m^{(i)}\sqrt{2\pi}})\, e^{\frac{-(m_t - \mu_m^{(i)})^2}{2(\sigma_m^{(i)})^2}} \right]}{\partial \mu_m^{(i)}} = 0 \qquad (5.10)$$

$$\frac{\partial \sum_{i=1}^{M}\sum_{t=1}^{N} P(I_t = i|\Theta^{old}, C_{obs}) \left[log(\frac{1}{\sigma_d^{(i)}\sqrt{2\pi}})\, e^{\frac{-(d_t - \mu_d^{(i)})^2}{2(\sigma_d^{(i)})^2}} \right]}{\partial \mu_d^{(i)}} = 0$$

Therefore, the mean value of the different modes for durations and magnitudes can be calculated as

$$(\mu_m^{(i)})^{New} = \frac{\sum_{t=1}^{N} m_t P(I_t = i|\Theta^{old}, C_{obs})}{\sum_{t=1}^{N} P(I_t = i|\Theta^{old}, C_{obs})} \qquad (5.11)$$

$$(\mu_d^{(i)})^{New} = \frac{\sum_{t=1}^{N} d_t P(I_t = i|\Theta^{old}, C_{obs})}{\sum_{t=1}^{N} P(I_t = i|\Theta^{old}, C_{obs})}$$

Variance of the different modes can also be calculated in a similar manner:

$$\frac{\partial \sum_{i=1}^{M}\sum_{t=1}^{N} P(I_t = i|\Theta^{old}, C_{obs}) \left[log(\frac{1}{\sigma_m^{(i)}\sqrt{2\pi}})\, e^{\frac{-(m_t - \mu_m^{(i)})^2}{2(\sigma_m^{(i)})^2}} \right]}{\partial \sigma_m^{(i)}} = 0 \qquad (5.12)$$

$$\frac{\partial \sum_{i=1}^{M}\sum_{t=1}^{N} P(I_t=i|\Theta^{old},C_{obs}) \left[log(\frac{1}{\sigma_d^{(i)}\sqrt{2\pi}})\ e^{\frac{-(d_t-\mu_d^{(i)})^2}{2(\sigma_d^{(i)})^2}} \right]}{\partial \sigma_d^{(i)}} = 0$$

Therefore:

$$((\sigma_m^{(i)})^{New})^2 = \frac{\sum_{t=1}^{N}(m_t-(\mu_m^{(i)})^{New})^2 P(I_t=i|\Theta^{old},C_{obs})}{\sum_{t=1}^{N} P(I_t=i|\Theta^{old},C_{obs})} \tag{5.13}$$

$$((\sigma_d^{(i)})^{New})^2 = \frac{\sum_{t=1}^{N}(d_t-(\mu_d^{(i)})^{New})^2 P(I_t=i|\Theta^{old},C_{obs})}{\sum_{t=1}^{N} P(I_t=i|\Theta^{old},C_{obs})}$$

The optimization problem to find α_{ij}, is constrained by $\sum_{j=1}^{M}\alpha_{ij}=1$ and as the result Lagrange multiplier λ is introduced:

$$\frac{\partial \left[\sum_{i=1}^{M}\sum_{j=1}^{M}\sum_{t=2}^{N} P(I_t=j,I_{t-1}=i|\Theta^{old},C_{obs}) log\alpha_{ij} + \lambda(\sum_{j=1}^{M}\alpha_{ij}-1)\right]}{\partial \alpha_{ij}} = 0 \tag{5.14}$$

Similarly, the optimization problem to find π_i, is constrained by $\sum_{j=1}^{M}\pi_i=1$ and as the result Lagrange multiplier λ is introduced:

$$\frac{\partial \left[\sum_{i=1}^{M} P(I_1=i|\Theta^{old},C_{obs}) log\pi_i + \lambda(\sum_{i=1}^{M}\pi_i-1)\right]}{\partial \pi_i} = 0 \tag{5.15}$$

Finally, the parameters of the Markov chain can be calculated as:

$$(\alpha_{ij})^{New} = \frac{\sum_{t=2}^{N} P(I_t=j,I_{t-1}=i|\Theta^{old},C_{obs})}{\sum_{j=1}^{M}\sum_{t=2}^{N} P(I_t=j,I_{t-1}=i|\Theta^{old},C_{obs})} \tag{5.16}$$

$$(\pi_i)^{New} = P(I_1=i|\Theta^{old},C_{obs})$$

The intermediate terms $P(I_t=j,I_{t-1}=i|\Theta^{old},C_{obs})$ and $P(I_t=i|\Theta^{old},C_{obs})$ in Equations 5.10 to 5.16 can be calculated according to Bayes rule and Markov property:

$$P(I_t=j,I_{t-1}=i|\Theta^{old},C_{obs}) = P(I_t=j,I_{t-1}=i|m_t,d_t,\Theta^{old},m_{t-1},...,m_1,d_{t-1},...,d_1) \tag{5.17}$$

$$= \frac{f(m_t|I_t=j,\Theta^{old})\ f(d_t|I_t=j,\Theta^{old})\ P(I_t=j|I_{t-1}=i,\Theta^{old})\ P(I_{t-1}=i|\Theta^{old})}{\sum_{i=1}^{M}\sum_{j=1}^{M} f(m_t|I_t=j,\Theta^{old})\ f(d_t|I_t=j,\Theta^{old})\ P(I_t=j|I_{t-1}=i,\Theta^{old})\ P(I_{t-1}=i|\Theta^{old})}$$

where in Equation 5.17, $P(I_t = j|I_{t-1} = i, \Theta^{old}) = (\alpha_{ij})^{old}$, $P(I_{t-1} = i|\Theta^{old})$ is obtained through the discrete-valued state propagation of Markov chain starting from the initial estimation of $P(I_1 = i|\Theta^{old}, C_{obs}) = (\pi_i)^{old}$, and $f(m_t|I_t = j, \Theta^{old})$ and $f(d_t|I_t = j, \Theta^{old})$ should be calculated based on the parameters in the previous iteration and Equation 5.5.

$P(I_t = i|\Theta^{old}, C_{obs})$ can be obtained from summation of $P(I_t = i, I_{t-1} = j|\Theta^{old}, C_{obs})$ over all the possible states for I_{t-1}:

$$P(I_t = i|\Theta^{old}, C_{obs}) = \sum_{j=1}^{M} P(I_t = i, I_{t-1} = j|\Theta^{old}, C_{obs}) \tag{5.18}$$

The initial values of the parameters are obtained assuming a mixture of Gaussian distributions for the magnitudes and durations and calculating the transition probabilities from occupation times in Equation 5.19.

$$\alpha_{ij} = \frac{C_{ij}}{C_i},\ i, j = 1, ..., M \tag{5.19}$$

where C_i is the number of times that the sequence is observed to be in states i and C_{ij} is the number of transitions from state i to state j.

5.5 Adaptive Fuzzification

As stated in Section 5.4, a drawback of using fixed membership functions for fuzzy classification is that they are dominated by modes with larger means and variances. Following an adaptive fuzzification procedure, we are looking for more precise discrete observations. A summary of adaptive fuzzification procedure introduced in this chapter is schematically presented in Figure 5.6.

As it is illustrated in Figure 5.6, in order to have more precise symbolic observations (triangles), the magnitudes and durations are divided to different modes based on their means and variances. Parameters of the fuzzy membership functions for each mode ($\sigma_{d^i,m^i}^{sml,\ ave,\ lrg}$ and $\mu_{d^i,m^i}^{sml,\ ave,\ lrg}$) are calculated based on a Fuzzy C-Means clustering (FCM) approach [98]. When a new observation for magnitude and duration is received, parameters of the membership functions for the different modes are weighted as a function of the posterior probability of each state given the new observation (Equations 5.20 and 5.21). The procedure to calculate this posterior probability is previously explained in Section 5.4. The final discretization step is based on these new membership functions.

$$w_i(t) = P(I_t = i|m_t, d_t, ..., m_0, d_0) \tag{5.20}$$

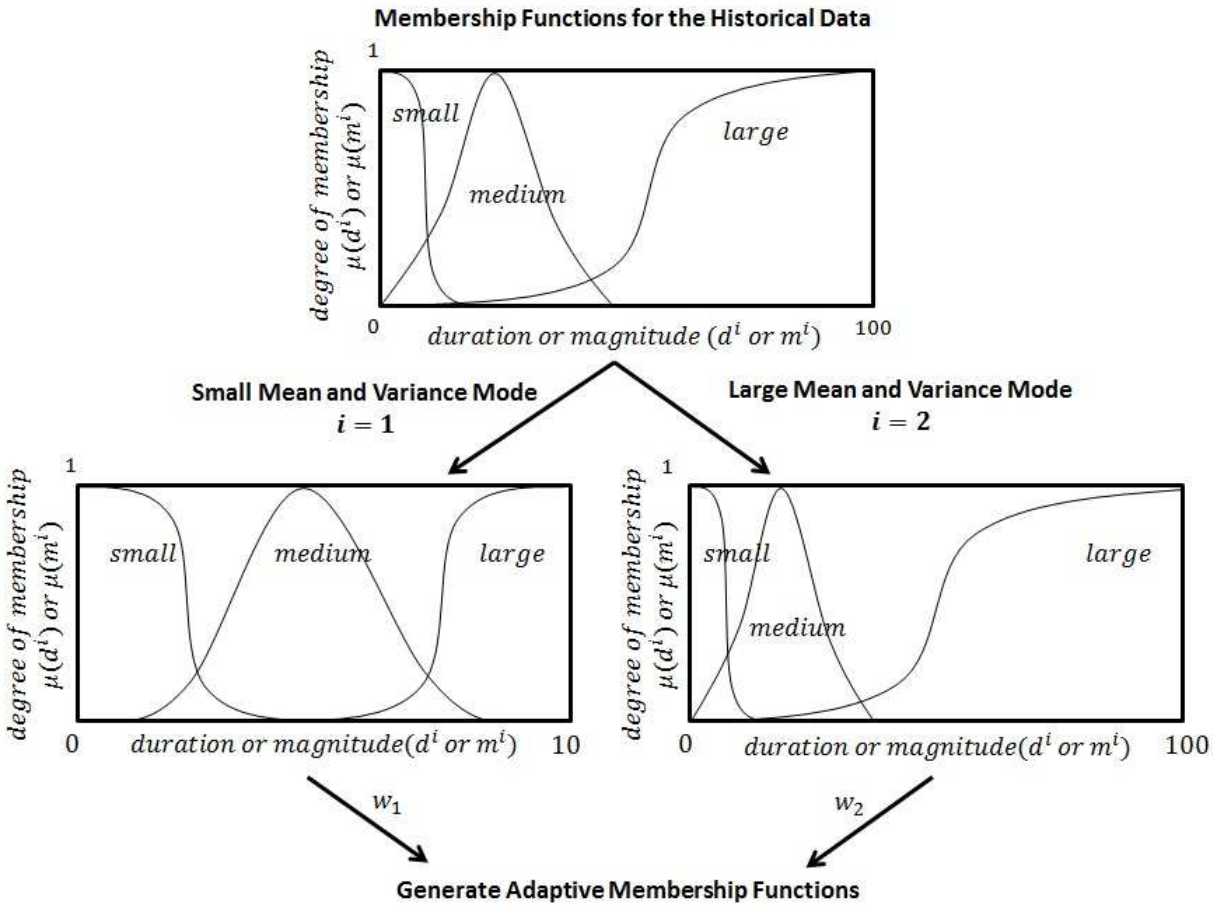

Figure 5.6: Combination of the fuzzy membership functions for different modes

As the result, parameters of the adaptive membership functions are calculated as following:

$$\sigma(t)_{d_t,m_t}^{sml,\ ave,\ lrg} = \sum_{i=1}^{M} w_i(t)\sigma_{d^i,m^i}^{sml,\ ave,\ lrg} \tag{5.21}$$

$$\mu(t)_{d_t,m_t}^{sml,\ ave,\ lrg} = \sum_{i=1}^{M} w_i(t)\mu_{d^i,m^i}^{sml,\ ave,\ lrg}$$

where $\sigma_{d^i,m^i}^{sml,\ ave,\ lrg}$ and $\mu_{d^i,m^i}^{sml,\ ave,\ lrg}$ are the means and variances of the Gaussian fuzzy membership functions for different modes previously calculated from the FCM approach.

Finally, using adaptive fuzzification and the method of triangular representation, a signal is converted to a sequence of symbols which contains 36 symbolic character alphabets. This symbolic representation of the signal is used in the next step to determine the faulty behavior.

5.6 Data Classification for Multiple Discrete Observations

In this section, assuming that the discrete observations are available from adaptive fuzzification and the method of triangular representation, a multivariate HMM procedure is adopted to model the normal and risky operations. These models are different from those in Section 5.4 in the sense that hidden states might not have physical meanings and are selected to best fit the data of the normal and risky operations. The data to train these models correspond to normal and risky operations which have previously occurred in history. After training the models, the only remaining step is to find the probability of process status from a window of multiple observations (N_W) given each model and decision making.

5.6.1 Data Classification Based on BPNNs

Wong et al. propose two approaches to solve the problem of multiple observations [20]. In the first approach, the classification of each individual variable based on the normal and abnormal models is used to generate a sequence of events for the individual variables. Based on this sequence of events, event sequencing HMMs are trained for overall classification. In the second approach, Back Propagation Neural Networks (BPNN) are applied for overall classification. This approach uses the probabilities of each classification category for each individual variable as the inputs to a neural network. The outputs of the neural network represent the probabilities that the overall classification belongs to each classification category. The overall classification of the plant is determined based on the highest output probability of the BPNN.

5.6.2 Data Classification Based on HMMs

The recent studies on hidden Markov models with multiple observations are based on the early study of Levinson et al. in American telephone and telegraph company [99]. Baggenstoss proposed a modified Baum Welch algorithm for training of hidden Markov models with multiple observations [100]. The training method can be further extended to second or higher order hidden Markov models [101].

Since process variables are correlated, the sequence resulted from the triangular representation must also be correlated. In this chapter, we propose to adopt a multivariate modeling approach to train HMMs. Li et al. has laid theoretical foundation on multivariate HMM modeling, which is adopted here [94].

Consider the following set of observation sequences:

$$O = \{O^1, O^2, ..., O^K\} \tag{5.22}$$

where

$$O^{(k)} = o_1^{(k)}, o_2^{(k)}, ..., o_{T_k}^{(k)},\ 1 \leq k \leq K \tag{5.23}$$

are individual observation sequences. As an example, for the industrial case study of this chapter in Section 5.9, K will be equal to 2, which corresponds to the variables interface level and underflow density. T_k the number of discrete observations used for training.

The following expression is always true when calculating probability of the observation sequence given the model:

$$P(O|\lambda) = P(O^{(1)}|\lambda)P(O^{(2)}|O^{(1)}, \lambda), ..., P(O^{(K)}|O^{(K-1)}, ..., O^{(1)}, \lambda) \tag{5.24}$$

$$P(O|\lambda) = P(O^{(2)}|\lambda)P(O^{(3)}|O^{(2)}, \lambda), ..., P(O^{(1)}|O^{(K)}, ..., O^{(2)}, \lambda)$$

...

$$P(O|\lambda) = P(O^{(K)}|\lambda)P(O^{(1)}|O^{(K)}, \lambda), ..., P(O^{(K-1)}|O^{(K)}, O^{(K-2)}, ..., O^{(1)}, \lambda)$$

Therefore, the probability of the multiple observations given the model can be written as the summation

$$P(O|\lambda) = \sum_{k=1}^{K} \omega_k P(O^{(k)}|\lambda) \tag{5.25}$$

where

$$\omega_1 = \frac{1}{K} P(O^{(2)}|O^{(1)}, \lambda), ..., P(O^{(K)}|O^{(K-1)}, ..., O^{(1)}, \lambda) \tag{5.26}$$

$$\omega_2 = \frac{1}{K} P(O^{(3)}|O^{(2)}, \lambda), ..., P(O^{(1)}|O^{(K)}, ..., O^{(2)}, \lambda)$$

...

$$\omega_K = \frac{1}{K} P(O^{(1)}|O^{(K)}, \lambda), ..., P(O^{(K-1)}|O^{(K)}, O^{(K-2)}, ..., O^{(1)}, \lambda)$$

are weights. Li et al. show that the following re-estimation formulas can be obtained for parameters of the hidden Markov model through the calculation of the auxiliary function at the E-step and maximizing it in the M-step [94]. If it is assumed that the individual observation sequences are serially independent of each other, i.e.,

$$P(O|\lambda) = \prod_{k=1}^{K} P(O^{(k)}|\lambda) \tag{5.27}$$

The combinatorial weights become

$$\omega_k = \frac{1}{K}\frac{P(O|\lambda)}{P(O^{(k)}|\lambda)},\ 1 \le k \le K \tag{5.28}$$

Substituting the above weights into equations, the following training equations will be obtained:

State transition probability:

$$a_{mn} = \frac{\sum_{k=1}^{K}\sum_{t=1}^{T_k-1}\xi_t^{(k)}(m,n)}{\sum_{k=1}^{K}\sum_{t=1}^{T_k-1}\gamma_t^{(k)}(n)},\ 1 \le m \le Q,\ 1 \le n \le Q \tag{5.29}$$

where Q is the number of states in the model. Other terms of Equation 5.29 (ξ and γ) have been previously discussed in Chapter 2.

Symbol emission probability:

$$b_n(m) = \frac{\sum_{k=1}^{K}\sum_{t=1}^{T_k-1}\delta(o_t^{(k)},\nu_m)\gamma_t^{(k)}(n)}{\sum_{k=1}^{K}\sum_{t=1}^{T_k-1}\gamma_t^{(k)}(n)},\ 1 \le m \le R,\ 1 \le n \le Q \tag{5.30}$$

where $\delta(o_t^{(k)},\nu_m)$ is equal to 1 if $o_t^{(k)} = \nu_m$ and 0 otherwise, and R is the number of observation symbols per state.

Initial state probability:

$$\pi_n = \frac{1}{K}\sum_{k=1}^{K}\gamma_1^{(k)}(n),\ 1 \le n \le Q \tag{5.31}$$

After training the models, as a new window of discrete observations (N_W) is received, the probability of the process status for the window given each model is calculated based on the forward-backward algorithm [40]. The decision on the operating condition (normal or risky) of the system is based on the model with a greater probability.

5.7 Data Classification Based on HMMs – A Moving Window Approach

In this section a moving window approach is proposed for process monitoring based on symbolic episode representation and hidden Markov models. All the materials of

this section are developed based on fixed fuzzy membership functions. More advanced approaches for the case of time varying membership functions are subjects of future studies.

Although BPNNs are limited to fixed window sizes, time varying input dimensions can be considered at each time step using HMMs [19]. Therefore, a search algorithm to find the optimal window size is proposed in this section. Unlike the previous approaches [102], this algorithm searches for a fixed episode of more informative observations in the window.

Selection of large window sizes has the drawback of remaining in transition zones for large time intervals where no decision can be made on the operating condition of the system. The minimum number of observations (N_{min}) required to thoroughly explain the operating condition might differ according to the level of noise removal, on-line sampling rate, etc. Although N_{min} contains a window of most recent observations, making the final decision only based on N_{min} may cause many false alarms and affect critical decisions. One solution to this problem is searching for N_{min} number of observations in a window of most recent data (N_W). Intuitively, using the proposed methodology in Figure 5.8, we are looking for a small window of observations which maximizes the difference of the likelihood given each model. Therefore, assuming $e_{max} = N_W - N_{min}$,

$$e_{opt} = argmax_{e\in[0,e_{max}]}\{(P(O|\lambda_{Normal}) - P(O|\lambda_{Abnormal}))^2|_{O=O(\tau+e:\tau+e+N_{min})}\} \quad (5.32)$$

where $\tau = t - N_W$, $O = O(\tau+e : \tau+e+N_{min})$ is the optimal episode of observations in the window and $\lambda_{Normal/Abnormal}$ represents the HMMs trained for normal or abnormal conditions. $P(O|\lambda)$ is calculated from the forward-backward algorithm [40].

Using the search algorithm in Equation 5.32, we are looking for an episode of N_{min} observations which best classifies the normal and abnormal operations in the window of N_W observations. A schematic of the proposed algorithm is illustrated in Figure 5.7.

A summary of the proposed algorithm is presented in Figure 5.8.

5.8 Simulation Case Study

This simulation case study is the same as the CSTRs in series example as explained in Chapter 3. For convenience, the diagram of the process is presented again in Figure 5.9.

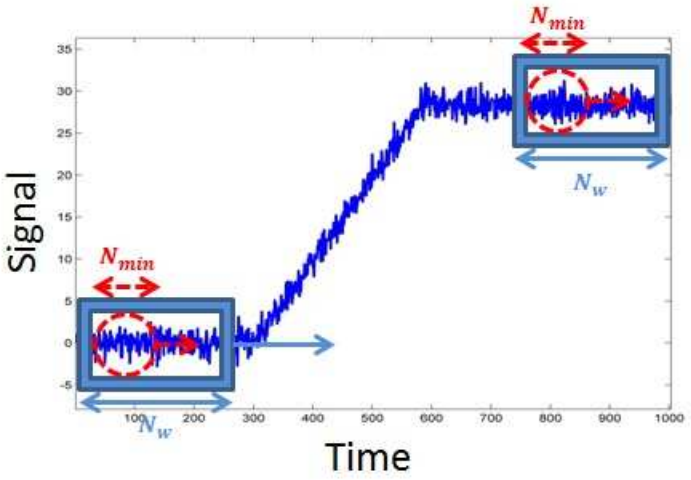

Figure 5.7: Optimal window size selection

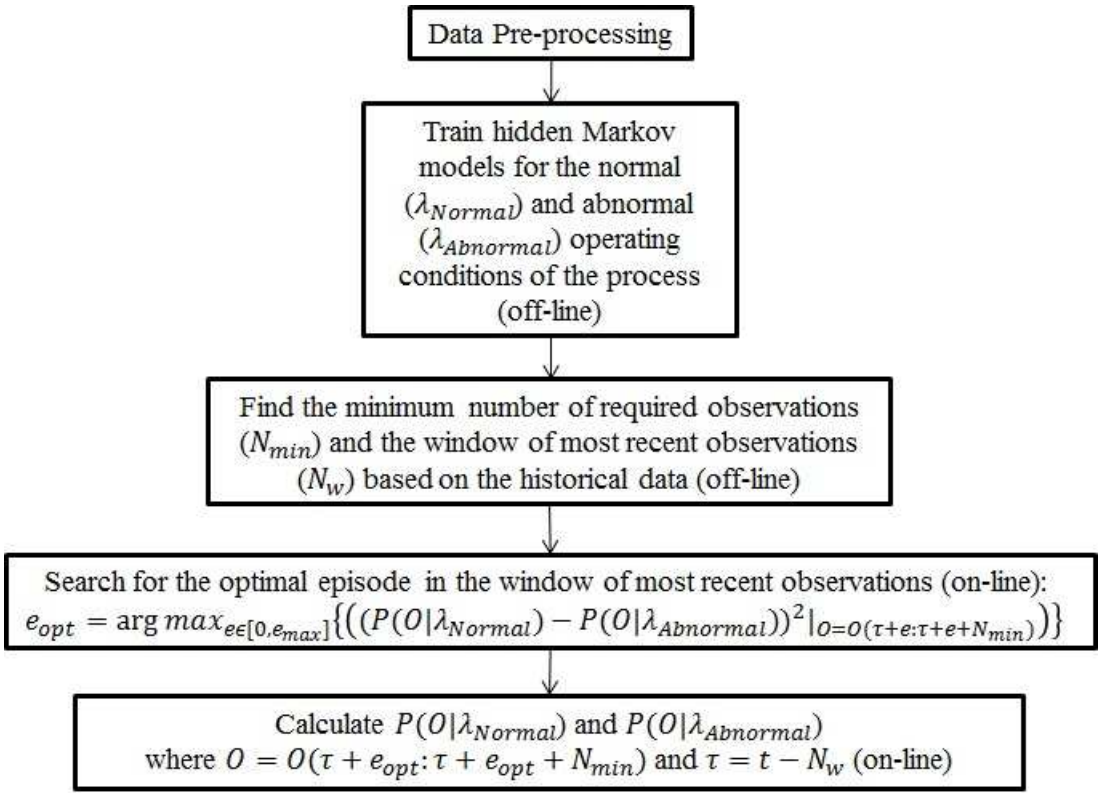

Figure 5.8: Summary of the proposed methodology of this paper

In this section, the initial value of C_{A2} is equal to 0.05 mol/L and the set-point is selected as 0.075 mol/L. As explained in [67], a PI controller with parameters $\tau_I = 0.25$ min and $K_C = 350$ $L^2/mol.min$ is implemented for control purposes. Here, it is assumed that a white noise disturbance with variance 0.5 always disturbs the

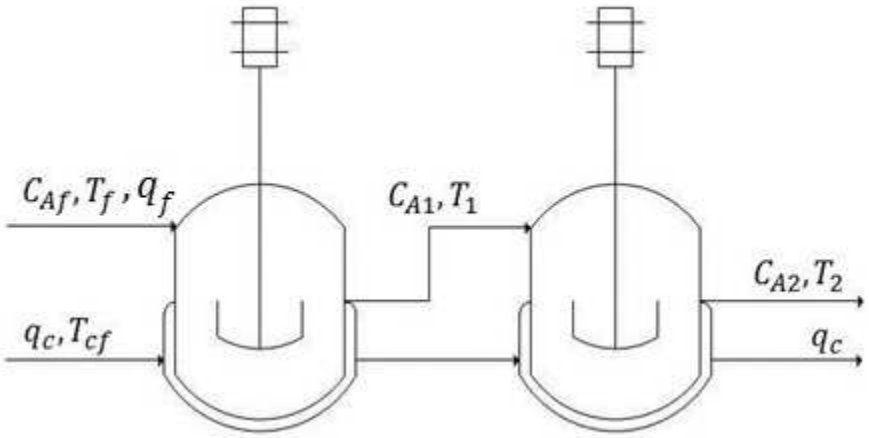

Figure 5.9: CSTR reactors in series [67]

feed flow-rate (q_f).

5.8.1 Comparison of the BPNN, HMM and HMM with Adaptive Fuzzification

In this simulation case study, first, using fixed fuzzy membership functions for triangular representation, the performances of the BPNN and HMM methods are compared. Next, the advantage of applying adaptive fuzzification for signal discretization is illustrated and compared to fixed fuzzy membership functions. It is assumed that a number of random pulse disturbances with random magnitude with mean 15 L/min and variance 4 occur in the feed flow-rate (q_f). The disturbance (q_f) and output signals C_{A2} and T_2 are presented in Figures 5.10 and 5.11.

Triangular representation of the signals applying fixed fuzzy membership functions are presented in Figures 5.12 and 5.13. Two direct red lines correspond to the start and end of abnormal operations of Figure 5.11.

Results of operating condition classification based on the BPNN and HMM methods using fixed fuzzy membership functions and a fixed window of $N_W = 5$ discrete observations are illustrated in Figures 5.14 and 5.15.

From the results in Figures 5.14 and 5.15, one could see that applying fixed fuzzy membership the HMM method could improve the performance and reduce the number of false alarms. The main disadvantage of the BPNN method for the purpose of overall classification ([20]) is no consideration of the interactions between different inputs in the training step. Therefore, each input affects the classification separately. This will cause a number of false alarms for the operating condition diagnosis.

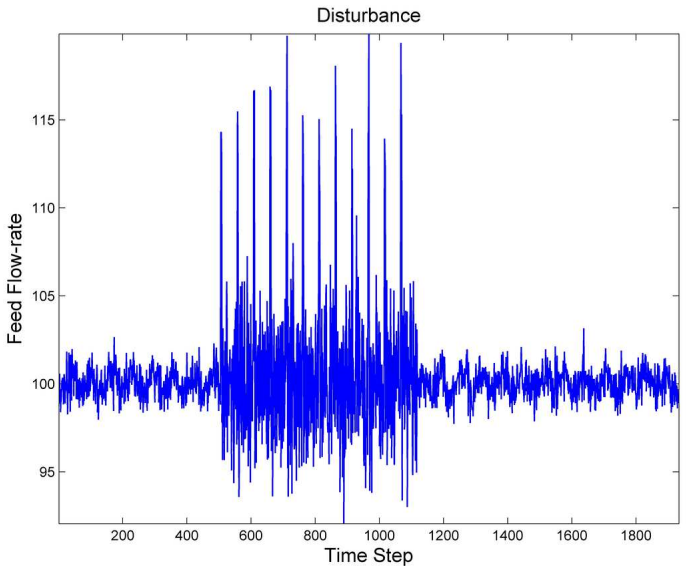

Figure 5.10: Normal and abnormal operations in the CSTRs in series - Feed flow rate

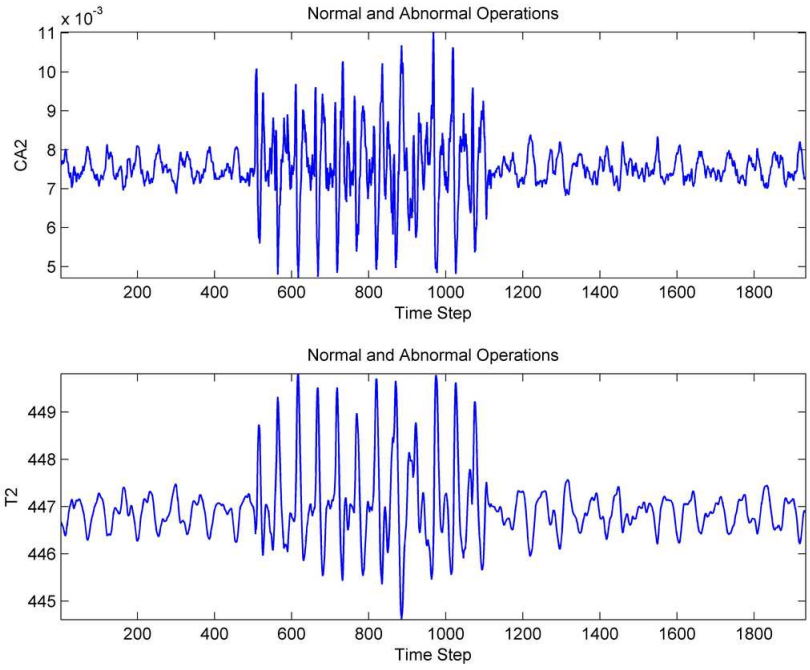

Figure 5.11: Normal and abnormal operations in the CSTRs in series - Output temperature and concentration

Results of the process classification can be further improved by applying adaptive fuzzy membership functions for signal discretization. Adaptive fuzzy membership functions are presented in Figures 5.16 and 5.17. The discretized observations are

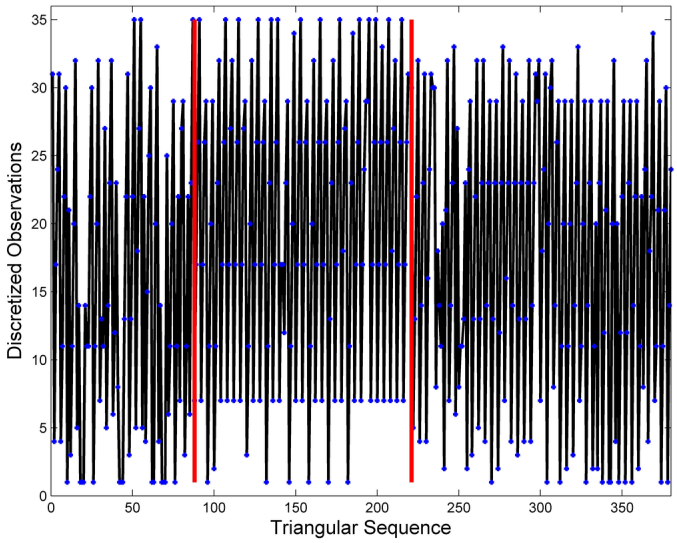

Figure 5.12: Triangular representation of output signals using fixed fuzzy membership functions - Output concentration (C_{A2})

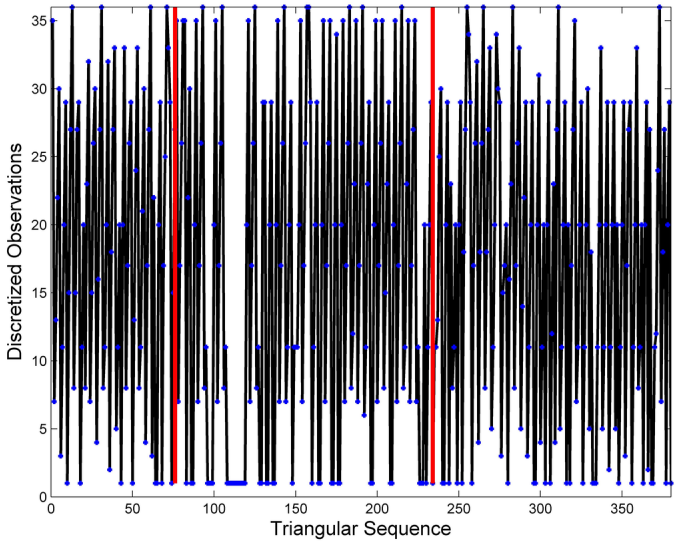

Figure 5.13: Triangular representation of output signals using fixed fuzzy membership functions - Output Temperature (T_2)

shown in Figures 5.18 and 5.19 respectively. From Figures 5.18 and 5.19, one could observe that the discrete observations (specifically in the normal mode) are generated with wider variety and less large type triangles (type 36 for example). In general, adaptive discretization has provided more distinguishable patterns.

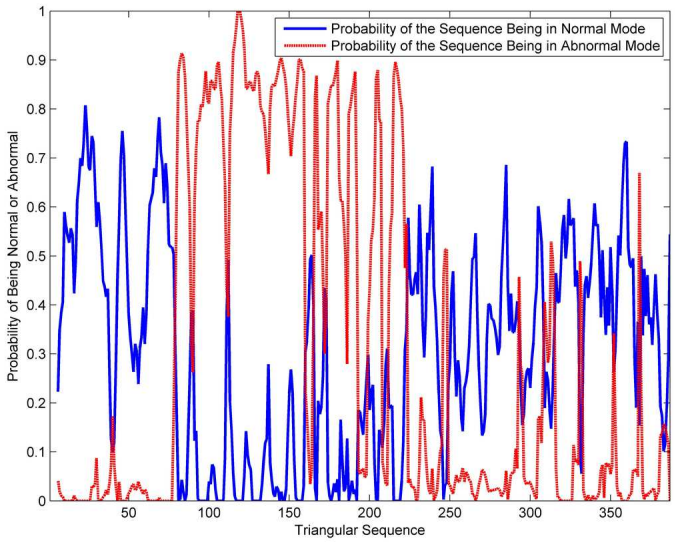

Figure 5.14: Results of the classification of normal and abnormal operating conditions based on fixed fuzzy membership functions ($NW = 5$) - BPNN method

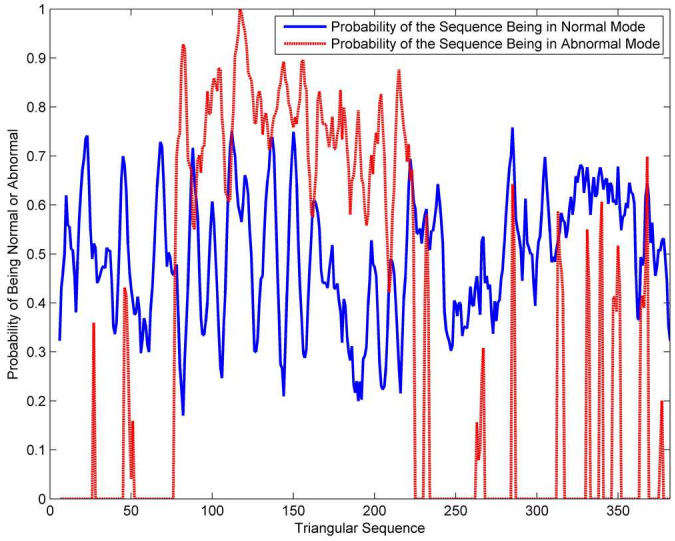

Figure 5.15: Results of the classification of normal and abnormal operating conditions based on fixed fuzzy membership functions ($NW = 5$) - HMM method

Results of overall classification based on the discrete observations in Figures 5.18 and 5.19 are presented in Figure 5.20.

From the results in Figure 5.20, it can be observed that having the same number of discrete observations in the moving window ($N_W = 5$), due to the presence of more precise discrete observations and patterns from adaptive fuzzification, the number of

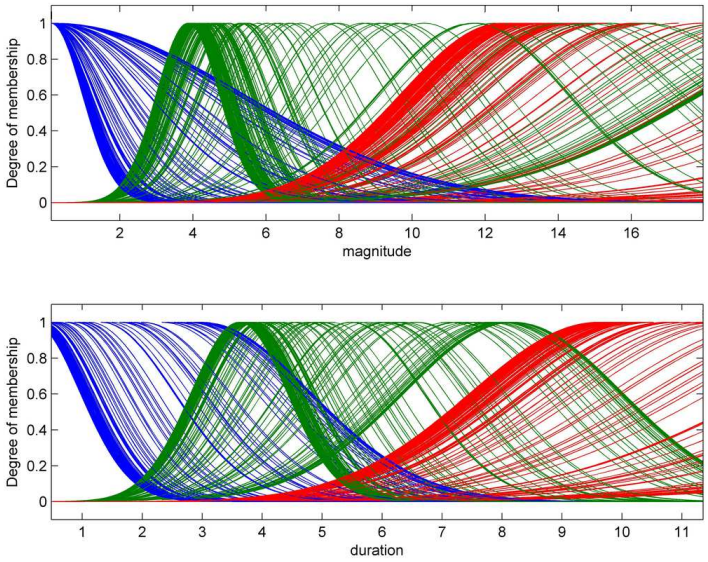

Figure 5.16: Adaptive fuzzy membership functions for the output signals (Figure 5.11) in different operating conditions - Output concentration (C_{A2})

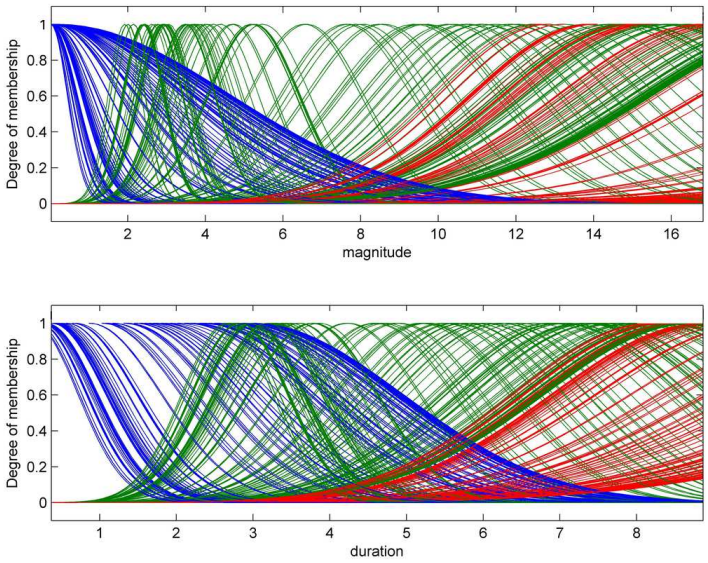

Figure 5.17: Adaptive fuzzy membership functions for the output signals (Figure 5.11) in different operating conditions - Output temperature (T_2)

false alarms has been greatly reduced.

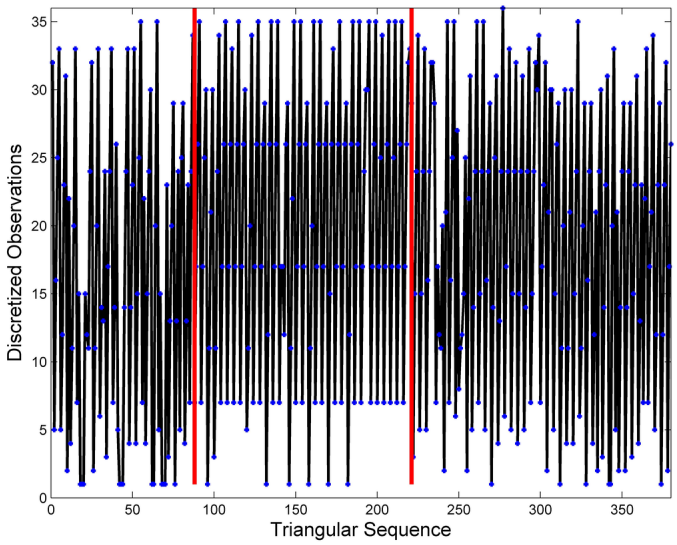

Figure 5.18: Triangular representation of output signals in Figure 5.11 using adaptive fuzzy membership functions - Output concentration (C_{A2})

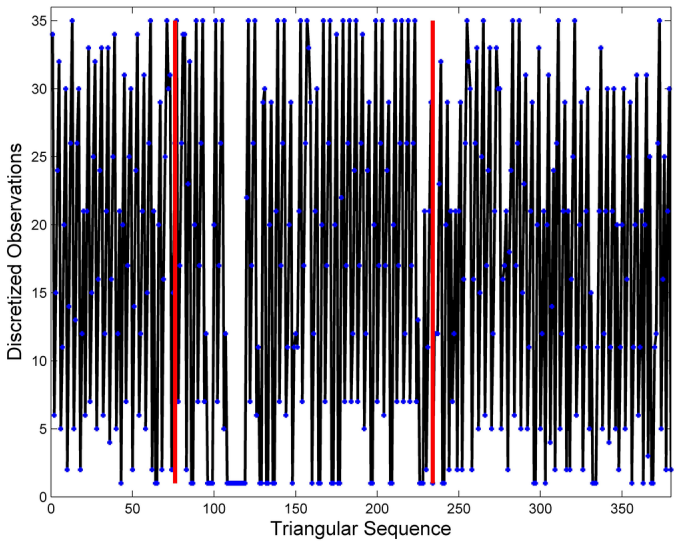

Figure 5.19: Triangular representation of output signals in Figure 5.11 using adaptive fuzzy membership functions - Output temperature (T_2)

5.8.2 Detection of Various Types of Faults Based on HMM and HMM with Adaptive Fuzzification

The merit of Adaptive fuzzification can be more clearly illustrated when different types of faults with random magnitudes and periods occur in the process. In this

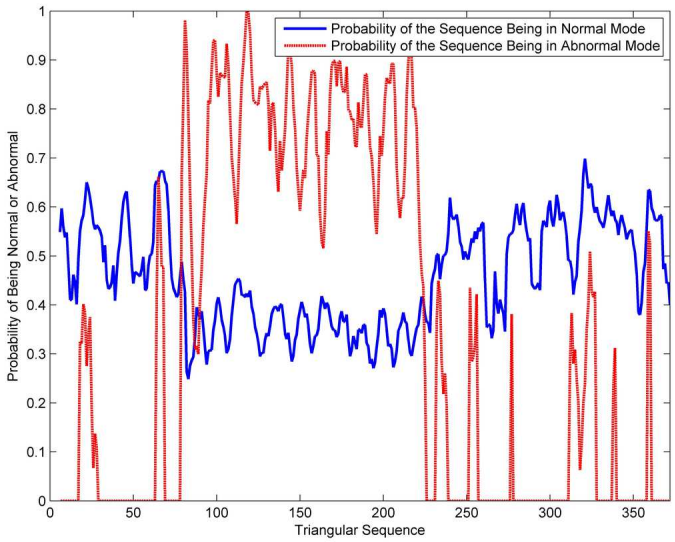

Figure 5.20: Results of the classification of normal and abnormal operating conditions in Figure 5.11 based on adaptive fuzzy membership functions ($N_W = 5$)

example, it is assumed that, first, some random impulse signals with mean 15 L/min and variance 4 occur in the feed flow rate. This disturbance is then followed by a number of random ramp type disturbances with maximum output value 3 L/min and variance 1, which cause some long duration and small magnitude faults in the process outputs. The normal and abnormal operations of the process in this example are presented in Figures 5.21 and 5.22. The two types of abnormal behaviors are named as abnormal conditions 1 and 2 respectively for the rest of this section.

First, both signals are discretized using fixed fuzzy membership functions. Results are presented in Figures 5.23 and 5.24.

Next, the same procedure as in the previous section is repeated to generate adaptive fuzzy membership functions. The membership functions are presented in Figures 5.25 and 5.26. The discretized observations are presented in Figures 5.27 and 5.28 respectively.

Comparing the fuzzy membership functions in Figures 5.25 and 5.26 with Figures 5.16 and 5.17, one could observe that the membership functions are shifted to larger durations and magnitudes due to the presence of more large triangles on average. Furthermore, similar to the discrete observations in Figures 5.18 and 5.19, comparing the discrete observations in Figures 5.23 and 5.24 with Figures 5.27 and 5.28, it can be seen that more variety of triangles are generated for the low mean and variance (normal) mode. In other words, adaptive fuzzification will adaptively change the fuzzy parameters. Therefore, the membership functions will not be dominated by the

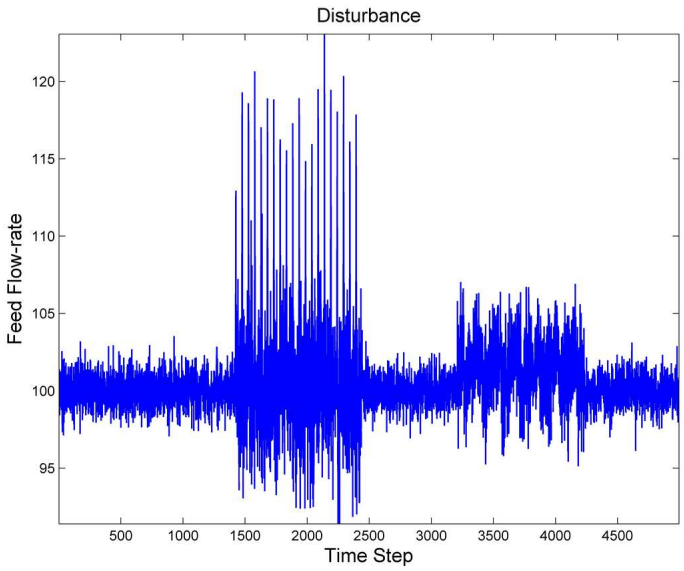

Figure 5.21: Normal, abnormal 1 and abnormal 2 operating conditions for the CSTRs in series - Feed flow rate

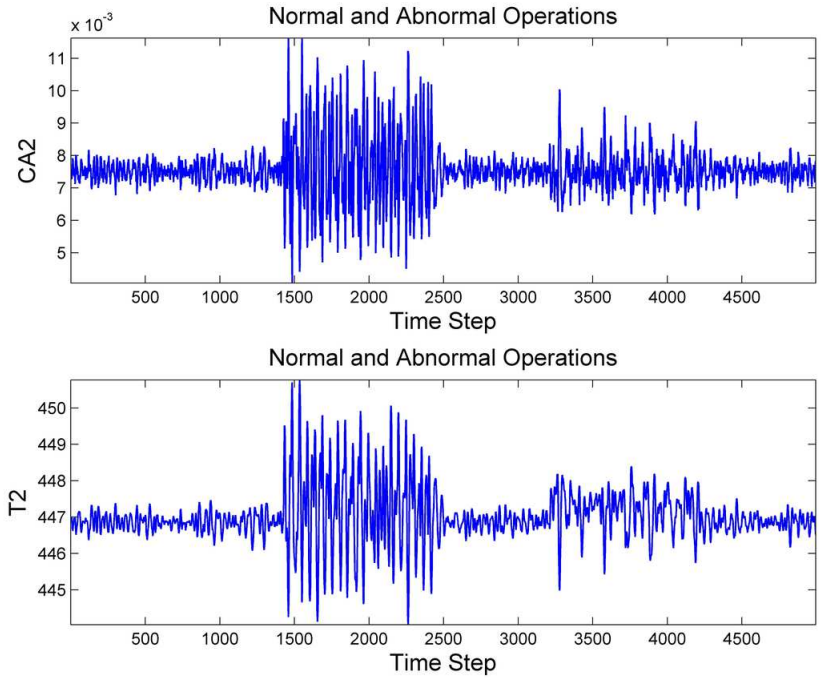

Figure 5.22: Normal, abnormal 1 and abnormal 2 operating conditions for the CSTRs in series - Output temperature and concentration

modes with larger duration and magnitude. Results of the overall classifications of the process based on HMMs using a fixed window of $N_W = 5$ observations, with and

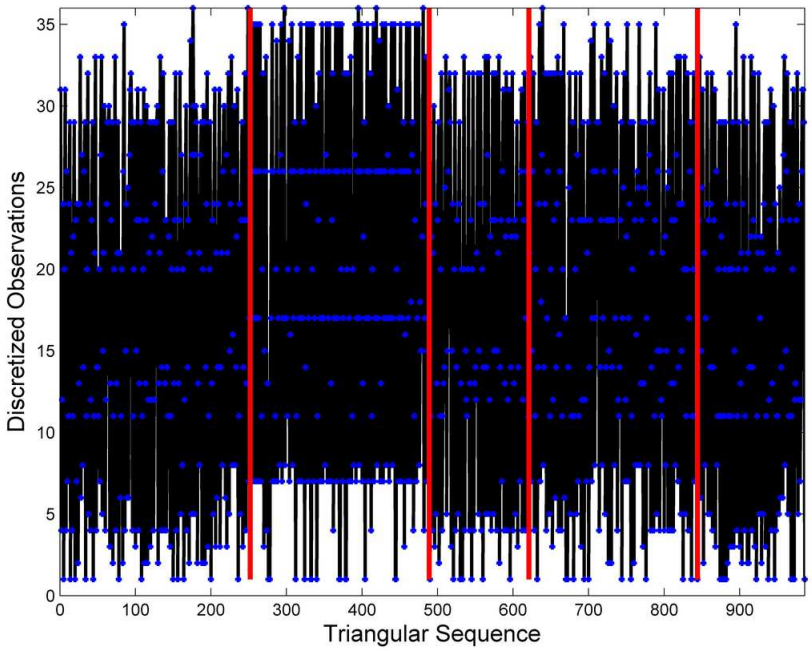

Figure 5.23: Triangular representation of normal, abnormal 1 and abnormal 2 output signals in Figure 5.22 using fixed fuzzy membership functions - Output concentration (C_{A2})

without using adaptive membership functions, are presented in Figures 5.29 and 5.30 respectively.

As presented in Figure 5.30, although still some false alarms appear in the results (in the time period between 200 and 300 for example), in general, applying adaptive membership functions, normal and abnormal modes (especially normal mode and abnormal mode 1) are more clearly distinguishable. As previously stated, this is due to the fact that application of adaptive fuzzy membership functions can provide more precise discrete observations and patterns considering different modes for the durations and magnitudes of the signals.

5.8.3 Detection of Size of the Faults Based on the HMM and Adaptive Fuzzification

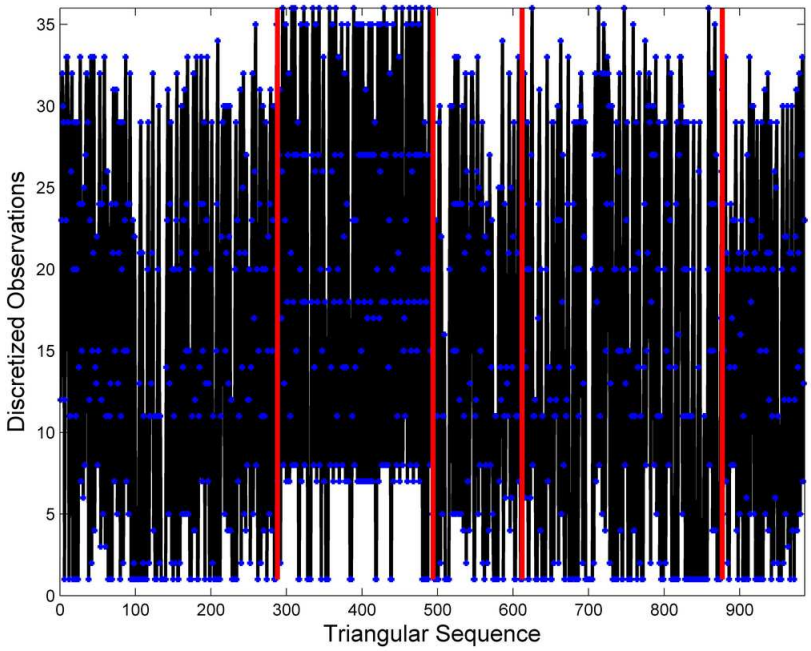

Figure 5.24: Triangular representation of normal, abnormal 1 and abnormal 2 output signals in Figure 5.22 using fixed fuzzy membership functions - Output temperature (T_2)

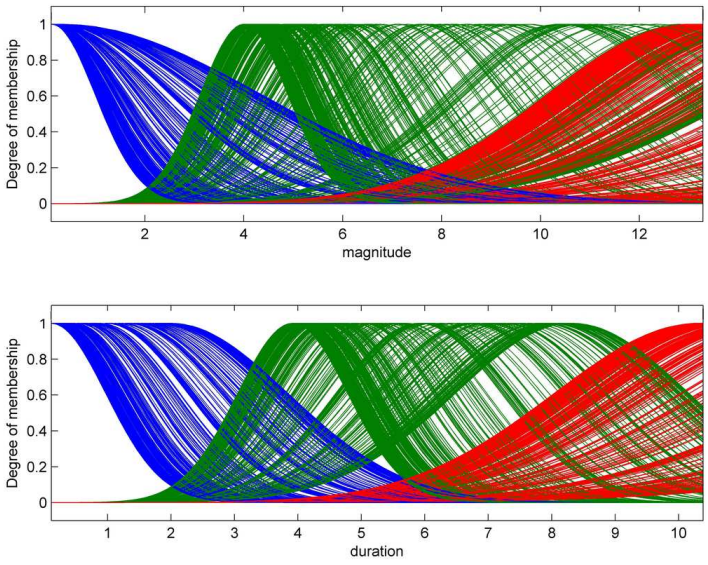

Figure 5.25: Adaptive fuzzy membership functions for the output signals of the Figure 5.22 in different operating conditions - Output concentration (C_{A2})

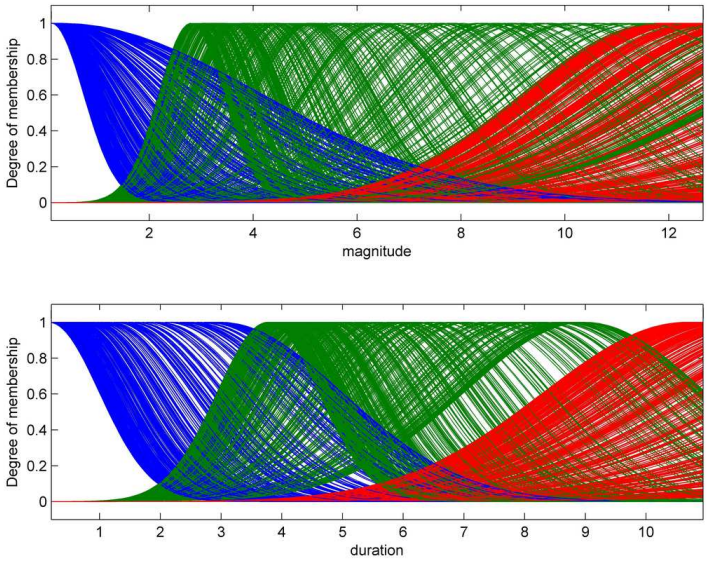

Figure 5.26: Adaptive fuzzy membership functions for the output signals of the Figure 5.22 in different operating conditions - Output temperature (T_2)

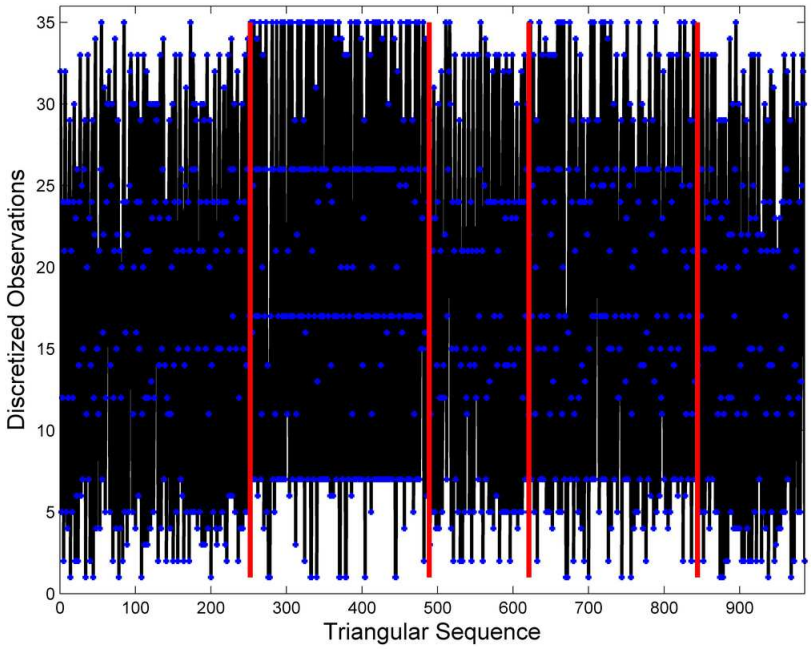

Figure 5.27: Triangular representation of normal, abnormal 1 and abnormal 2 output signals in Figure 5.22 using adaptive membership functions - Output concentration (C_{A2})

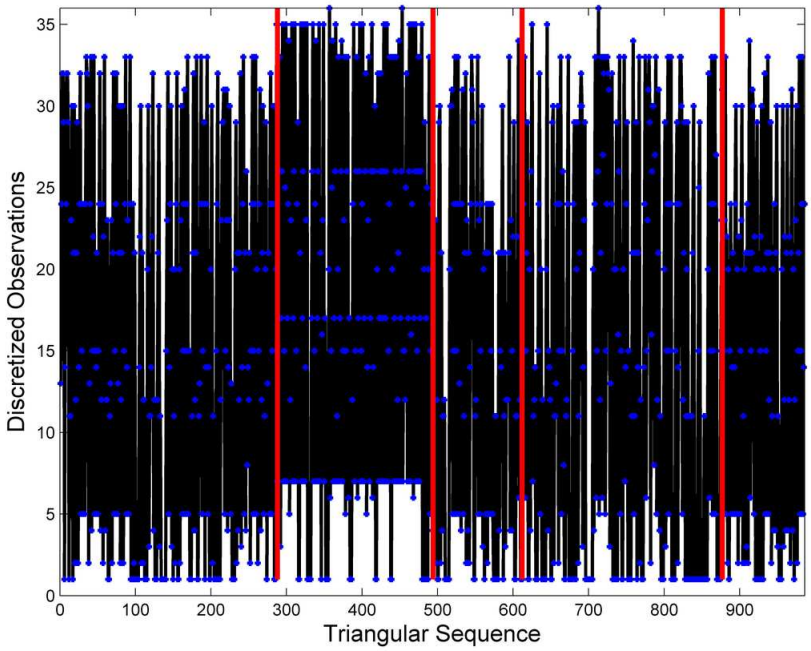

Figure 5.28: Triangular representation of normal, abnormal 1 and abnormal 2 output signals in Figure 5.22 using adaptive membership functions - Output temperature (T_2)

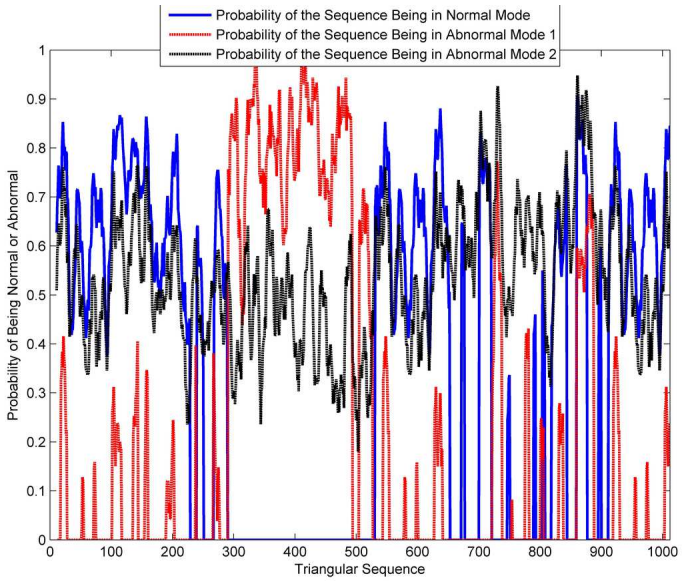

Figure 5.29: Results of the classification of normal, abnormal 1 and abnormal 2 operating conditions in Figure 5.22 based on the HMM method ($NW = 5$) - Fixed fuzzy membership functions

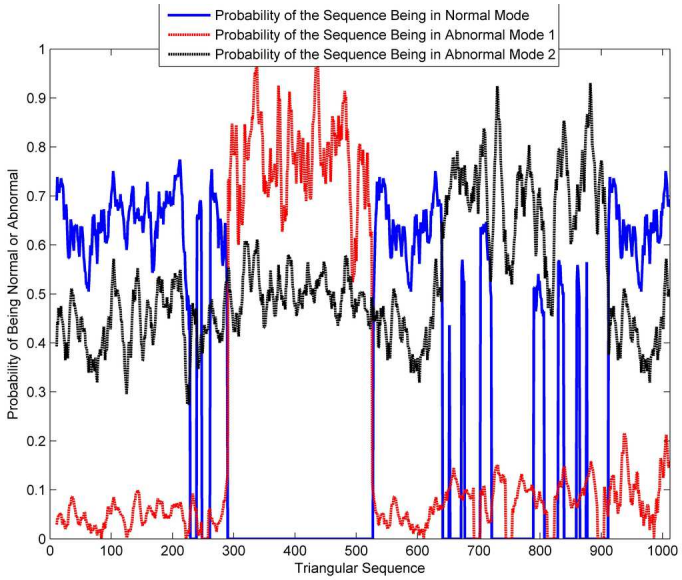

Figure 5.30: Results of the classification of normal, abnormal 1 and abnormal 2 operating conditions in Figure 5.22 based on the HMM method ($NW = 5$) - Adaptive fuzzy membership functions

In this section, the ability of the proposed method in detection of faults with different magnitudes is investigated. It is assumed that two random ramp type disturbances, one with maximum output value 8 L/min and variance 3, and the other one with maximum output value 3 L/min and variance 1 occur in the feed flow-rate. Similar to the previous section, operating conditions of the process are named as normal, abnormal 1 and abnormal 2. Different operating conditions of the process are presented in Figures 5.31 and 5.32.

Using adaptive fuzzy membership functions, discrete observations are generated as in Figures 5.33 and 5.34.

Results of overall classification of the process based on HMMs and a fixed window of five observations ($N_W = 5$) are presented in Figure 5.35.

Based on the results in Figure 5.35, it can be concluded that to some large extent the method is able to determine the size of the faults. However, a number of false alarms appear and there are some periods during which an unknown pattern arises (the time period around 500 for example). Therefore, using the method of triangular representation for complete isolation of faults might have some mis-detection. Increasing the number of modes for adaptive fuzzification could be a solution to provide more precise patterns and classifications for faults with different magnitudes. However, the increasing computational cost will be unavoidable. As the result, using the proposed method of this chapter for high accuracy fault isolation purposes might

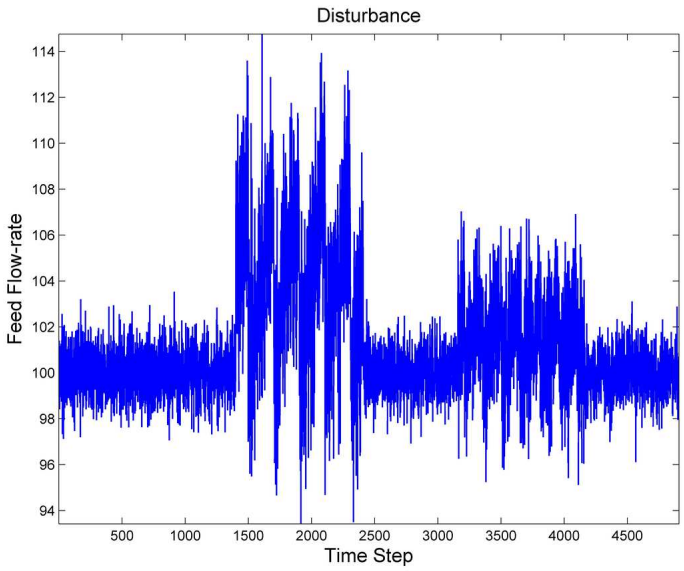

Figure 5.31: Normal, abnormal 1 and abnormal 2 operating conditions for the CSTRs in series - Feed flow rate

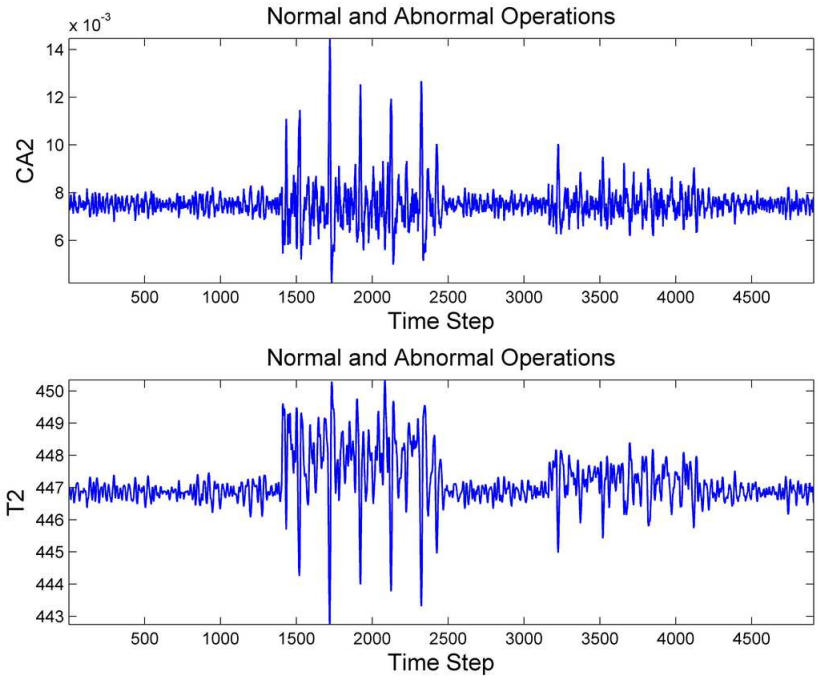

Figure 5.32: Normal, abnormal 1 and abnormal 2 operating conditions for the CSTRs in series - Output temperature and concentration

suffer from the computation limit in an on-line application. Furthermore, as the results in this section show, if the training data is selected sufficiently informative, the

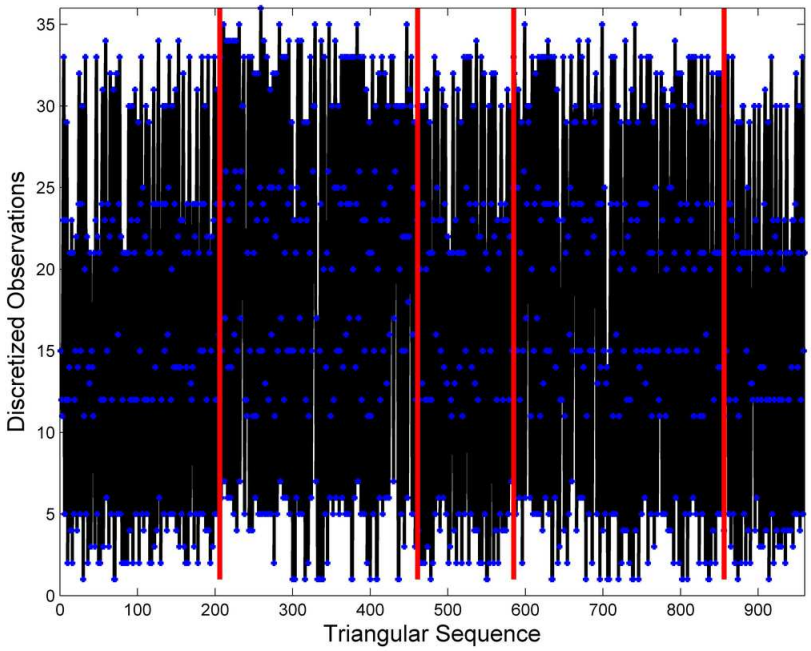

Figure 5.33: Triangular representation of normal, abnormal 1 and abnormal 2 output signals in Figure 5.32 using adaptive membership functions - Output concentration (C_{A2})

method is robust in abnormal condition diagnosis.

5.8.4 Data Classification Based on a Moving Window of Observations

In this section, a pulse disturbance with the amplitude 15 (L/min) and period of five samples in time steps between $610 - 1210$ is assumed to occur in the feed flow rate which results in overshoots in process outputs. A combination of normal and abnormal operating regions after reaching the desired set-point is presented in Figure 5.36.

After removing the high frequency noise in two levels using wavelet analysis, and normalizing the data, minimum, maximum and inflection points of the signals are calculated. Then, using appropriate fixed fuzzy membership functions and rules for durations and magnitudes, signals are converted to discrete observations. Discretized observations of the output concentration and temperature signals are presented in Figures 5.37 and 5.38.

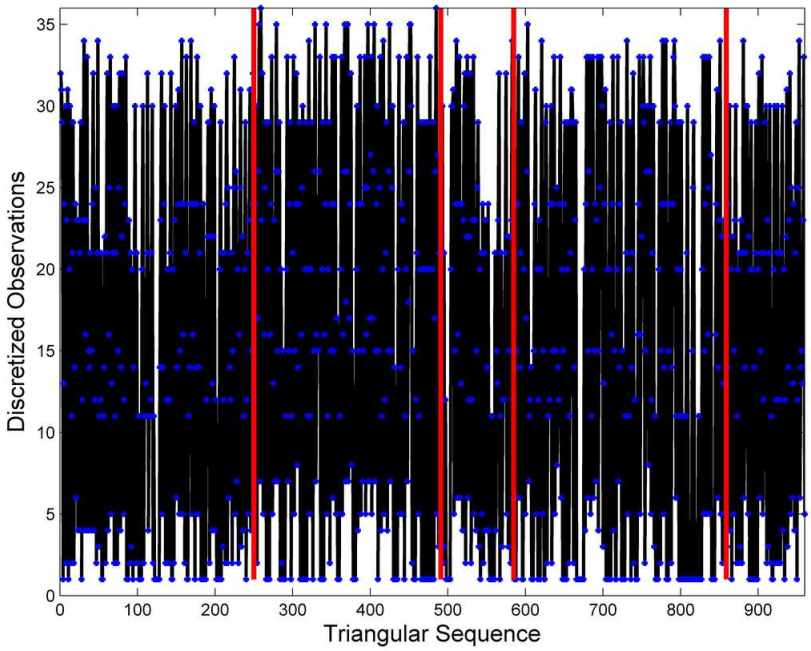

Figure 5.34: Triangular representation of normal, abnormal 1 and abnormal 2 output signals in Figure 5.32 using adaptive membership functions - Output temperature (T_2)

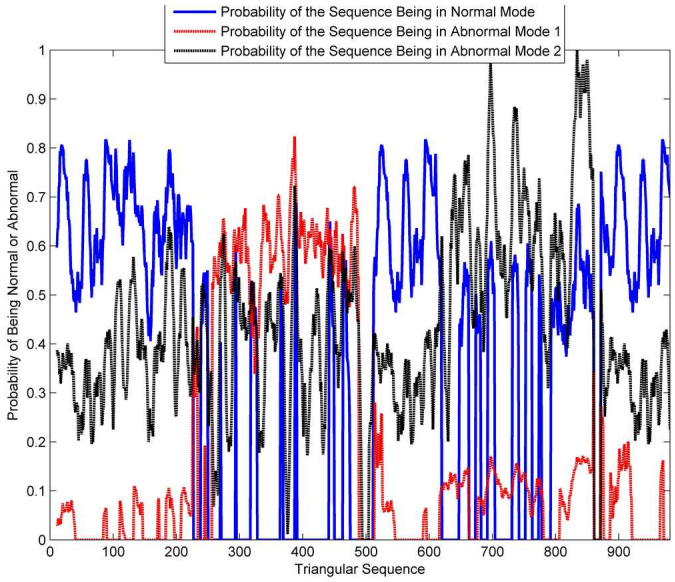

Figure 5.35: Results of the classification of normal and abnormal operating conditions in Figure 5.32 based on adaptive fuzzy membership functions ($NW = 5$)

Large Window of Input Data (non-adaptive window sizes)

Large window of input data for process classification provides similar results between the BPNN and HMM approaches with fixed fuzzy membership functions. Figures 5.39

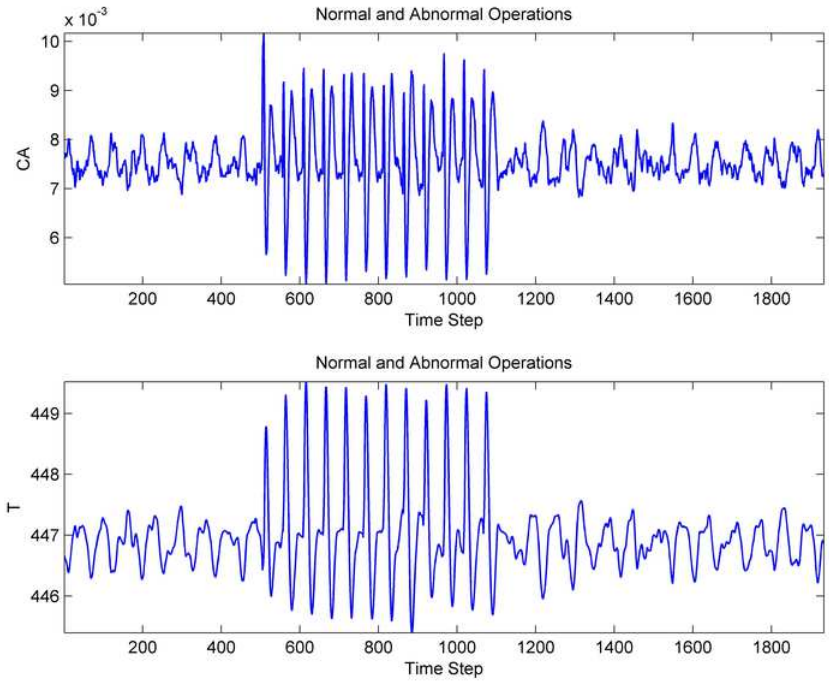

Figure 5.36: Normal and abnormal operations after reaching the desired set-point

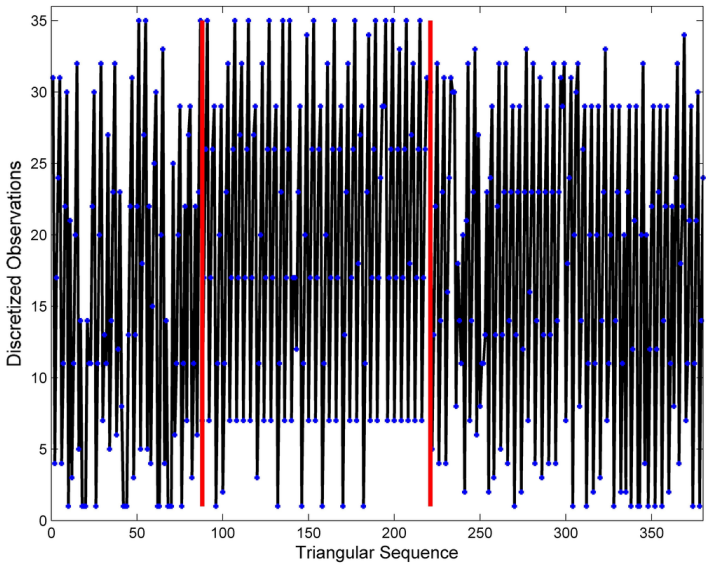

Figure 5.37: Discritized observations for the output concentration in Figure 5.36

and 5.40 compare the results of the two approaches considering 10 last observations as the input to the classification algorithms. A total number of 772 discrete observations, including 579 observations for normal and 193 for abnormal regions, are used to train

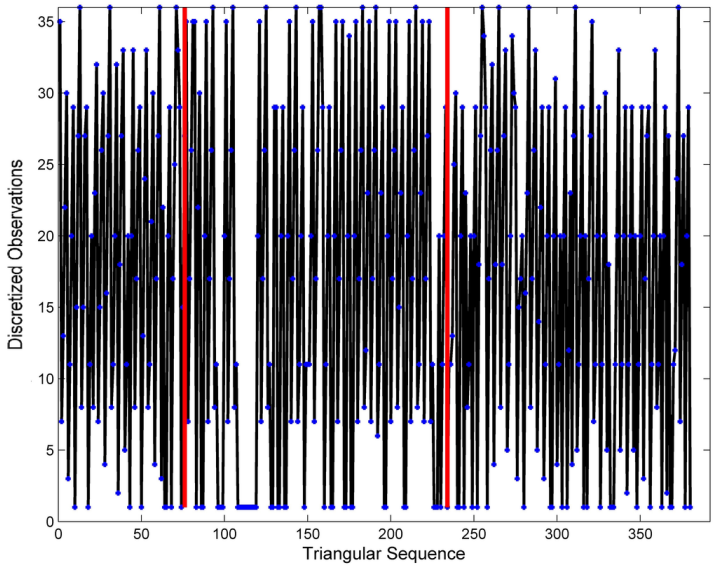

Figure 5.38: Discretized observations for the output temperature in Figure 5.36

the models of the normal and abnormal operating conditions.

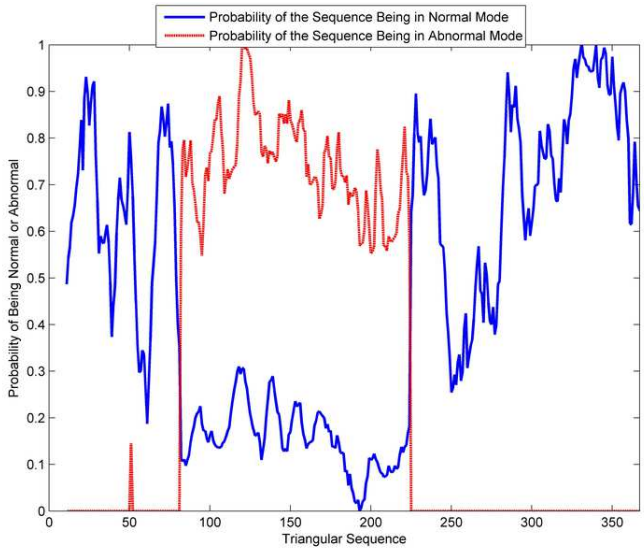

Figure 5.39: Normalized probability of the observation sequences in Figure 5.36 belonging to normal and abnormal regions using HMMs with fixed window of data ($N_W = N_{min} = 10$)

As presented in Figures 5.39 and 5.40, with a large window of the input data, both approaches provide similar results.

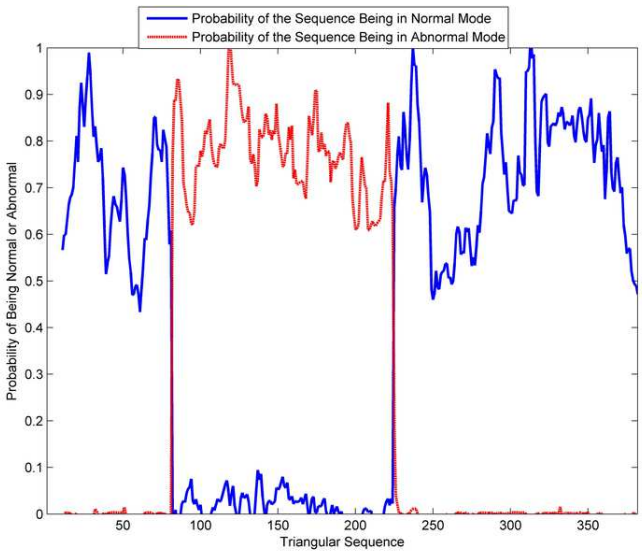

Figure 5.40: Normalized probability of the observation sequences in Figure 5.36 belonging to normal and abnormal regions using the BPNN approach ($N_W = N_{min} = 10$)

Small Window of Input Data (adaptive window sizes)

Using large window sizes, a long time will be required for any classification algorithm to capture the most recent behavior of the process due to a large amount of old data in the window. Consequently, the classification algorithm remains in transition zones where no decision can be made on the operating condition of the system. Figures 5.41 to 5.43 present the results of the overall classification with 5 observations as the input of the classification system.

As it is clear from Figure 5.41, a large number of false alarms appear when the window size is reduced to half using the BPNN approach.

Figure 5.42 shows the result of overall classification based on the proposed moving window method. The number of false alarms is reduced. However, the likelihood ratio is also decreased. In other words, the overall decision making has been improved while the individual effect of each variable is reduced. The number of shifts to find the optimal window (e_{opt}) is presented in Figure 5.43. This number varies between 1 and $e_{max} = N_W - N_{min} = 9 - 5 = 4$ and indicates the $O_{opt} = O(\tau + e_{opt} : \tau + e_{opt} + N_{min})$ sequence of observations which are selected for overall classification.

Following the proposed moving window procedure, the final decision will be based on the more informative observations in the window and the old information will not affect the overall decision making.

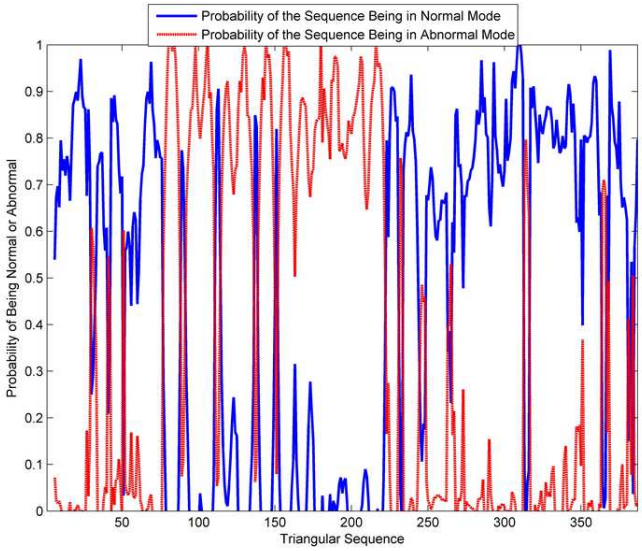

Figure 5.41: Normalized probability of the observation sequences in Figure 5.36 belonging to normal and abnormal regions using the BPNN approach ($N_W = N_{min} = 5$)

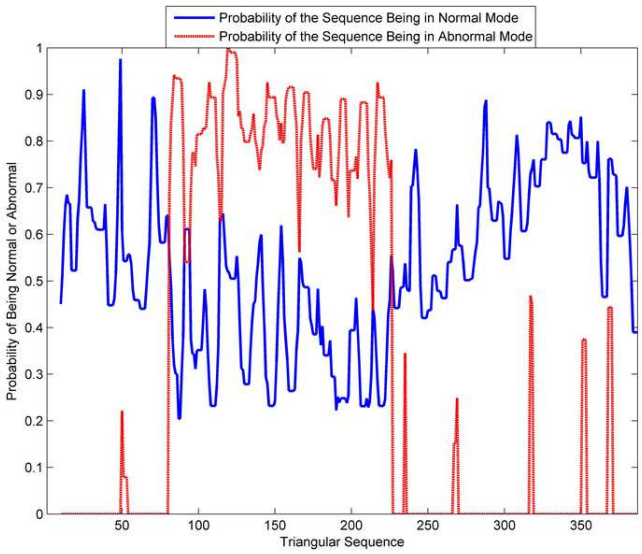

Figure 5.42: Normalized probability of the observation sequences in Figure 5.36 belonging to normal and abnormal regions using the proposed moving window approach ($N_W = 9, N_{min} = 5$)

5.9 Industrial Case Study

This industrial case study will be presented in Chapter 6, and compared to the methods in other chapters of the thesis.

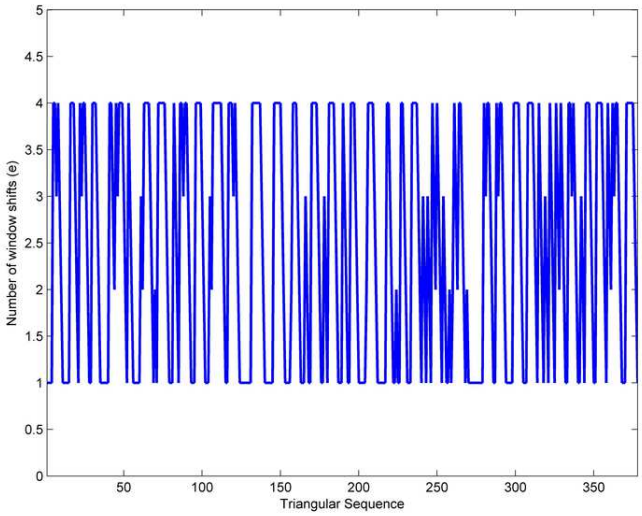

Figure 5.43: Number of shifts in Figure 5.42 to find the optimal window

5.10 Conclusion

In this chapter, a new method for overall classification of the process operation status based on hidden Markov models is introduced. First, means and variances of the different modes for the magnitudes and durations of the triangles and the parameters of the Markov chain are calculated using the EM algorithm. Next, applying Hamiltons filter, posterior probability of each state given the new magnitude and duration is calculated. Using this posterior probability, the fuzzy membership functions are weighted adaptively to provide an accurate discretization based on the method of triangular representation. A multivariate HMM scheme, with moving a window of observations, is also implemented for the overall classification of the process status.

The adaptive fuzzification strategy proposed in this chapter provides more accurate discrete observations when using the method of triangular representation by considering different modes for the durations and magnitudes. Furthermore, applying a multivariate scheme to train HMMs for multiple discrete observations automatically considers the transition behavior of different variables and the number of false alarms is reduced. Application of the proposed method in simulation and industrial case studies shows that it is a promising approach to detect the normal and abnormal operations of the process.

Chapter 6

Applications to an Industrial Scale Oil Sands Primary Separation Vessel

In this chapter, it is shown that a combination of semi-empirical equations and data driven methods provides an appropriate solution for estimation of the required critical velocity to avoid sand deposition and line plugging in underflow of an industrial scale Primary Separation Vessel (PSV). Sections 6.1 to 6.5 provide an overview of the process, the proposed strategy and results of on-line implementation. In Sections 6.6 and 6.7, the proposed methods of Chapters 3 and 5 are tested on the historical data of the PSV, and the deficiencies of data driven methods to address this complex industrial problem are explained. Finally, in Section 6.8, we draw the final conclusion of the thesis.

6.1 Introduction

6.1.1 Problem Statement

Most of previous studies on the required critical minimum velocity to move solid beds inside the slurry lines are related to the process design step where the dynamics of

Short version of a part of this chapter has been published in N. Sammaknejad, B. Huang, R. S. Sanders, Y. Miao, F. Xu, A. Espejo (2015). Adaptive Soft Sensing and On-line Estimation of the Critical Minimum Velocity with Application to an Oil Sand Primary Separation Vessel. Proceedings of the IFAC 9^{th} International Symposium on Advanced Control of Chemical Processes (ADCHEM). Whistler, Canada [124].

Complete version of a part of this chapter is to be submitted as N. Sammaknejad, B. Huang, R. S. Sanders, Y. Miao, F. Xu, A. Espejo. Adaptive Prediction of Critical Minimum Velocity of Slurry Flow with Application to an Oil Sand Primary Separation Vessel. Journal of Process Control.

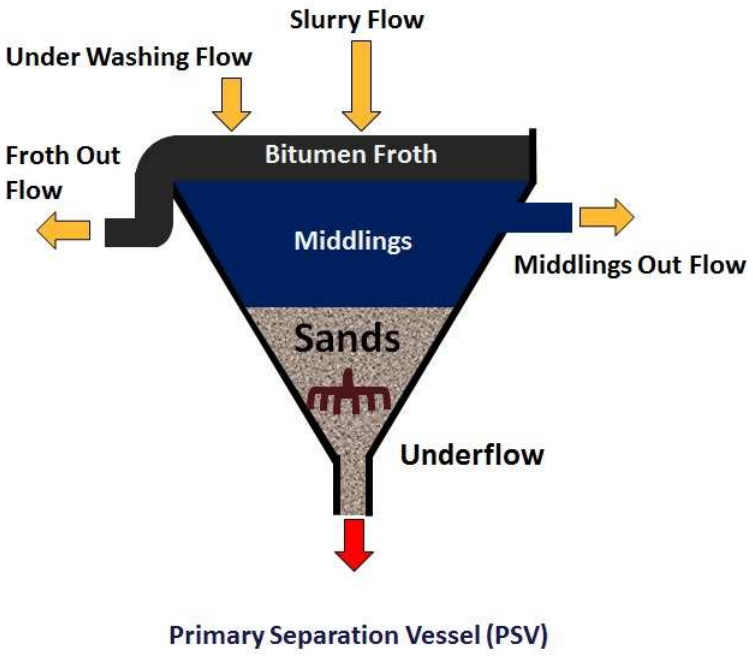

Figure 6.1: Three layers of the PSV unit

the process are not thoroughly considered. Here, a general framework for on-line estimation of the critical minimum velocity is proposed and applied to the underflow of a Primary Separation Vessel (PSV) in oil sands industry. A probabilistic framework is introduced to treat missing observations. Statistical methods are used to design a soft sensor and modify the predictions. Results are compared with other data driven methods.

6.1.2 Process Overview

The industrial application of this chapter is critical minimum velocity estimation in the underflow of an industrial scale PSV in oil sands industry. The PSV unit is presented in Figure 6.1.

The PSV unit is a large settling vessel to separate the feed into three different streams. The slurry feed, which includes aerated bitumen aggregates, water, coarse sand and fine solids, enters at the center of the unit. The bitumen floats over a weir circling the top for further froth treatment. Coarse sand particles settle to the bottom and form the underflow stream. A third outlet stream, which usually contains fines, bitumen aggregates and water, is removed from the middle of the vessel and referred to as the middlings. Both middlings and underflow streams are transferred to secondary recovery units using two variable speed pumps through two different

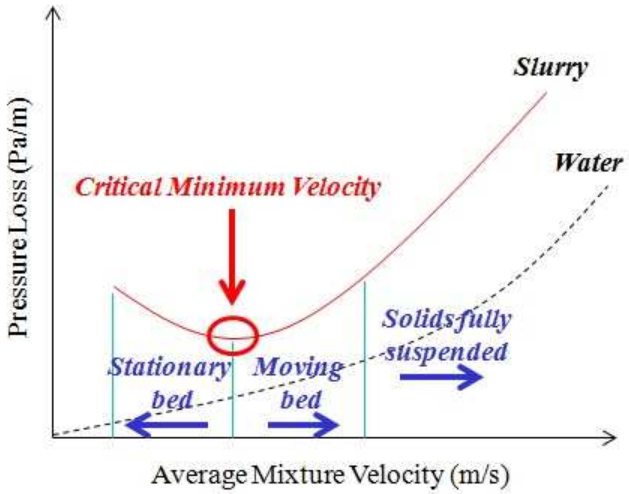

Figure 6.2: pressure loss verses average mixture velocity inside the slurry line

process lines.

6.1.3 Definition of the Critical Minimum Velocity

The critical minimum velocity, also known as the deposition velocity, is the operational velocity at which the stationary bed of solids first forms. The pressure loss inside the line ($\frac{Pa}{m}$) verses average mixture velocity ($\frac{m}{s}$) for water and the slurry flow is presented in Figure 6.2 ([104]).

In Figure 6.2, one could observe that in the case of having water inside the process line, the pressure loss increases while raising the average mixture velocity. However, in the case of slurry flow, there is a velocity below which the particles start to make a stationary bed, and right above it, the bed of particles starts to move. This critical point, where the pressure loss inside the slurry line becomes the minimum, is known as the critical minimum velocity or the deposition velocity.

6.1.4 Importance of the Critical Minimum Velocity

Since the underflow stream in the PSV unit usually contains coarse sand particles, there is a concern of sand deposition and line plugging. Complete plugging of the line, which occurs at flow-rates below the critical velocity, is referred to as the "sanding" phenomena in oil sands industry. In addition to the sanding phenomena, operating velocities below the critical minimum velocity will cause excessive erosion in the lower part of the line ([106]).

On the other hand, operating velocities greater than the critical minimum velocity are uneconomical as more pump power will be required ([106]).

On-line estimation of the critical minimum velocity and comparison with the current operating velocity will provide a lower limit for the operator to avoid near sanding as well as sanding regions. Also, it will help to avoid conservative high flow rate operations in the underflow stream, and therefore, improve PSV bitumen recovery.

6.1.5 Solution Strategy

Unlike previous applications where the critical minimum velocity equations are only used in the design step, in this chapter, a novel approach for on-line estimation of the critical minimum velocity with application to the underflow stream of the PSV unit is introduced. When the on-line estimation becomes greater than the current operating velocity, a near-sanding alarm is generated.

The proposed solution strategy is as follows:

First, the appropriate semi-empirical equation for on-line estimation of the critical minimum velocity is selected. Next, the effective variables for the estimation are obtained. Since one of the key variables, carrier fluid density, is difficult to measure on-line, a soft sensor is developed to provide a parallel on-line measurement for this variable. The recursive Partial Least Squares (rPLS) method is used to develop this soft senor. Also, an adaptive approach based on Hidden Markov Models (HMMs) is used to adaptively change the sensitivity of the critical velocity estimations. Due to the presence of unknown operating modes, the Expectation Maximization (EM) algorithm is used to train the HMM. A procedure to treat the missing observations through the iterations of the EM algorithm is proposed. In the historical data, some observations appear as "Not a Number", or "NaN", and "Error", in the server. The new method automatically considers the effect of the missing observations during the parameter estimation step. Finally, the algorithm is tested in on-line environment through the communication of the Distributed Control System (DCS), OPC server and MATLAB. The solution strategy is summarized in Figure 6.3.

6.2 Critical Minimum Velocity Estimation

Numerous semi-empirical equations have been developed for the purpose of deposition velocity estimation in literature. They are based on both force balance and laboratory analysis. Quality of these models depends primarily on the quality of the experimental data. One of the earliest and most practically used correlations is the

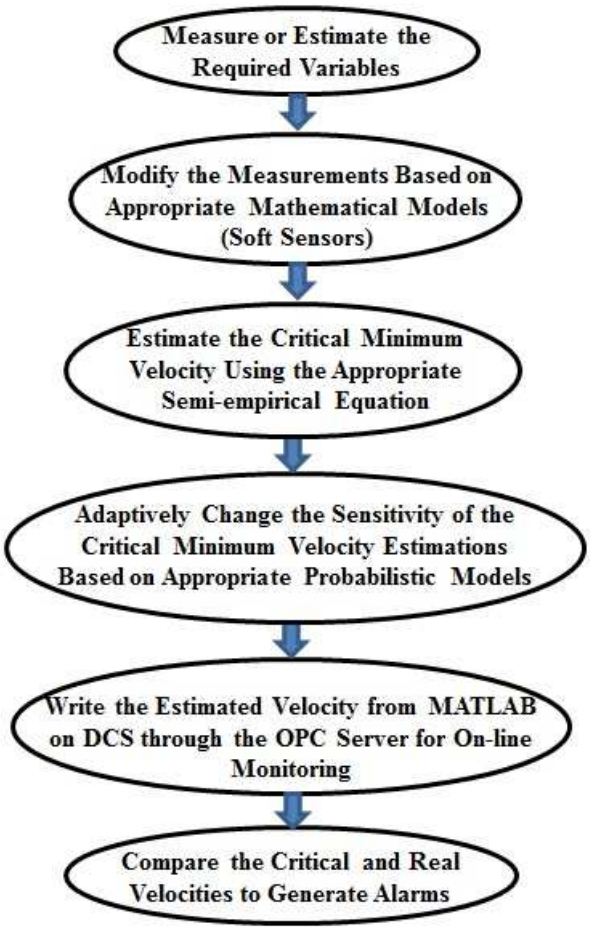

Figure 6.3: Solution strategy for on-line estimation of the critical minimum velocity

critical minimum velocity equation proposed by Durand [107]:

$$V_C = F_L\sqrt{2gD(s-1)} \tag{6.1}$$

where F_L is a constant which is obtained graphically as a function of particle diameter ($0.44 < d < 2.04$ $[mm]$) and volumetric solid concentration ($0.05 < C_v < 0.15$). D is the diameter of the line ($0.04 < D < 0.58$ $[m]$), g is the gravitational constant, and s is the specific gravity of solids. The solid particles used to develop this equation mostly include coal and sand.

Based on a data set containing 864 experimental results, Turian et al. developed another equation for the estimation of the critical minimum velocity [103]. The equation contains a term to represent the effect of the carrier liquid viscosity.

$$\frac{V_C}{[2gD(s-1)]^{0.5}} = X_1 C_v^{X_2}(1-C_v)^{X_3}\{\frac{D\rho[gD(s-1)]^{0.5}}{\mu}\}^{X_4}(\frac{d}{D})^{X^5} \tag{6.2}$$

where μ is the carrier liquid viscosity and ρ is the carried liquid density. X_i's are the coefficients obtained from the regression analysis. Although a large data base was used in the development of this equation, only a few data for large particles and lines were involved and the effect of particle size was not thoroughly considered [108].

As explained in [104], Shook et al. developed a correlation between the Archimedes and Froude numbers to estimate the critical minimum velocity. The Archimedes number is defined as the ratio of the gravitational forces to the viscous forces on a particle, i.e.,

$$Ar = \frac{4}{3}\frac{gd^3(S-1)\rho_f^2}{\mu_f^2} \tag{6.3}$$

where μ_f is the carrier fluid viscosity, ρ_f is the carrier fluid density and S is the ratio of the solid density to carrier fluid density $\frac{\rho_s}{\rho_f}$.

The Froude number is defined as the ratio of the kinetic forces to the gravitational forces on particle:

$$Fr = \frac{V_C}{\sqrt{gD(S-1)}} \tag{6.4}$$

Shook et al. introduce the Archimedes number as an independent variable and provide the following relation between the Archimedes and Froude numbers ([109], [110]):

$$540 < Ar,\ Fr = 1.78Ar^{-0.019} \tag{6.5}$$

$$160 < Ar < 540,\ Fr = 1.19Ar^{0.045}$$

$$80 < Ar < 160,\ Fr = 0.197Ar^{0.4}$$

This equation, which is known as the SRC (Saskatchewan Research Council) equation, is based on a large data-base with properties similar to the underflow of the PSV unit. It is developed based on high quality experimental data and is applicable to the turbulent flows and a variety of line diameters from 0.05 to 0.5 $[m]$. Combination of all these properties makes the SRC equation an appropriate choice for the industrial application of this chapter.

There are other similar studies in literature to develop appropriate semi-empirical equations for specific conditions ([105], [111]).

The key variables for estimation of the critical minimum velocity using the SRC equation are as follows:

6.2.1 Carrier Fluid Density

Carrier fluid is a portion of the slurry which contains particles with diameter less than 44 $[\mu m]$ [112]. In the PSV unit, as presented in Figure 6.1, coarser particles are usually dragged to the tailings stream while fines and bitumen aggregates enter the middlings stream. Therefore, the middlings stream provides a good indication to the carrier fluid properties. In this study, middlings density, which is measured through an on-line analyzer, is used as an indication to the carrier fluid density in the SRC equation.

6.2.2 Carrier Fluid Viscosity

Carrier fluid viscosity is known to be a function of carrier fluid solid concentration in fluid particle systems literature. Having the value of the carrier fluid density as explained in the previous section, the carrier fluid solid concentration can be obtained from Equation 6.6:

$$\rho_f = C_f \rho_s + (1 - C_f)\rho_l(T) \tag{6.6}$$

where C_f is the carrier fluid solid concentration, ρ_s is the density of the solid phase, which is often selected as 2650 $[\frac{kg}{m^3}]$ in oil sands industry as an average, and $\rho_l(T)$ is the density of the liquid phase (water is the dominant component) as a function of the PSV temperature (T). $\rho_l(T)$ is obtained as follows [113]:

$$\rho_l(T) = a_5[1 - \frac{(T + a_1^2)(T + a_2)}{a_3(T + a_4)}] \tag{6.7}$$

where $a_1 = -3.983035°C$, $a_2 = 301.797°C$, $a_3 = 522528.9°C^2$,$a_4 = 69.34881°C$, and $a_5 = 999.974950$ $[\frac{kg}{m^3}]$.

Having the carrier fluid solid concentration, various semi-empirical correlations

exist in literature to find the carrier fluid viscosity. Among them, our investigation on the historical data of the PSV shows that Equation (6.8), which is developed in conditions close to the middlings stream [114], is appropriate for this industrial application. The linear behavior of this equation provides smooth estimations for the critical minimum velocity.

$$\mu_f = \mu_l(T)(1 + 14.7C_f) \tag{6.8}$$

where $\mu_l(T) = A \times 10^{\frac{B}{(T-C)}}[Pa.s]$, with $A = 2.414 \times 10^{-5}[Pa.s]$, $B = 247.8K$ and $C = 140K$. $\mu_l(T)$ is the viscosity of the liquid phase (water is the dominant component) as a function of temperature (T) in Kelvin [115].

6.2.3 Coarse Particle Diameter

Coarse particle diameter plays an important role in estimation of the critical minimum velocity. Intuitively, coarse particle diameter should have a positive correlation with the ratio of volumetric concentration of coarse solids to fines in the mixture as in Equation 6.9. Similar intuitive correlations have been previously explained in literature [116].

$$d \propto X = \frac{C_{mix} - C_{fines}}{C_{fines}} \tag{6.9}$$

where d is the coarse particle diameter, C_{mix} is the volumetric concentration of solids in the mixture and C_{fines} is the volumetric concentration of fines in the mixture.

In Equation 6.9, one could observe that in the case of having no coarse particles in the mixture ($C_{mix} = C_{fines}$), the X factor becomes zero. When the ratio of $(\frac{C_{mix}}{C_{fines}})$ starts to increase, i.e., there are more coarse particles in the mixture, the X factor starts to grow. Since d is the median of all available particle diameters in the mixture, it is expected to have a positive correlation with the X factor.

As previously stated in the introduction section, fines and bitumen aggregates usually enter the middlings stream while the coarser particles tend to go directly to the tailings. Therefore, the X factor in Equation 6.9 can be written as

$$X = \frac{C_{und} - C_{mid}}{C_{mid}} \tag{6.10}$$

where "und" and "mid" refer to underflow and middlings streams respectively.

In this study, the positive correlation between coarse particle diameter and the X factor is assumed to be linear. More complicated correlations will be a subject

of future studies. Having the minimum and maximum value of the coarse particle diameter [117], the minimum and maximum value of the X factor from historical data, and the linearity assumption, the on-line estimation of coarse particle diameter can be calculated as in Equation 6.11.

$$\frac{X - X_{min}}{X_{max} - X_{min}} = \frac{d - d_{min}}{d_{max} - d_{min}} \tag{6.11}$$

Since this equation is subject to many uncertainties, it is only applied when the X factor is within two standard deviations of X_{mean}. Otherwise the coarse particle diameter is assumed to be constant (d_{mean}).

6.3 Soft Sensor Development

From previous sections, one could observe that carrier fluid density, which is available through middlings density analyzer, plays a key role in on-line estimation of the critical minimum velocity. However, there are several short periods in the historical data where this on-line measurement is not available. Maintenance of the PSV and exceeding the measurement limits are the main reasons of such circumstances. During such periods, the data which appears in DCS represents the lower limit of the online analyzer, and not the true value. As a result of such situations, the Archimedes number suddenly decreases. This results in a sudden spike, and a false alarm in on-line estimation of the critical minimum velocity.

Consequently, providing another on-line measurement parallel to the density on-line analyzer will help to avoid such false alarms and malfunctions in the case of on-line analyzer failure (see Figures 6.6 and 6.7 for more information).

Lab data for the middlings density is available every two hours. Therefore, if meaningful correlations exist between other process variables and the middlings density lab data, it will be possible to develop a mathematical model (soft sensor) and provide another parallel measurement for the middlings density.

Due to frequent challenges in flow-rate measurement, it is difficult to develop first-principle models for this soft sensor. However, data-driven approaches provide acceptable results. Density of the middlings stream is strongly correlated with the density of other layers, e.g., feed, froth and the underflow. Therefore, linear regression techniques like Partial Least Squares (PLS) provide appropriate data-driven models to solve this problem. However, since the PSV unit shows a time varying behavior according to the historical data (working conditions of the PSV might change due to the changes in the feed properties) model updating is necessary.

6.3.1 Recursive Exponentially Weighted PLS (rPLS)

Many different approaches have been reported in literature in order to update the model in the on-line application. One of the simplest ways, known as the model coefficients recalculation, is to add the new samples to the training data-set, and re-identify the model so that the model is able to adapt to new operating conditions [118]. But, such methods will cause some delay in model updating. Furthermore, it will be necessary to assign more weights to the new samples when working on large data bases. Successful applications of the recursive Partial Least Squares (rPLS) method in industrial processes are reported in literature [119]. Unlike the model coefficients recalculation methods, the rPLS method significantly weights every new sample to the data-base and continuously updates the model. Therefore, the model will more rapidly adapt to new process conditions.

The rPLS method used in this chapter is based on the study on the improved PLS kernel algorithm [121]. In this updating strategy, a forgetting factor is used to exponentially discount the past data and takes into account the effect of the recent observations [120]. The procedure of covariance matrix updating is as follows [122]:

$$R_{xx}(t) = \lambda R_{xx}(t-1) + \tilde{x}(t)^T \tilde{x}(t) \tag{6.12}$$

$$R_{xy}(t) = \lambda R_{xy}(t-1) + \tilde{x}(t)^T \tilde{y}(t)$$

where the forgetting factor ($0 \leq \lambda \leq 1$) reflects the rate of discounting the old data.

The mean centered data ($\tilde{x}(t)$ and $\tilde{y}(t)$) for the new available inputs and outputs are obtained as in Equation 6.13.

$$\tilde{x}(t) = x(t) - \bar{x}(t) \tag{6.13}$$

$$\tilde{y}(t) = y(t) - \bar{y}(t)$$

where the mean vectors are updated as follows [123]:

$$\bar{x}(t) = \frac{N-1}{N}\bar{x}(t-1) + \frac{1}{N}x(t) \tag{6.14}$$

$$\bar{y}(t) = \frac{N-1}{N}\bar{y}(t-1) + \frac{1}{N}y(t)$$

$R_{xx}(0)$ and $R_{xy}(0)$ are the initial covariance matrices for the historical input and output mean centered data (X, y), i.e.,

$$R_{xx}(0) = X^T X \tag{6.15}$$

$$R_{xy}(0) = X^T y$$

N is the length of the data in X and y.

When the new covariance matrices are available from Equation 6.12, the regression coefficients (b) are obtained following a fast kernel PLS calculation. Details of the method can be found in literature [121]. Having the regression coefficients available, the mean centered final prediction ($\hat{y}_t$) in on-line application is obtained as,

$$\hat{y}_t = bX_t$$

where X_t is the mean centered vector of input variables at time t.

Results of the rPLS algorithm will be compared to the fixed PLS algorithm in the Results Section. One could observe that using the updating rule in Equation 6.12, the model is able to adapt to new operating modes and conditions and the recursive method shows a superior performance in comparison to the fixed method. Outlier removal based on the 3σ rule and data smoothing based on a moving average filter are used in both training and test steps for the soft sensor inputs.

6.4 Adaptive Sensitivity Levels for Critical Velocity Estimation

The main idea of the work presented in this section is to adaptively change the sensitivity of the critical velocity estimations according to the operating mode of the process. Consequently, more sensitive predictions will be generated when the process is operating more abnormally and the prediction sensitivity decreases when the process is in normal operating condition.

In order to avoid false alarms and provide more sensitive predictions, it is necessary to adaptively select the $K(t)$ value in Equation 6.16.

$$Q_S(t) = Q_C(t) + K(t) \times \sigma_{Q_C} \tag{6.16}$$

where $Q_S(t)$ is the sensitive estimation of $Q_C(t)$ (critical flow-rate) based on the current operating mode of the real flow-rate at time t (F_t), and σ_{Q_C} is the standard deviation of the critical flow-rate estimation error from the historical data obtained from a Monte Carlo simulation. Note that, having the diameter of the line, velocity can be converted to flow-rate.

The lower and upper bounds of the K value ($K_{L/U}$) in Equation 6.16 can be obtained by solving the optimization problem in Equation 6.17 based on different sensitivity values, e.g., use $\alpha_L = 0.7$ to find K_L and $\alpha_U = 1$ to find K_U, etc.

$$K_{L/U} = argmin_K \|Q_S - \alpha_{L/U} F_{Normal}\| \tag{6.17}$$

where F_{Normal} is the vector of normal flow rates of the process from historical data.

The historical data for the underflow flow rate can be divided to three operating modes ($I_t's$), i.e., mode 1 ($I_t = 1$) is low flow rate, mode 2 ($I_t = 2$) is average flow rate and mode 3 ($I_t = 3$) is high flow rate. In this work, mode 1 (near sand deposition and plugging) and mode 3 (impact on the bitumen recovery) are considered as upset operations. In order to avoid such regions, the $K(t)$ value will be adaptively selected according to Equation 6.18 as follows:

$$\frac{K(t) - K_L}{K_U - K_L} = 1 - P(I_t = 2|F_t, ..., F_0) = P(Upset\ Modes) \tag{6.18}$$

where $F_t, ..., F_0$ are the underflow flow-rate observations from time 0 to t, and $P(Upset\ Modes)$ is the probability of the upset operating modes to occur.

Adaptive selection of the K value according to Equation 6.18 will increase the sensitivity of the estimations in the upset operating modes, while reducing the sensitivity in the normal modes. Using this adaptive technique, the number of false alarms will be greatly reduced. See Figures 6.8 and 6.9 for more information.

6.4.1 Flow rate Mode Diagnosis

In this section, the problem of calculating the probability of the current operating mode given flow rate observations ($P(I_t|F_t, ..., F_0)$) is addressed. It is assumed that operating modes of the flow rate can transit to each other following a Markov chain model with mean values, variances, state transition probabilities and initial state distributions given as μ_i, σ_i^2, α_{ij} and π_i where i and j ($1 \leq i, j \leq M = 3$) are indicators of the operating mode ($I_t = i, j,\ 1 \leq i, j \leq M$). The training procedure to obtain these parameters will be explained in the next section.

In order to calculate $P(I_t|F_t, ..., F_0)$, Hamilton's filtering strategy is used to infer the operating mode of the process for the on line diagnosis application [46]. Probability of the hidden operating mode given flow rate observations is calculated as in Equation 6.19.

$$P(I_t|F_t, ..., F_0) = \sum_{I_{t-1}=1}^{M} P(I_t, I_{t-1}|F_t, ..., F_0) \tag{6.19}$$

where M is the number of available operating modes ($M = 3$ in Equation (6.18) for this study).

The following expression provides the joint probability of the states I_t, I_{t-1} given the information up to time $t - 1$:

$$P(I_t, I_{t-1}|F_{t-1}, ..., F_0) = P(I_t|I_{t-1})P(I_{t-1}|F_{t-1}, ..., F_0) \tag{6.20}$$

$P(I_{t-1}|F_{t-1},,F_0)$ is the output of the filter at the previous sample time starting from the initial state distribution of the Markov chain model (π), and $P(I_t|I_{t-1})$ is the transition probability (α_{ij}) obtained from the Markov assumption of the model. The joint probability in (6.20) is updated when having a new observation at time t, i.e.,

$$P(I_t, I_{t-1}|F_t, ..., F_0) = \frac{P(F_t, I_t, I_{t-1}|F_{t-1}, ..., F_0)}{P(F_t|F_{t-1}, ..., F_0)} \tag{6.21}$$

where

$P(F_t, I_t, I_{t-1}|F_{t-1}, ..., F_0) = P(F_t|I_t, I_{t-1}, F_{t-1}, ..., F_0) \times P(I_t, I_{t-1}|F_{t-1}, ..., F_0)$.

$P(I_t, I_{t-1}|F_{t-1}, ..., F_0)$ is obtained previously from Equation 6.20 and $P(F_t|I_t, I_{t-1}, F_{t-1}, ..., F_0) = P(F_t|I_t)$ is calculated using the Gaussian distribution assumption ($(F_t|I_t = i; \mu_i, \sigma_i) \sim N(\mu_i, \sigma_i^2),\ i = 1, ..., M$). Finally, the denominator in Equation 6.21 is determined as follows:

$$P(F_t|F_{t-1}, ..., F_0) = \sum_{I_t=1}^{M} \sum_{I_{t-1}=1}^{M} P(F_t, I_t, I_{t-1}|F_{t-1}, ..., F_0) \tag{6.22}$$

6.4.2 HMM Training

In this section, the procedure of training the HMM to model the transitions of the flow-rate in the presence of missing observations is introduced. Due to the existence of the unknown operating modes and missing observations, the Expectation Maximization (EM) algorithm provides appropriate solutions to this problem. The EM algorithm solves the maximum likelihood estimation problem by iteratively switching between the Expectation (E) and Maximization (M) steps [28].

Hamilton is one of the forerunners in applications of HMMs to infer the current regime of observations [44]. Other researchers have performed similar studies in this area ([47], [49] and [55]). Our recent studies provide a time varying solution to such problems in the presence of both discrete and continuous observations ([45] and [71]).

There has also been a great effort in handling the problem of missing observations recently. Missing data usually occurs due to sensor malfunctions, network connection interruptions and measurement errors [59]. As explained in chapter 3, three types of missing data have been introduced in literature [60]. In the case of missing at random (MAR), the probability of missingness might only depend on the observed section of the data. If the data is missed completely at random (MCAR), the distribution of missingness will not depend on the observed, nor the missing data sets. Finally, in the case of missing not at random (MNAR), probability of missingness will depend on the missing data set. In a very recent article, a novel approach is proposed for adaptive identification of nonlinear processes in the presence of missing observations

[57]. Missing data treatment is based on the MCAR assumption in this article.

Following the proposed strategy of this section, means and variances of the different operating modes of the flow-rate (μ_i's and σ_i^2's), the transition probabilities (α_{ij}'s) and the initial state distribution of the Markov chain model (π), in the presence of missing observations, will be obtained. These parameters have been introduced in the previous section, and will be indicated by $\Theta = \{\mu_i,\ \sigma_i,\ \alpha_{i,j},\ \pi_i\}$, $1 \leq i, j \leq M$ during the parameter estimation.

As previously stated, in the historical data of the flow-rate, some observations appear as "NaN" or "Error". In the training data set, these observations are considered as MCAR missing values. Consequently, the complete data set $F = \{F_1, ..., F_N\}$ will be divided to the observed $F_O = \{F_{t_1}, ..., F_{t_\alpha}\}$, and missing $F_M = \{F_{m_1}, ..., F_{m_\beta}\}$ sections. Obviously, the union of F_O and F_M is F. Different hidden operating modes of the flow-rate, for example, low, medium and high, at different time instants, are presented by $I = \{I_1, ..., I_N\}$. The observed and missing or hidden data sets are presented by $C_{obs} = F_O$ and $C_{mis} = \{I, F_M\}$ respectively.

Flow-rate observations are assumed to follow a normal distribution as follows:

$$(F_t|I_t = i; \mu_i, \sigma_i) \sim N(\mu_i, \sigma_i^2), i = 1, ..., M \tag{6.23}$$

where all the parameters in Equation 6.23 have been defined previously.

In order to model flow-rate transitions between different operating modes, the observations are considered to follow a Markov chain model with parameters given as,

$$\alpha_{ij} = P(I_t = j|I_{t-1} = i),\ i, j = 1, ..., M \text{ and } t = 1, ..., N \tag{6.24}$$
$$\pi_i = P(I_1 = i), i = 1, ..., M$$

In the Expectation (E) step of the EM algorithm the conditional expectation of complete data log likelihood function is calculated, that is,

$$Q(\Theta|\Theta^{old}) = E_{C_{mis}|(\Theta^{old}, C_{obs})}\{logP(C_{obs}, C_{mis}|\Theta)\} \tag{6.25}$$

where the superscript *old* refers to the parameters in the previous iteration of the EM algorithm, and Θ is the set of unknown parameters to be obtained in the maximization step.

In the Maximization (M) step, the set of parameters that maximize the Q-function are calculated:

$$\Theta^{new} = argmax_\Theta Q(\Theta|\Theta^{old}) \tag{6.26}$$

Equation 6.25 can be formulated for the problem of this chapter as follows,

$$Q(\Theta|\Theta^{old}) = E_{I,F_M|(\Theta^{old}, F_O)}\{logP(F_{1:N}, I_{1:N}|\Theta)\} \tag{6.27}$$

According to the chain rule, $P(F_{1:N}, I_{1:N}|\Theta)$ in Equation 6.27 can be written as,

$$P(F_{1:N}, I_{1:N}|\Theta) = P(F_{1:N}|I_{1:N}, \Theta) \times P(I_{1:N}|\Theta) \tag{6.28}$$

where different terms of Equation 6.28 can be separately simplified as follows:

$$P(F_{1:N}|I_{1:N}, \Theta) = \prod_{t=1}^{N} P(F_t|F_{t-1}, ..., F_1, I_{1:N}, \Theta) = \prod_{t=1}^{N} P(F_t|I_t, \Theta) \tag{6.29}$$

$$P(I_{1:N}|\Theta) = \prod_{t=1}^{N} P(I_t|I_{t-1}, ..., I_1, \Theta) = \prod_{t=1}^{N} P(I_t|I_{t-1}, \Theta) = P(I_1) \prod_{t=2}^{N} P(I_t|I_{t-1}, \Theta) \tag{6.30}$$

where Equation 6.29 is obtained based on the Gaussian assumption in Equation 6.23, and Equation 6.30 is obtained based on the Markov property of the model.

From Equations 6.27 to 6.30, and using the properties of log operator, the Q-function can be written as

$$Q(\Theta|\Theta^{old}) = E_{I,F_M|(\Theta^{old},F_O)}\{logP(I_1)\} + E_{I,F_M|(\Theta^{old},F_O)}\{\sum_{t=1}^{N} logP(F_t|I_t, \Theta)\} \tag{6.31}$$

$$+E_{I,F_M|(\Theta^{old},F_O)}\{\sum_{t=2}^{N} logP(I_t|I_{t-1}, \Theta)\}$$

In the first step, expected value of the expression in (6.31) is calculated with respect to the hidden operating mode I, that is,

$$Q(\Theta|\Theta^{old}) = E_{F_M|(\Theta^{old},F_O,I)}\{\sum_{i=1}^{M} P(I_1 = i|\Theta^{old}, F_O) log\pi_i\} \tag{6.32}$$

$$+E_{F_M|(\Theta^{old},F_O,I)}\{\sum_{i=1}^{M}\sum_{t=1}^{N} P(I_t = i|\Theta^{old}, F_O) \times logP(F_t|I_t = i, \mu_i, \sigma_i)\}$$

$$+E_{F_M|(\Theta^{old},F_O,I)}\{\sum_{i=1}^{M}\sum_{j=1}^{M}\sum_{t=2}^{N} P(I_t = j, I_{t-1} = i|\Theta^{old}, F_O) \times log\alpha_{ij}\}$$

Next, the expected value is calculated with respect to the missing observations (F_M) as follows:

$$Q(\Theta|\Theta^{old}) = \sum_{i=1}^{M} P(I_1 = i|\Theta^{old}, F_O) log\pi_i \tag{6.33}$$

$$+\sum_{i=1}^{M}\sum_{t=t_1}^{t_\alpha} P(I_t = i|\Theta^{old}, F_O) \times logP(F_t|I_t = i, \mu_i, \sigma_i)$$

$$+\sum_{i=1}^{M}\sum_{t=m_1}^{m_\beta} P(I_t=i|\Theta^{old},F_O)\times\int P(F_t|\Theta^{old},I_t=i)\times logP(F_t|I_t=i,\mu_i,\sigma_i)dF_t$$

$$+\sum_{i=1}^{M}\sum_{j=1}^{M}\sum_{t=2}^{N} P(I_t=j,I_{t-1}=i|\Theta^{old},F_O)\times log\alpha_{ij}$$

Calculating the integral term in (6.33), it is not difficult to show that the final expression of the Q-function is as follows:

$$Q(\Theta|\Theta^{old})=\sum_{i=1}^{M} P(I_1=i|\Theta^{old},F_O)log\pi_i \tag{6.34}$$

$$+\sum_{i=1}^{M}\sum_{t=t_1}^{t_\alpha} P(I_t=i|\Theta^{old},F_O)\times logP(F_t|I_t=i,\mu_i,\sigma_i)$$

$$+\sum_{i=1}^{M}\sum_{t=m_1}^{m_\beta} P(I_t=i|\Theta^{old},F_O)\times(-\frac{1}{2}log(2\pi\sigma_i^2)-\frac{1}{2\sigma_i^2}((\sigma_i^{old})^2+(\mu_i-\mu_i^{old})^2))$$

$$+\sum_{i=1}^{M}\sum_{j=1}^{M}\sum_{t=2}^{N} P(I_t=j,I_{t-1}=i|\Theta^{old},F_O)\times log\alpha_{ij}$$

In the M-step, the update formulas are obtained by taking the derivative of the Q-function with respect to unknown parameters, and then, setting them zero. Consequently, the mean value and variance will be updated as follows:

$$\mu_i^{new}=\frac{\sum_{t=t_1}^{t_\alpha} F_tP(I_t=i|\Theta^{old},C_{obs})}{\sum_{t=1}^{N} P(I_t=i|\Theta^{old},C_{obs})} \tag{6.35}$$

$$+\frac{\sum_{t=m_1}^{m_\beta}\mu_i^{old}P(I_t=i|\Theta^{old},C_{obs})}{\sum_{t=1}^{N} P(I_t=i|\Theta^{old},C_{obs})}$$

$$(\sigma_i^{new})^2=\frac{\sum_{t=t_1}^{t_\alpha}(F_t-\mu_i^{new})^2P(I_t=i|\Theta^{old},C_{obs})}{\sum_{t=1}^{N} P(I_t=i|\Theta^{old},C_{obs})} \tag{6.36}$$

$$+\frac{\sum_{t=m_1}^{m_\beta}((\sigma_i^{old})^2+(\mu_i^{new}-\mu_i^{old})^2)P(I_t=i|\Theta^{old},C_{obs})}{\sum_{t=1}^{N} P(I_t=i|\Theta^{old},C_{obs})}$$

The optimization problems to find π_i and α_{ij} are constrained by $\sum_{i=1}^{M}\pi_i=1$ and $\sum_{j=1}^{M}\alpha_{ij}=1$ respectively. Therefore, the Lagrange multiplier should be introduced, and the final parameter estimation results are as follows:

$$\alpha_{ij}^{new}=\frac{\sum_{t=2}^{N} P(I_t=j,I_{t-1}=i|\theta^{old},C_{obs})}{\sum_{j=1}^{M}\sum_{t=2}^{N} P(I_t=j,I_{t-1}=i|\theta^{old},C_{obs})} \tag{6.37}$$

$$\pi_i^{new}=P(I_1=i|\theta^{old},C_{obs})$$

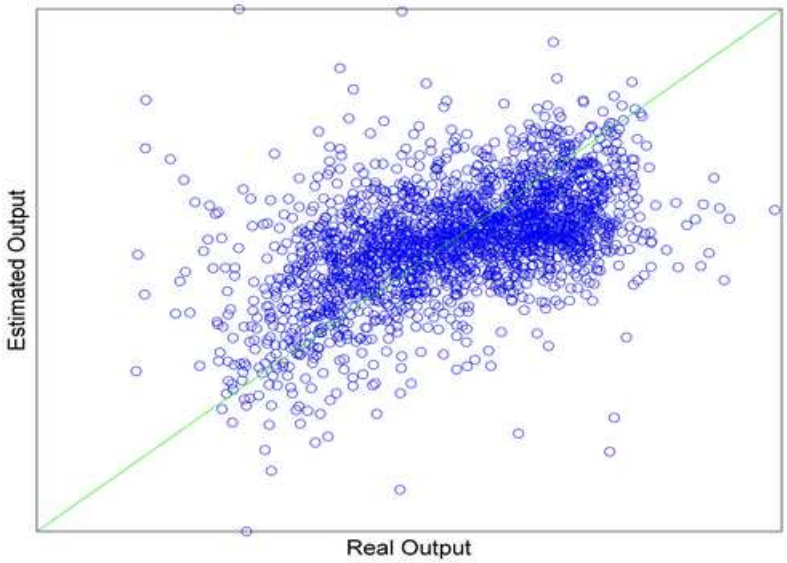

Figure 6.4: Middlings density ($\frac{g}{cm^3}$) soft sensor estimations verses lab data (scatter plot)

The procedure to find posterior distributions $P(I_t = j, I_{t-1} = i|\theta^{old}, C_{obs})$ and $P(I_t = i|\Theta^{old}, C_{obs})$ in the presence of missing observations is available in Chapter 3. The training algorithm is iteratively repeated until a certain convergence criterion is satisfied.

6.5 Results of the Proposed Method

In all the industrial data of this section, the Y axis is masked for proprietary consideration.

6.5.1 Soft Sensor Performance

In this section, results of the middlings density soft sensor are presented. They are compared to the soft sensor with fixed parameters (rPLS versus PLS).

Figures 6.4 and 6.5 illustrate the results of the soft sensor predictions in 2012 historical data (scatter plot and time trend). From these figures, one could observe that the soft sensor is able to track the trend of the lab data well.

Comparison of the results between the fixed and recursive soft sensors is presented in Table 6.1. The model performance is evaluated by Root Mean Square Error of

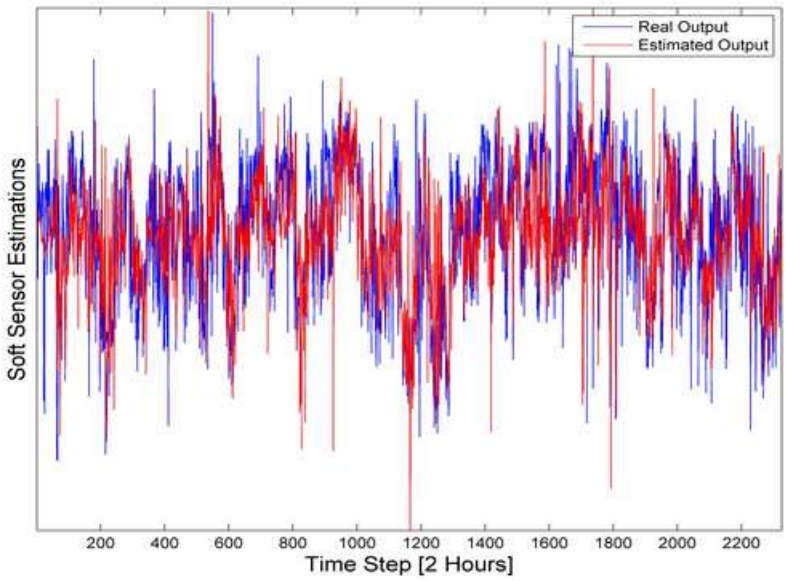

Figure 6.5: Middlings density ($\frac{g}{cm^3}$) soft sensor estimations verses lab data (time trend plot)

Prediction (RMSEP) and the correlation coefficient R.

Table 6.1: Comparison between the performance of the fixed and recursive PLS soft sensors

Soft Sensor	PLS	rPLS
RMSEP	0.2119	0.1054
R	0.4273	0.6077

The current results for the rPLS soft sensor satisfy the need to have a parallel measurement for the middlings density on line analyzer. This parallel measurement will help to avoid false alarms and sudden spikes in the predictions of the critical velocity due to unavailability of the on line analyzer as explained in Section 6.3. Figure 6.6 presents a case of the on line analyzer unavailability. Figure 6.7 shows how the results have been improved after having a parallel measurement from the soft sensor.

6.5.2 Critical Velocity Estimation

In this section, results of both on-line testing and off-line verification for operating mode diagnosis and estimation of the critical velocity are presented. Since it is not

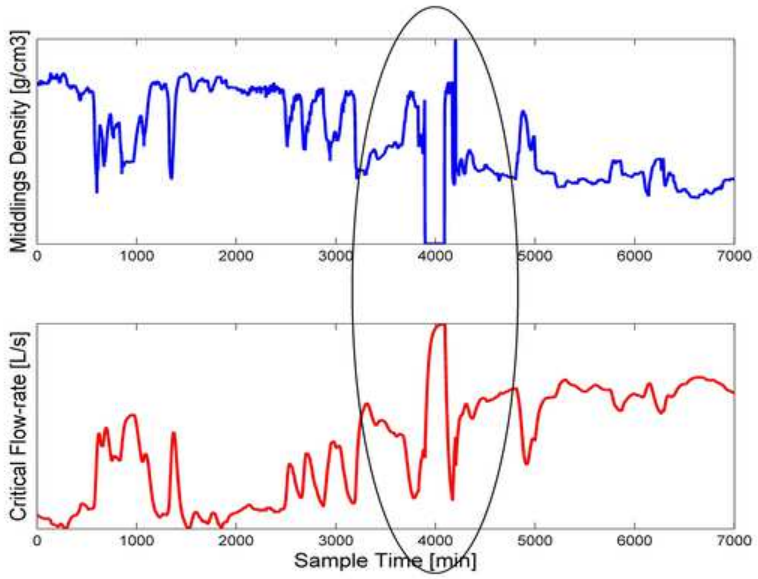

Figure 6.6: Sudden spike in the prediction of the critical minimum velocity due to the on-line analyzer unavailability

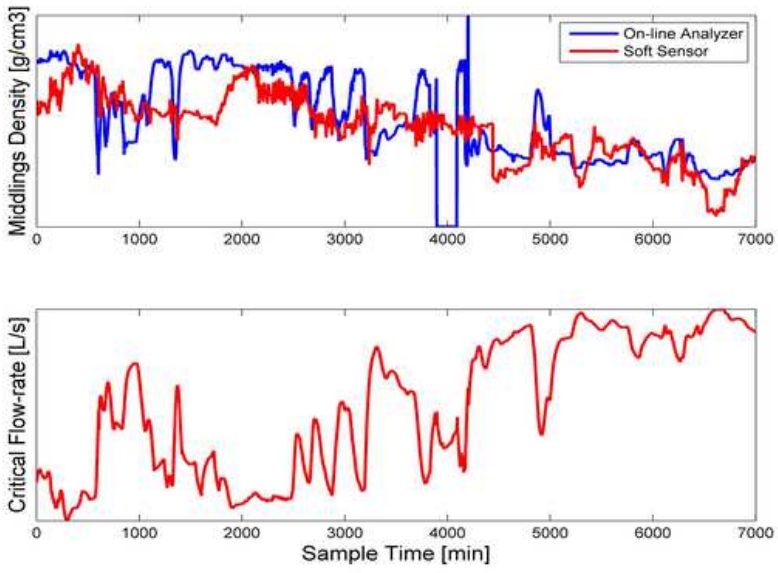

Figure 6.7: Modified critical velocity estimation results in the case of analyzer unavailability

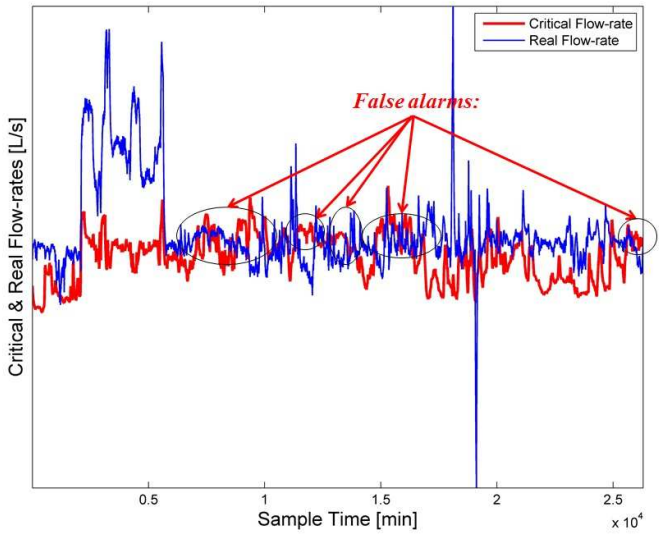

Figure 6.8: On-line estimation of the critical velocity without adaptive sensitivity levels

possible to present the results for all the historical data, only some special cases which contain a combination of normal and upset operating regions are illustrated.

Examples of the critical velocity estimation in the historical data of the PSV, without and with adaptive sensitivity levels, are presented in Figures 6.8 and 6.9 respectively. In these figures real (blue line) and the critical (red line) flow rates are compared. One could observe that by applying time varying sensitivity levels the number false alarms in the normal process operation are greatly reduced. For the case of abnormal operation, estimations are provided with a high sensitivity.

A case of combination of normal and upset regions which has occurred in 2013 is presented in Figure 6.10.

Figure 6.10 presents two cases of upset operations which have occurred in the historical data. In both cases, the operator has increased the flow rate to avoid sand deposition in the underflow. However, it can be observed that in the first abnormal operation, the flow-rate has been increased very conservatively. This might introduce bitumen loss in the process. Increasing velocities that are close to the critical velocity are usually sufficient to avoid near sanding regions.

Operating modes of the flow rate to provide critical velocity estimations with adaptive sensitivities in Figure 6.10 are presented in Figure 6.11.

In Figure 6.11, different operating modes of the flow-rate based on the filtering algorithm introduced in Section 6.4.1 are presented. As previously mentioned in Sec-

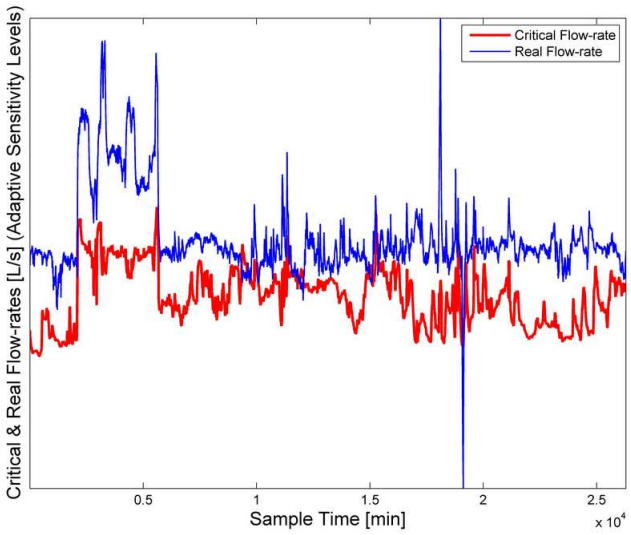

Figure 6.9: On-line estimation of the critical velocity after applying adaptive sensitivity levels

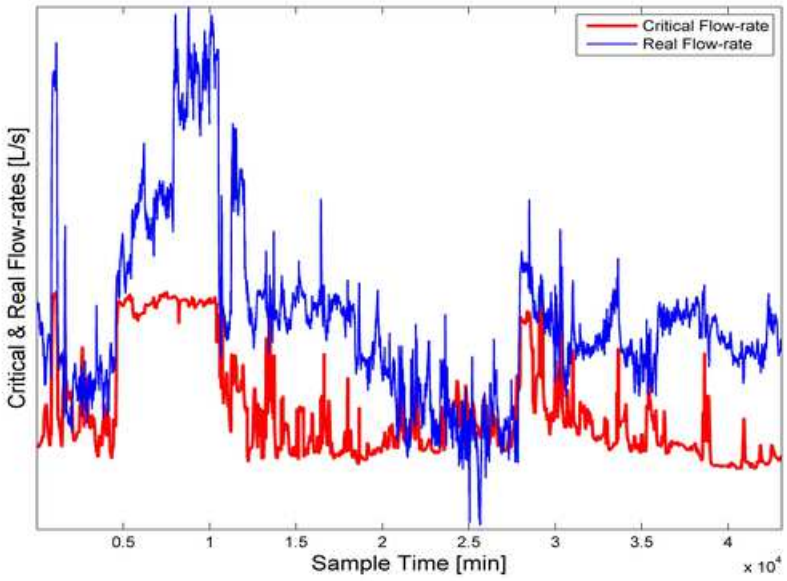

Figure 6.10: A case of upset operation in 2013 data-set

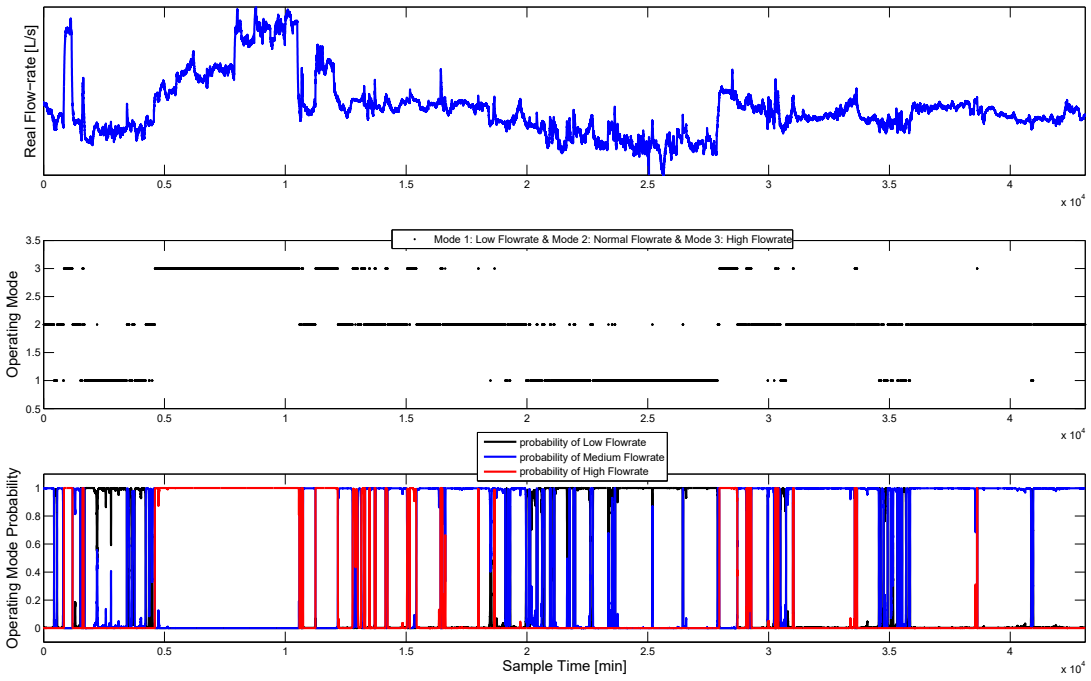

Figure 6.11: Flow-rate operating modes for the data in Figure 6.10

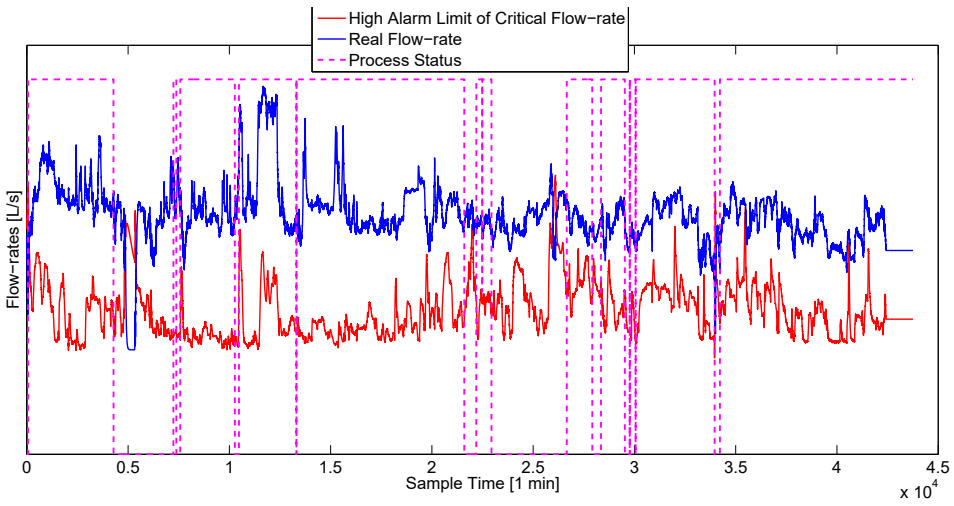

Figure 6.12: Results of on-line testing of the algorithm in 2014 data set

tion 6.4, adaptive modification of estimation sensitivity based on the operating mode will greatly reduce the number of false alarms.

Figure 6.12 illustrates the results of the on-line testing of the algorithm in 2014. During this period underflow flow rate was high (almost twice as much as the previous historical data) due to changes in process condition. Therefore, critical velocity estimations are generated with the highest sensitivity level, which is, the real estimation plus three standard deviations of the estimation error. The dashed line represents the process status where zero indicates process shut down and one indicates process in operation. It is observed that sudden changes in process status are usually accompanied by spikes in the estimation of the critical velocity.

The alarm report for time intervals where the critical flow rate crosses the real flow rate shows that a number of alarms have been generated in such periods. At the beginning, some alarms occur in the underflow density. Next, alarms appear in the underflow pump current, indicating that there are some obstacles in moving the solid bed.

The illustrated results in this section provide a potential of the proposed method for monitoring of the PSV underflow. Since on line estimation of the critical velocity is a combination of several influential variables, it will provide a summary of the status of the effective variables on line, and assist on site operators to maintain normal operating conditions.

6.6 PSV Operating Mode Diagnosis Based on Time Varying HMMs

In this section, the proposed method of Chapter 3, i.e., operating mode diagnosis based on HMMs with time varying transition probabilities is applied on the PSV unit.

As previously stated, the critical minimum velocity is a function of different variables including carrier fluid's density, carrier fluid's viscosity, coarse particle volumetric fraction, coarse particle diameter, etc. The middlings stream is a good indicator for the carrier fluid properties. On the other hand, the tailings stream, which usually contains coarse particles, is a good indicator for the depositing material's properties. Thus, properties of these two layers play the key roles in sanding detection.

Density of the middlings and underflow streams can be measured through on-line analyzers. In the previous sections of this chapter, it is observed that all the other necessary variables which can impact the critical minimum velocity are directly, or indirectly, functions of these two variables. Consequently, the densities might be directly observed to infer the operating condition of the process and avoid sanding conditions. In this section, these two variables are selected as the sanding indicators. Similar to Section 6.4, the tailings flow-rate, which can provide some pre-indication to the operating condition of the proces, is selected as the scheduling variable.

When the process approaches an upset operating condition, the underflow density starts to gradually increase. Since the tailings flow-rate is in closed loop with underflow density, when the underflow density exceeds some high limit, the pump RPM starts to increase to remove the deposited sand and return the operation to its normal condition. This causes a sudden decrease in the middlings and underflow densities. If in the early stage of such circumstances the operators can be notified about the operating mode of the process, they can add water through the cone flush water (Figure 6.1) to assist the suspension of solid deposits and avoid complete sanding of the line. An example of an upset operating region which has occurred in the past is presented in Figure 6.13.

From Figure 6.13, one could observe that the abnormal (intermediate) and beginning of the upset operating conditions, which are denoted by a red circle, are close to each other and far from the normal operation of the process. Existence of such asymmetric behavior provides a good example to show the merit of using time varying over fixed transition probabilities.

In this industrial case study, 8 cases of upset operating conditions similar to Fig-

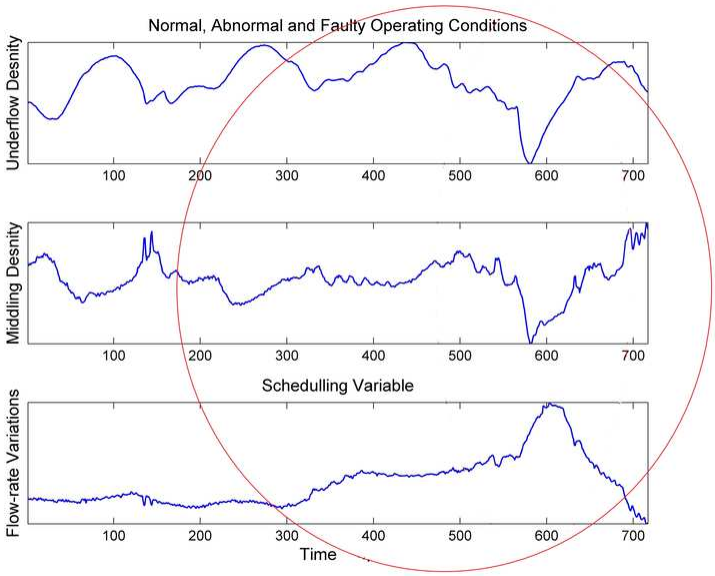

Figure 6.13: A case of upset operating condition in the primary separation vessel from historical data

ure 6.13, which have occurred during a year in the historical data, are selected for training purposes. A case of an upset operating region which has occurred in a month of a different year is selected as the test data-set. During the on-line measurement process, some observations appear as "Not a Number" or "NaN" value in the server. These observations are treated as the "missing observations" in this case study.

Following the proposed estimation method of Chapter 3, the parameters are obtained as in Table 6.2.

Table 6.2: Estimated parameters for the industrial case study using the EM algorithm

$$\pi_0 = [0.3333, 0.3333, 0.3333]$$
$$\gamma_{11} = 0.9934, \gamma_{21} = 0.1987, \gamma_{22} = 0.9887, \gamma_{33} = 0.9929,$$
$$\mu_1 = [1.4483\ 1.4965], \mu_2 = [1.2757\ 1.5246], \mu_3 = [1.1490\ 1.4227]$$
$$\Sigma_1 = \begin{pmatrix} 0.0022 & 0.0003 \\ 0.0003 & 0.0013 \end{pmatrix}$$
$$\Sigma_2 = \begin{pmatrix} 0.0041 & -0.0018 \\ -0.0018 & 0.0032 \end{pmatrix}$$
$$\Sigma_3 = \begin{pmatrix} 0.0043 & 0.0014 \\ 0.0014 & 0.0085 \end{pmatrix}$$
$$\sigma_{H_1} = 4.681 \times 10^3, \sigma_{H_2} = 1.257 \times 10^3, \sigma_{H_3} = 3.7051 \times 10^4$$

The validation data-set is presented in Figure 6.14. The upset behavior occurs at the time period around 12000 where the densities suddenly start to decrease due to

the reaction of the pump to an abnormal event (red oval). Results of the operating mode recognition based on the proposed method in Chapter 3 using Hamilton's filtering algorithm are presented in Figure 6.15.

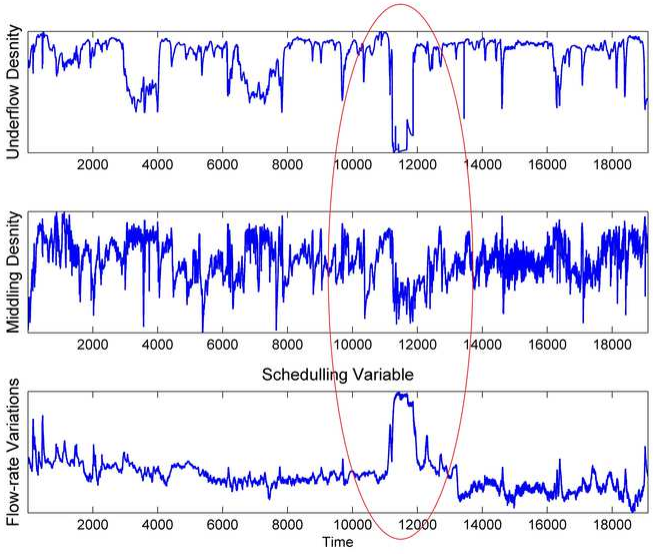

Figure 6.14: Validation data-set for the industrial case study

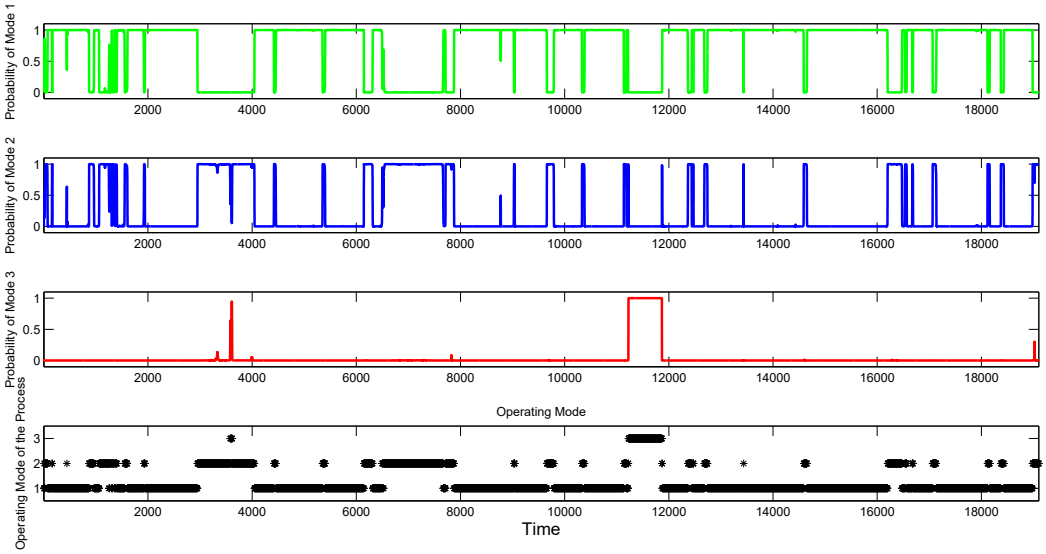

Figure 6.15: Operating modes of the process for the industrial case study based on the proposed method of Chapter 3

In order to compare the results with the case of conventional HMMs, the same

training and test data-sets are utilized. However, only the observed part is used since conventional HMMs cannot deal with missing observations. Results of the operating mode diagnosis based on conventional HMMs are presented in Figure 6.16.

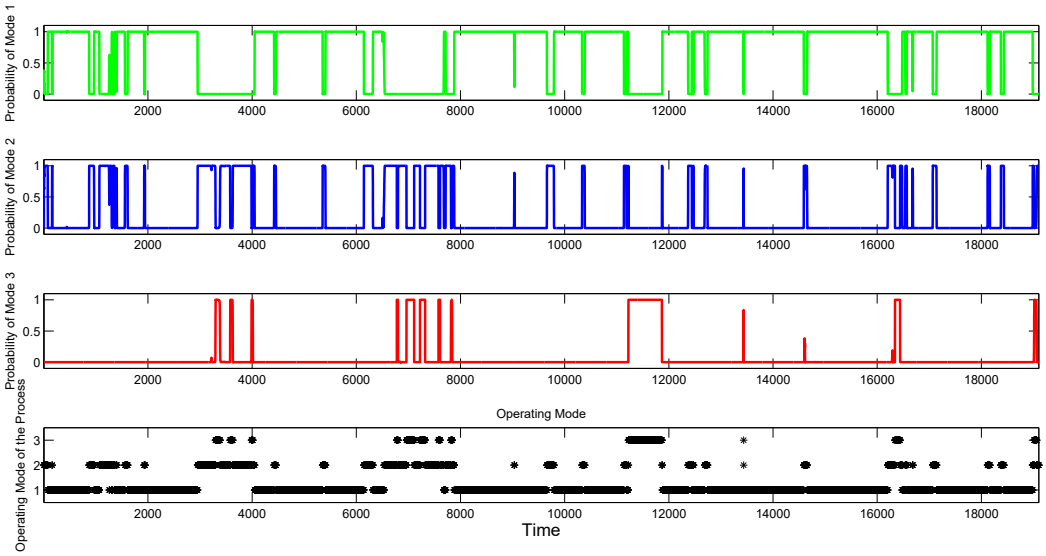

Figure 6.16: Operating modes of the process for the industrial case study based on conventional HMMs

Again, it is observed from Figure 6.16 that a number of false alarms appear in the predictions and the abnormal and faulty modes are not clearly diagnosed from each other when using conventional HMMs. The better performance of the proposed method of Chapter 3 (compare the results in Figures 6.15 and 6.16) can be understood from the behavior of the scheduling variable (Flow-rate) in Figure 6.14. One could see that in the faulty mode, the scheduling variable suddenly increases and helps in operating condition diagnosis. However, in the normal and abnormal operations, the scheduling variable correctly affects the transition probabilities to remain in the normal and abnormal modes and avoid false alarms.

6.7 PSV Monitoring Based on Symbolic Episode Representation and HMMs

In this section, application of the proposed method in Chapter 5, i.e., a combination of HMMs and symbolic episode representation based on a fixed window of observations, is tested on the tailings line of the primary separation vessel (PSV).

6.7.1 Upset and Normal Operating Conditions and Variable Selection

As previously mentioned, among all the variables which can affect the critical minimum velocity estimation, the middlings and underflow densities can be measured on-line. The historical data shows that the middlings interface level is correlated with the middlings density in upset operating conditions. Furthermore, it provides faster responses in comparison to the middlings density and more clear patterns. Therefore, middlings interface level can be used as a potential sanding indicator. Underflow density also shows an obvious fast increasing trend in upset regions. An example of an upset region in May 2011 is depicted in Figure 6.17. The industrial data are normalized due to proprietary reason.

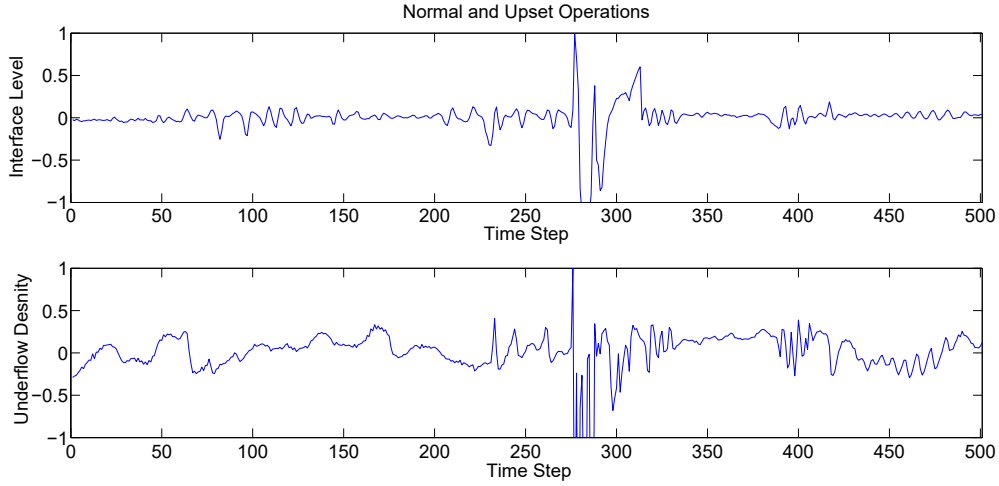

Figure 6.17: Underflow density and the interface level of the PSV unit in an upset region

As illustrated in Figure 6.17, in upset regions, there is a sudden jump in the interface level and an increase in the underflow density. This phenomenon can be understood from the process behavior. When solid particles accumulate in the underflow stream, the underflow density increases and results in a higher level of the middlings. Finally, underflow density and the middlings interface level are selected as two potential sanding indicators.

6.7.2 Training Data for the Normal and Upset Operations

Five cases of normal operating regions followed by an upset region in the historical data of 2011–2012 have been used to train HMMs for the normal and upset operating conditions. In the case of having an upset region, the system can return to the normal

operation if the operator makes an appropriate fast reaction. But in the case of having a complete sanding the system must be shut down for repair and sand removal which will cause a lot of physical effort and financial loss. An example of an upset operation in the tailings line which ended up to a complete sanding in June 2011 is presented in Figure 6.18. The upset region between sample times 100 and 200 has been passed by increasing the tailings flow rate. However, the sanding region after sample time 350 has ended up to complete shut-down of the process.

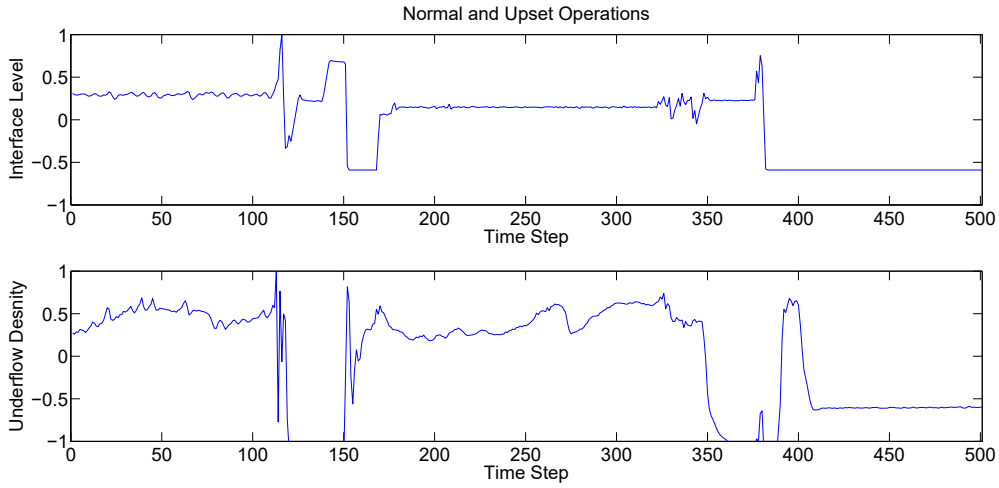

Figure 6.18: Interface level and underflow density - combination of normal and upset operating regions occurred in July 2011

6.7.3 Adaptive Fuzzification and Decision Making

The methodology proposed in Chapter 5 is tested on the historical data of the primary separation vessel.

A total number of 652 discrete observations are used to train the normal and upset hidden Markov models. It includes 320 observations for normal and 332 observations for upset regions corresponding to five normal/ upset operating conditions. In each data set two third of the data is used for training purposes and one third is used for validation.

As explained in Section 5.6, the number of states in each HMM is an important design issue. Small number of states allows a faster training but incomplete classification. Large number of states need more computational time and might cause over fitting problems. The number of hidden states is usually selected as the average number of symbols in the sequence of the training set [20]. In this problem, considering 6 and 8 hidden states for the normal and upset regions respectively will provide an

appropriate performance.

A combination of normal and upset operating regions occurred in July 2011 is presented in Figure 6.19.

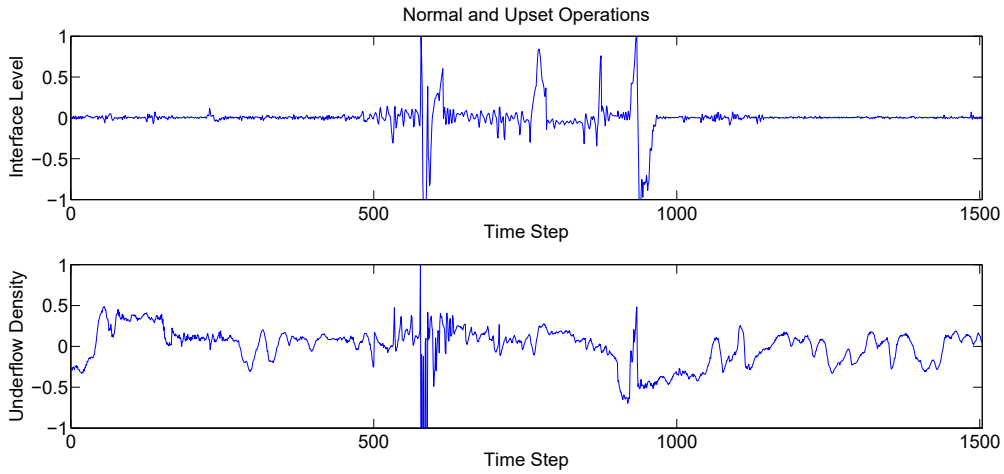

Figure 6.19: Interface level and underflow density - combination of normal and upset operating regions occurred in July 2011.

Wavelet analysis is used to remove the noise from signals. The level of noise removal is very important when using wavelets. Our experience in the historical data of the underflow density and interface level of the PSV unit shows that removing the noise in two levels is sufficient to capture the main information of the signals.

Small and large variance (normal and abnormal) states for the magnitudes and durations of the interface level and underflow density signals and the adaptive fuzzy membership functions are presented in Figures 6.20 to 6.23. The mean and variance of the membership functions vary according to Equation 5.20. Means and variances of the different modes obtained from the EM algorithm are presented in Tables 6.3 and 6.4.

Table 6.3: Parameters of the different modes from the historical data (interface level)

$$\mu_m^{(1)} = 1.4391,\ \mu_d^{(1)} = 3.5337,\ \mu_m^{(2)} = 11.0559,\ \mu_d^{(2)} = 6.7118$$
$$\sigma_m^{(1)} = 1.8066,\ \sigma_d^{(1)} = 6.7118,\ \sigma_m^{(2)} = 13.6903,\ \sigma_d^{(2)} = 7.2399$$
$$A = \begin{pmatrix} 0.9776 & 0.0224 \\ 0.2627 & 0.7373 \end{pmatrix}$$

Finally, Based on the durations and magnitudes of the episodes obtained from maximum, minimum and inflection points and adaptive fuzzification, the continuous signals in Figure 6.19 are converted to discrete observation sequences from 1 to 36,

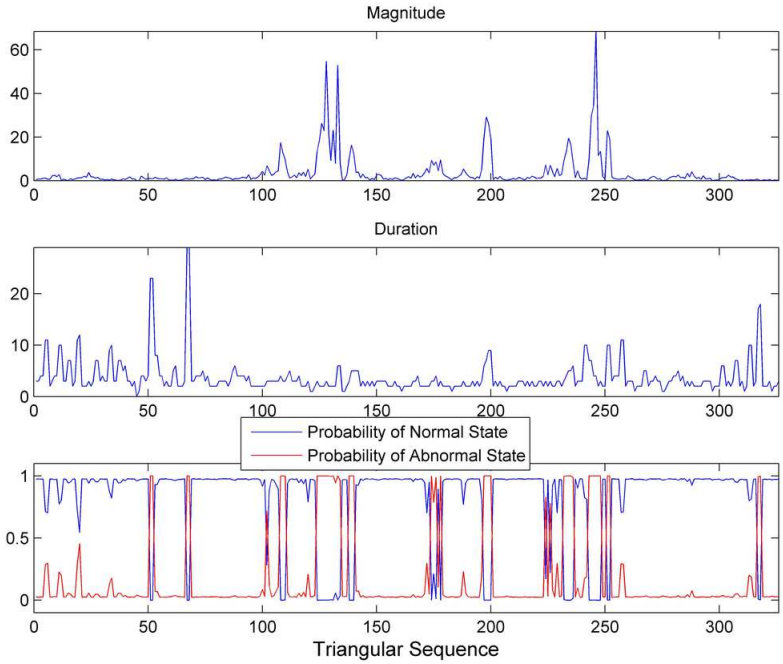

Figure 6.20: Different states for the durations and magnitudes in the interface level signal

Table 6.4: Parameters of the different modes from the historical data (underflow density)

$$\mu_m^{(1)} = 0.4481,\ \mu_d^{(1)} = 3.4534,\ \mu_m^{(2)} = 2.33572,\ \mu_d^{(2)} = 7.8026$$
$$\sigma_m^{(1)} = 0.5531,\ \sigma_d^{(1)} = 1.9811,\ \sigma_m^{(2)} = 2.9804,\ \sigma_d^{(2)} = 4.9794$$
$$A = \begin{pmatrix} 0.9855 & 0.0145 \\ 0.2052 & 0.7948 \end{pmatrix}$$

where each of these numbers corresponds to a certain type of triangle. Results are presented in Figures 6.24 and 6.25.

The upset regions are specified with two direct red lines. As it is clear from the interface level, when the system enters the upset region, more large type triangles (33-36, 23-27, 14-18 and 5-9) are produced. The underflow density also shows some changes in the pattern at sample times around 150 and 220 which correspond to the upset regions in Figure 6.19. Moreover, due to the normalization of the underflow density in the previous section, almost the same number of triangular sequences is generated from both signals. However, in a real time application, it is not necessary to have the same number of discrete observations at each time step, e.g., a signal might have more fluctuations than the other signal. The final decision will always be

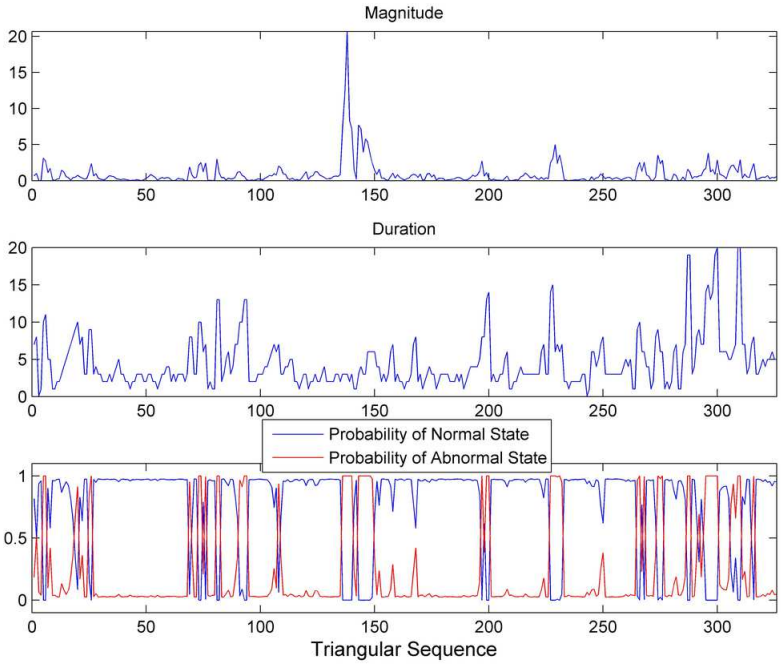

Figure 6.21: Different states for the durations and magnitudes in the underflow density signal

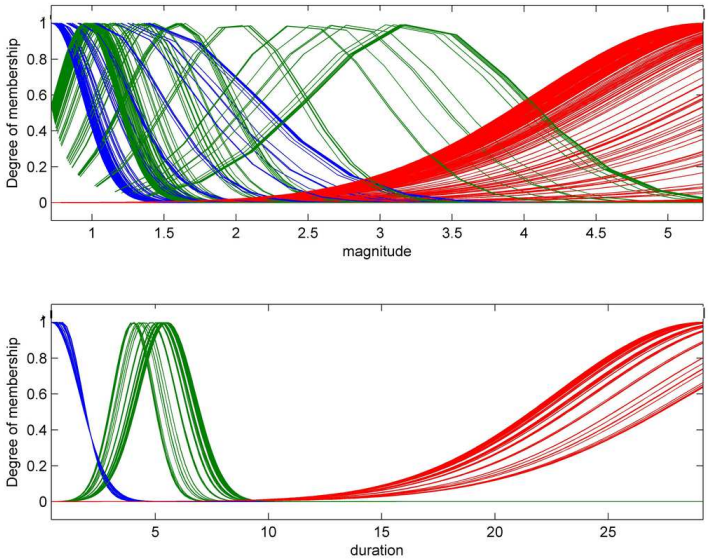

Figure 6.22: Adaptive fuzzy membership functions for the interface level signal

made based on the most recent observations of the window.

Simulation results of decision making based on a window size (N_W) of 15 and 7

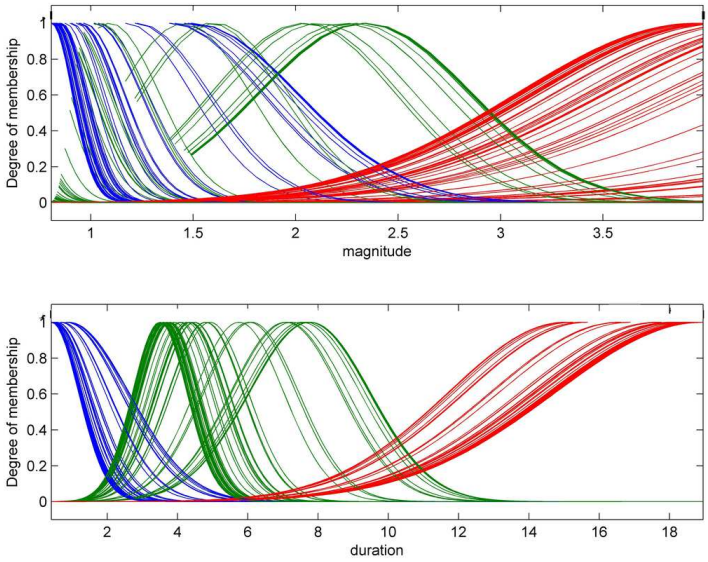

Figure 6.23: Adaptive fuzzy membership functions for the underflow density signal

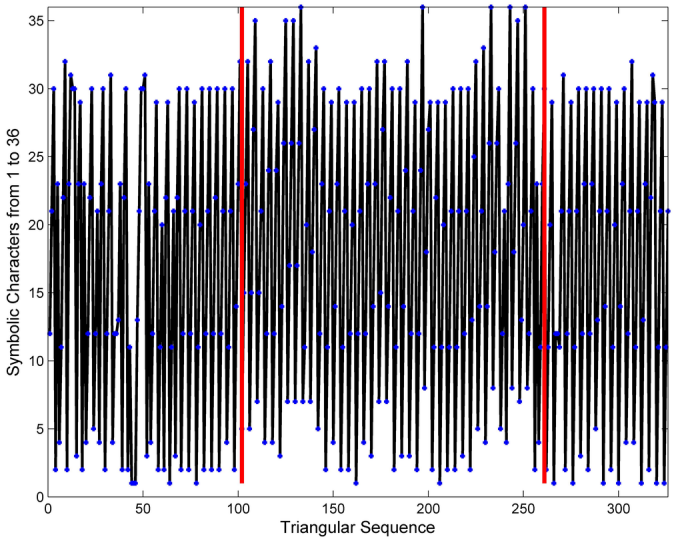

Figure 6.24: Discretized observations of the interface level using appropriate fuzzy rules and membership functions

recent observations are presented in Figures 6.26 and 6.27.

Results show that the system starts from a normal operation, then switches to an upset operation and finally returns to the normal operation again. Increase in the probability of the normal mode at the upset region shows that the normal and upset regions have similar patterns at some periods. In time steps around 130, suddenly

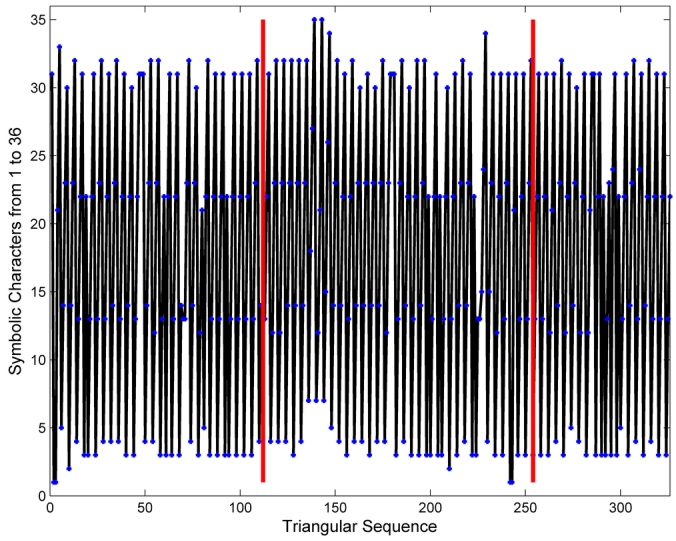

Figure 6.25: Discretized observations of the underflow density using appropriate fuzzy rules and membership functions

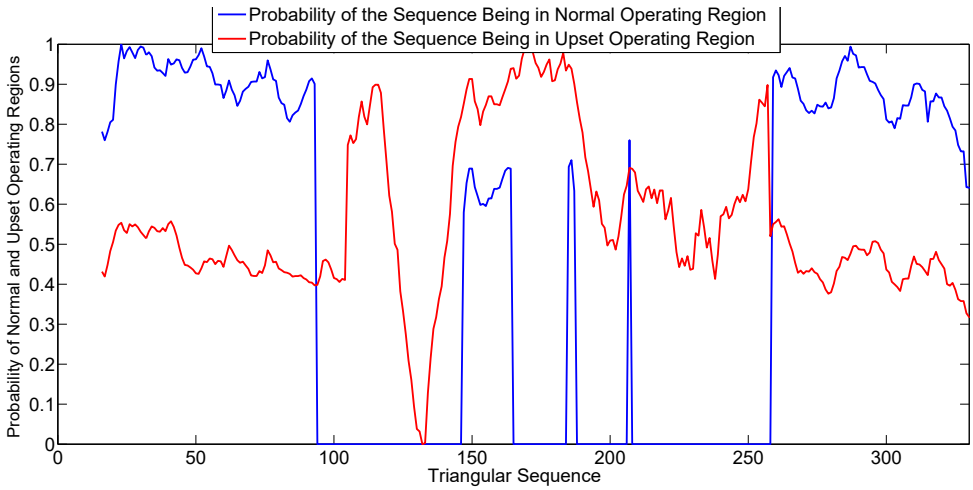

Figure 6.26: Overall classification of the process for different window sizes - $N_W = 15$

both probabilities become zero, which is an indicator for the existence of an unknown pattern in the data.

Decreasing the window size will result in a classification based on more recent observations in the window. However, the number of false alarms might increase accordingly. Therefore, the window size should be selected for different processes according to the importance of the old observations in the window. The problem of optimal window size selection for the case of fixed fuzzy membership functions is addressed in detail in Section 5.7 [95]. Finding the optimal window of observations for

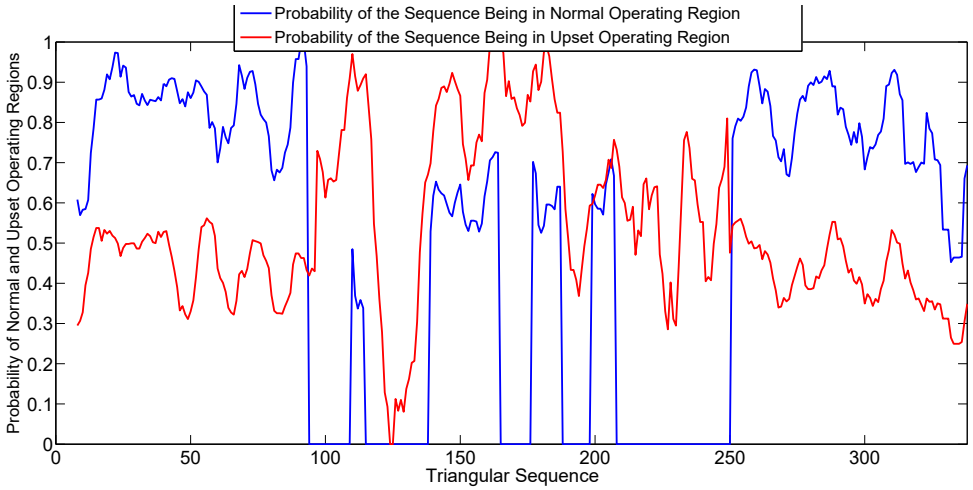

Figure 6.27: Overall classification of the process for different window sizes - $N_W = 7$

the case of adaptive fuzzy membership functions remains as the future study in this field. In this section, N_W is tentatively selected as an average which is small enough to show the current operating condition and large enough to avoid unnecessary alarms for fault detection purposes.

6.8 Conclusion

In Sections 6.1 to 6.5 of this chapter, a novel procedure for on-line estimation of the critical minimum velocity of of slurry flow is introduced. The method is applied on the PSV underflow. A soft sensor is developed to correct the measurements from the on-line analyzer for the key process variable. In order to reduce false alarms, an adaptive scheme based on HMMs is proposed to determine the sensitivity of the critical velocity estimations. A general method for missing data treatment during HMM training is proposed.

The proposed method is tested both on-line and on the historical data of the PSV unit, and shows acceptable performance in detection of operating modes of the process. In upset operating conditions, the estimated value of the critical velocity increases. A caution alarm is generated when the value of the critical velocity is higher than the current flow rate. Increasing the underflow flow rate can help to return the process to the normal operating condition.

In Sections 6.6 and 6.7, the proposed process monitoring methods of Chapters 3 and 5 have been tested on the historical data of PSV respectively. For each method, the appropriate variable selection strategy is explained. Although both methods are successful in recognition of certain types of patterns, a general fault diagnosis strategy

based on data driven methods may not be possible. This is due to the wide variety of patterns that might occur in upset regions, i.e., one should develop numerous models for various patterns if relying only on data driven methods. Furthermore, each data driven framework is appropriate to extract a certain type of pattern, e.g., the proposed method of Chapter 3 is is able to extract continuous changes in means and variances of key variables, while the proposed method of Chapter 5 is searching for sudden spikes and shifts.

On the other hand, when using data driven methods in conjunction with first principle knowledge, the fundamental underlying relations of variables will be used for prediction purposes, while the predictions are modified on-line based on historical data trends. Accordingly, this combination shows a more appropriate performance.

Chapter 7

Concluding Remarks and Future Directions

In this chapter, the conclusions drawn from various chapters of the thesis are summarized. Furthermore, connection of various chapters to the core idea of the thesis is explained.

7.1 Concluding Remarks

The main focus of this thesis is Fault Detection and Isolation (FDI) for chemical processes based on HMMs. Some of the developed methods have been applied on the underflow line of the PSV unit of an oil sand extraction process which has a concern of sand deposition and pipeline plugging.

In Chapter 2 the mathematical fundamentals of the thesis are explained. Due to the existence of various operating regimes and missing variables, the EM algorithm is used to train HMMs. On-line mode diagnosis is based on either the forward-backward algorithm or Hamilton's filtering strategy.

To apply HMMs in industrial applications, three main theoretical extensions have been proposed:

- The first theoretical contribution, as introduced in Chapter 3, is the development of time varying HMMs with a specific Transition Probability Matrix (TPM) structure in presence of missing observations. Consideration of the TPM elements as a function of an auxiliary scheduling variable provides more flexibility in hidden Markov modeling. In the case of having operating modes with a temporal behavior far from the majority of modes, the proposed method assists to more precisely classify various data into clusters. Furthermore, the imposed conditions on the TPM structure reduce the possibility of sudden jumps

between the modes which are far from each other while decreasing the number of parameters to be estimated. Accordingly, the computational cost will be significantly reduced when dealing with processes with a large number of operating modes. Also, the method is able to treat the missing information in the historical data set. As the result, instead of solving a complex optimization problem, the missing variables are integrated out in the iterations of the EM algorithm, and the optimization problem becomes analytically tractable. The proposed method of Chapter 3 has been tested on various examples including the historical data of the PSV unit, and the advantages of the proposed method over conventional techniques are explained.

- The second contribution, as presented in Chapter 4, is to provide a framework to deal with negative effects of outliers in time varying HMMs. This framework makes the model adaptive with respect to both data quality and process operating mode. It is usually expected to have lower quality data when operating near faulty modes. Therefore, according to the data quality in each mode, the proposed method assigns an appropriate scholar weight for the covariance matrix to downweight the negative effect of outliers. The introduced scalar weight follows a *gamma* distribution as a function of the degree of freedom (ν). By integrating out this scalar weight from the joint density function of observations and the weight, the general form of the density function of observations in each operating mode, which is a Student t distribution, is obtained. The Student t distribution has heavier tails in comparison to the Gaussian version. The smaller the assigned degree of freedom (ν), the heavier the tails of the t distribution. Therefore, the percentage of outliers will have a reverse correlation with the assigned degree of freedom, and effect of outliers will be automatically considered during parameter estimation.

- The third contribution, as demonstrated in Chapter 5, is an adaptive FDI strategy based on the combination of Qualitative Trend Analysis (QTA) and HMMs. First, wavelets are applied to remove the high frequency noise of the signals. Next, based on appropriate fuzzy membership functions, the continuous time signals are converted to some triangular episodes. The fuzzification step has a hierarchical structure, in which, based on the durations and magnitudes of the triangles, and fixed fuzzy membership functions in various modes, time varying membership functions are generated to have more accurate from continuous to discrete mappings. The main advantage of triangular representation is to reduce the resolutions of observations, i.e., to focus on main features rather

than details. Furthermore, the set of discrete observations are appropriate for classification systems such as HMMs. Having the set of discrete observations available, HMMs are trained for normal and abnormal operations of the process. In an on-line application, the decision on the current regime of the process is made based on a window of N_W recent observations and the forward-backward algorithm. A search algorithm to find the optimal window of observations is also proposed. Similar to Chapter 3, the proposed method is tested on various examples including the historical data of the PSV unit. The method shows a superior performance over others in all the examples.

Chapters 3, 4 and 5 are in parallel in the sense that they all provide various FDI methods for chemical processes based on HMMs.

In Chapter 6, first, the developed methods in various chapters are applied on the PSV unit. Our experience shows that data driven frameworks are not sufficient for a successful FDI in the PSV. This is due to the existence of numerous patterns and operating modes in the operation of the PSV unit.

Therefore, data driven methods are combined with some first-principle knowledge of the process to provide a general FDI strategy. First, the appropriate semi-empirical equation is selected to estimate the required critical velocity to constantly move the solid bed inside slurry pipelines and avoid pipeline plugging. Since one of the key variables, the carrier fluid density, has some measurement inaccuracies, a soft sensor based on the recursive Partial Least Squares (rPLS) method is developed to provide a parallel measurement for this variable. Next, the estimated velocity is modified based on the operating mode of the underflow flow rate. The idea is to infer the mode of the underflow flow rate using HMM classification, and then, decide on the estimation error of the predicted velocity based on the estimated operating mode. It is shown that such consideration significantly reduces the number of false alarms. This strategy has been tested in on line environment and illustrated successful results.

To summarize, the main focus of this thesis is the development of HMM based techniques 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. Moreover, consideration of different emission density functions at various modes of HMMs assist to extract various types of information from the data, e.g., the Student t distribution treats the negative effect of low quality data, or discrete probability mass functions assist to extract the key features and some certain type of discrete information from the data. For the industrial case study of this thesis, it is observed that a combination of data driven approaches with

the first principle knowledge provides the most accurate process monitoring results. This method is used in the online environment consequently.

7.2 Future Directions

7.2.1 Number of Operating Modes

The number of operating modes or local models is usually assumed to be known *a-priori* in multiple model identification problems. The main disadvantage of such an assumption is the overfitting issue in complex problems, i.e., modeling the random noise instead the true underlying correlation [85].

One approach to address such a problem is to use Variational Bayesian (VB) methods. Having appropriate priors, VB marginalizes the likelihood function over model parameters. The model can then be maximized with respect to model size providing an optimal structure for the process [85]. In such an approach, no prior is considered for the model size. Instead, the variational inference procedure will be run for various model sizes, and among them, the model with the largest variational free energy (an approximation of the log marginal likelihood) will be selected. This approach is computationally less expensive in comparison to the EM algorithm which usually uses cross-validation techniques for model size selection [85].

Although VB has been previously used in Statistics literature to find the optimal number of components in a Gaussian mixture model or some particular types of HMMs ([79, 85]), its application in a general HMM based fault diagnosis, e.g., based on discrete observations, is still sparse, and can be considered as a new direction in this area.

7.2.2 Uncertain/ Discrete Scheduling Variable

As an indicator of the process operating mode, the scheduling variable plays a critical role in local model selection, and an uncertain scheduling variable can cause many modeling inaccuracies. To address this issue, Kalman smoother, or Sequential Monte-Carlo (SMC) method can be used to consider various dynamics of the scheduling variable [125].

Other than this, it is possible to consider appropriate priors for parameters of the scheduling variable, and solve the problem under the VB framework. Such a consideration will integrate out the effect of noise while estimating process parameters. Furthermore, it is possible to use particle filters when dealing with non-Gaussian distributions of the scheduling variable [57].

In addition, a practical approach to reduce the scheduling variable uncertainty is to, first, remove the noise using appropriate filters, and then, generate the scheduling variable in the form of some discrete symbols. This will greatly reduce both the noise and resolution of the scheduling variable, and satisfy the need to have an overall indication to the true operating mode. One approach to perform such resolution reduction is to use the triangular representation method introduced in Chapter 5.

7.2.3 Latent Variable Models in Conjunction with HMMs

In two very recent studies, multiple Principal Component Analyzers (PCA) have been used for the purpose of fault classification [21, 22]. Latent variable models such as PCA are appropriate to extract static information from high dimensional data sets. Therefore, having them in conjunction with HMMs will provide a condition to further obtain the temporal information.

Similarly, HMMs can be used in conjunction with Partial Least Squares (PLS), and then be applied in soft sensor or process monitoring applications. Having such consideration, the temporal process information will be further considered when developing a multiple PLS model.

7.2.4 Conditional Random Fields

Logistic regression models are appropriate classification tools to predict possible categories of a dependent variable given a set of independent variables. Therefore, they have the capability to extract the static information in the data [126]. The general form of a multinomial logistic regression model is presented in Equation 7.1.

$$ln\frac{\pi_j(x_i)}{\pi_J(x_i)} = \beta_j^T x_i,\ j = 1, ..., J-1 \tag{7.1}$$

where in this equation $x_i = (x_{i0}, ..., x_{ip})^T$ denotes the explanatory variables for subject $1 \leq i \leq n$ and $\beta_j = (\beta_{j0}, ..., \beta_{jp})$, $1 \leq j \leq J-1$, is the regression parameters for the j^{th} category. The baseline category J is usually selected as the most common category. $y_i = (y_{i1}, ..., y_{iJ})$ can be considered as a multinomial trial for subject i. The trial y_{ij} is equal to one whenever a trial occurs in category j. Every trial might occur only in one category, i.e., $\sum_{j=1}^{J} y_{ij} = 1$. Consequently, $\pi_j(x_i) = P(y_{ij} = 1|x_i)$ in Equation 7.1 [126].

According to Equation 7.1, $exp(\beta_j^T x_i) > 1$ represents the trial to occur in category

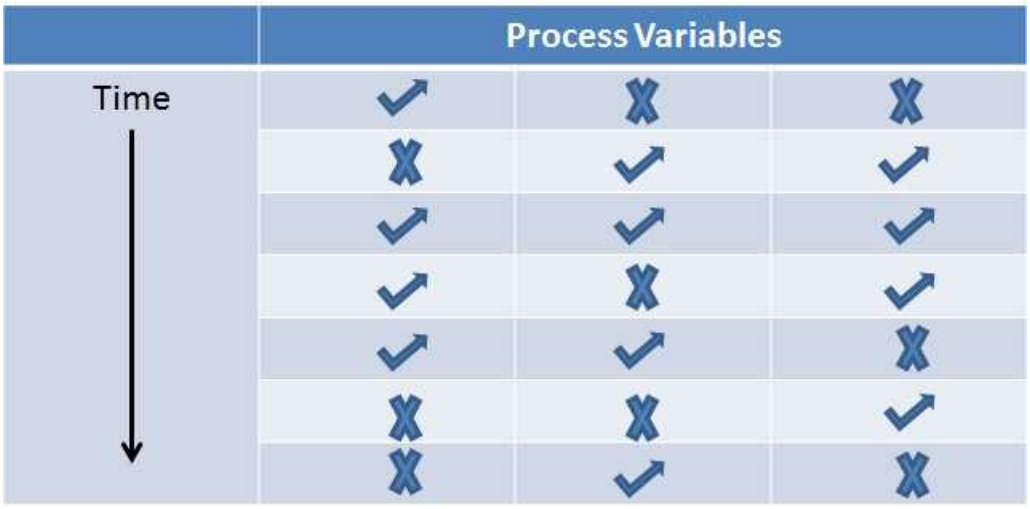

Figure 7.1: Various configurations of missing observations in the historical data

j against J. $\pi_j(x_i)$ can be written as in Equation 7.2.

$$\pi_j(x) = \frac{exp(\beta_j^T x)}{1 + \sum_{h=1}^{T-1} exp(\beta_h^T x)} \tag{7.2}$$

Performing logistic regression according to discrete modes of a factor graph like HMM creates a "Conditional Random Field (CRF)" [127]. CRFs have a wide range of applications including natural language processing, computer vision and informatics [127]. Accordingly, they can be widely used in chemical processes to classify various operating regions.

7.2.5 Missing Data Treatment

The Expectation Maximization (EM) algorithm, which is an iterative optimization technique, provides a condition to integrate out the missing observations during parameter estimation. Consequently, the likelihood (cost function) surface is reshaped, and unknown parameters become tractable [29]. Following such an approach, many identification procedures have been introduced to develop mathematical models for industrial processes [45, 57].

Partially missing observations in a multivariate vector of data is an interesting topic to be further considered in future studies [29], i.e, considering Missing Not at Random (MNAR) missing data instead Missing at Random (MAR) or Missing Completely at Random (MCAR). Various possible configurations of missing observations in historical data are presented in Figure 7.1.

In order to solve the parameter estimation problem under the EM framework, the Q function should be obtained as previously explained in various chapters of the thesis. However, formulation of the Q function according to Figure 7.1 will result in many difficulties [29]. As an example, according to Figure 7.1, various possibilities exist for calculation of the joint distribution of the observed and missing segments. Furthermore, calculation of the expected value of missing variables given the observed part, when the observed and missing segments are correlated, will cause many calculation difficulties [29]. These problems still remain as subjects of future studies in this area.

7.2.6 Model Switching Mechanism

Although HMMs have been widely used in some areas such an Economics, their application in process identification is quite recent, and there are many contributions to be further considered [54], e.g., subject of time varying HMMs to monitor chemical processes as introduced in Chapter 3 of the thesis. In such a structure, the transition probabilities are considered to be time varying, i.e., a_{ij} is replaced by $a_{ij}(t)$ where t represents the current sample time.

One possible improvement to the considered structure for time varying HMMs in this thesis is to use more sophisticated TPM structures. One could use logistic regression models in such a structure where the diagonal elements of the TPM (a_{ii}) are the baseline categories of the logistic regression model, and other elements of the row ($a_{ij,i\neq j}$) are functions of the baseline transition probability [126]. Application of such a structure will further improve the flexibility of the hidden Markov modeling. More information on logistic regression is provided in Equations 7.1 and 7.2.

Bibliography

[1] R. Isermann, Model-Based Fault-Detection and Diagnosis - Status and Applications, Ann. Rev. Contr. 29 (2005) 71–85.

[2] A. Pernestål, A Bayesian Approach to Fault Isolation with Application to Diesel Engine Diagnosis, KTH (2007)

[3] V. Venkatasubramanian, R. Rengaswamy, S. N. Kavuri, K. Yin, A Review of Process Fault Detection and Diagnosis Part III: Process History Based Methods, Comp. Chem. Eng. 27 (2003) 327–346.

[4] P. Smyth, Hidden Markov Models for Fault Detection in Dynamic Systems, Patt. Recogn. 27(1) (1994) 149–164.

[5] R. Isermann, Process Fault Detection Based on Modeling and Estimation Methods - A Survey, Automatica 20(4) (1984) 387–404.

[6] R. N. Clark, A Simplified Instrument Failure Detection Scheme, IEEE Trans. Aeros. Electron. Syst. 14(4) (1978) 558–563.

[7] A. S. Willsky, A Survey of Design Methods for Failure Detection in Dynamic Systems, Automatica 12 (1976) 601–611.

[8] J. Gertler, Fault Detection and Isolation Using Parity Relations, Control Eng. Practice 5(5) (1997) 653–661.

[9] K. Watanabe, I. Matsuura, M. Abe, M. Kubota, D. M. Himmelblau, Incipient Fault Diagnosis of Chemical Processes via Artificial Neural Networks, AIChE J. 35(11) (1989) 1803–1812.

[10] J. C. Hoskins, K. M. Kaliyur, D. M. Himmelblau, Fault Diagnosis in Complex Chemical Plants Using Artificial Neural Networks, AIChE J. 37(1) (1991) 137–141.

[11] D. M. Himmelblau, Fault Detection and Diagnosis in Chemical and Petrochemical Processes, Amsterdam: Elsevier Press (1978).

[12] S. Yin, S. X. Ding, X. Xie, H. Luo, A Review on Basic Data-Driven Approaches for Industrial Process Monitoring, IEEE Trans. Ind. Electron. 61(11) (2014) 6418–6428.

[13] S. Yin, X. Li, H. Gao, O. Kaynak, Data-Based Techniques Focused on Modern Industry: An Overview, IEEE Trans. Ind. Electron. 62(1) (2015) 657–667.

[14] D. Garca-Alvarez, Fault Detection Using Principal Component Analysis (PCA) in a Waste Water Treatment Plant (WWTP), Proc. Inter. Stud. Scient. Conf. (2009)

[15] P. Geladi, B. R. Kowalski, Partial Least Squares Regression: A Tutorial, Analytica Chemica Acta 185 (1986) 1–17.

[16] E. Parzen, On Estimation of a Probability Density Function and Mode, Ann. Math. Stat. 33(3) (1962) 1065–1076.

[17] K. Fukunaga, Introduction to Statistical Pattern Recognition, New York: Academic Press (1972).

[18] S. Verron, T. Tiplica, A. Kobi, Fault Detection of Univariate Non-Gaussian Data with Bayesian Network, IEEE Inter. Conf. Indust. Tech. (2010)

[19] J. C. Wong, K. A. McDonald, A. Palazoglu, Classification of Process Trends Based on Fuzzified Symbolic Representation and Hidden Markov Models, J. Proc. Cont. 8(5–6) (1998) 395–408.

[20] J. C. Wong, K. A. McDonald, A. Palazoglu, Classification of Abnormal Plant Operation Using Multiple Process Variable Trends, J. Proc. Cont. 11 (2001) 409–418.

[21] J. Zhu, Z. Ge, Z. Song, Dynamic Mixture Probabilistic PCA Classifier Modeling and Application for Fault Classification, J. Chemometrics (2015), DOI: 10.1002/cem.2714.

[22] J. Zhu, Z. Ge, Z. Song, HMM-Driven Robust Probabilistic Principal Component Analyzer for Dynamic Process Fault Classification, IEEE Trans. Ind. Electron. (2015), DOI: 10.1109/TIE.2015.2396877.

[23] R. Jiang, J. Yu, V. Makis, Optimal Bayesian Estimation and Control Scheme for Gear Shaft Fault Detection, Comput. Indust. Eng. 63 (2012) 754–762.

[24] J. Ying, T. Kirubarajan, K. R. Pattipati, A. Patterson-Hine, A Hidden Markov Model-Based Algorithm for Fault Diagnosis with Partial and Imperfect Tests, IEEE Trans. Syst. Man Cyber. 30(4) (2000) 463–473.

[25] A. Ghasemi, S. Yacout, M. Ouali, Parameter Estimation Methods for Condition-Based Maintenance With Indirect Observations, IEEE Trans. Reliab. 59(2) (2010) 426–439.

[26] W. C. Wong, J. H. Lee, Fault Detection and Diagnosis Using Hidden Markov Disturbance Models, Ind. Eng. Chem. Res. 49 (2010) 7901–7908.

[27] S. Borman, The Expectation Maximization Algorithm-a Short Tutorial, Submitted for publication (2004). 1–9.

[28] A. P. Dempster, N. M. Laird, D. B. Rubin, Maximum Likelihood from Incomplete Data via the EM Algorithm, J. Roy. Stat. Soc. Ser. B 39(1) (1977) 1–38.

[29] G. McLachlan, T. Krishnan, The EM algorithm and Extensions, Wiley, New York (1997).

[30] S. N. Rai, D. E. Matthews, Improving the EM Algorithm, Biometrics 49(2) (1993) 587–591.

[31] L. Xu, M. I. Jordan, On Convergence Properties of the EM Algorithm for Gaussian Mixtures, Neural Computation 8 (1996) 129–151.

[32] R. A. Redner, H. F. Walker, Maximum Likelihood and the Em Algorithm, SIAM Review 26(2) (1984) 195–239.

[33] R. Salakhutdinov, S. Roweis, Z. Ghahramani, Optimization with EM and Expectation-Conjugate-Gradient, Proceedings of the Twentieth International Conference on Machine Learning (ICML) (2003)

[34] D. Karlis, E. Xekalaki, Choosing Initial Values for the EM Algorithm for Finite Mixtures, Comput. Stat. Data Anal. 41 (2003) 577–590.

[35] G. J. McLachlan, On the Choice of Starting Values for the EM Algorithm in Fitting Mixture Models , J. Roy. Stat. Soc. S. D 37(4/5) (1988) 417–425.

[36] B. G. Leroux, Consistent Estimation of a Mixing Distribution, Ann. Stat. 20(3) (1992) 1350–1360.

[37] W. A. Woodward, W. C. Parr, W. R. Schucany, H. Lindsey, A Comparison of Minimum Distance and Maximum Likelihood Estimation of a Mixture Proportion, J. Amer. Stat. Assoc. 79(387) (1984) 590–598.

[38] D. Bohning, E. Dietz, R. Schaub, P. Schlattmann, B. G. Lindsay, The Distribution of the Likelihood Ratio for Mixtures of Densities from the One Parameter Exponential Family, Ann. Inst. Statist. Math. 46(2) (1994) 373–388.

[39] A. J. Bilmes, A Gentle Tutorial of the EM Algorithm and Its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models, Inter. Comp. Scie. Inst. 4 510 (1998) 126.

[40] L. R. Rabiner, A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition, Proc. IEEE 77(2) (1989) 257–286.

[41] E. F. Lussier, Markov models and hidden Markov models: A brief tutorial, International Computer Science Institute (1998)

[42] L. Gu, EM and HMM, https://www.cs.cmu.edu/ epxing/Class/10701-08s/recitation/em-hmm.pdf

[43] T. Kanungo, Hidden Markov Models, Center of Automation Research, University of Maryland, http://www.kanungo.com/software/hmmtut.pdf

[44] J. D. Hamilton, Analysis of Time Series Subject to Changes in Regimes, J. Econometrics 45(1–2) (1990) 39–70.

[45] N. Sammaknejad, B. Huang, W. Xiong, A. Fatehi, F. Xu, A. Espejo, Operating Condition Diagnosis Based on HMM with Adaptive Transition Probabilities in Presence of Missing Observations, AIChE J. 61(2) (2015) 477–491.

[46] J. D. Hamilton, Rational-Expectations Econometric Analysis of Changes in Regime: An Investigation of the Term Structure of Interest Rates, J. Econ. Dynam. Cont. 12 (1998) 385–423.

[47] N. P. B. Bollen, S. F. Gray, R. E. Whaley, Regime Switching in Foreign Exchange Rates: Evidence from Currency Option Prices, J. Econometrics 94 (2000) 239–276.

[48] A. Ang, G. Bekert, Regime Switches in Interest Rates, J. Busin. Econ. Stat. 20(2) (2002) 163–182.

[49] D. Pelletier, Regime Switching for Dynamic Correlations, J. Econometrics 131 (2006) 445–473.

[50] T. D. Mount, Y. Ning, X. Cai, Predicting Price Spikes in Electricity Markets Using a Regime Switching Model with Time Varying Parameters, Energy Economics 28 (2006) 62–80.

[51] F. X. Diebold, J. Lee, G. C. Weinbach, Regime Switching with Time Varying Transition Probabilities, Non-stat. Time Ser. Anal. Cointeg. (1994) 283–302.

[52] A. J. Filardo, Business-Cycle Phases and Their Transitional Dynamics, J. Busin. Econ. Stat. 12(3) (1994) 299–308.

[53] E. Otranto, The Multi-chain Markov Switching Model, J. Forecasting 24 (2005) 523–537.

[54] O. L. V. Costa, M. D. Fragoso, R. P. Marques, Discrete-time Markov Jump Linear Systems, Springer (2005).

[55] X. Jin, B. Huang, Identification of Switched Markov Autoregressive Exogenous Systems with Hidden Switching State, Automatica 48(2) (2012) 436–441.

[56] M. J. Kim, V. Makis, R. Jiang, Parameter Estimation for Partially Observable Systems Subject to Random Failure, Appl. Stochas. Models Busin. Indust. 29 (2013) 279–294.

[57] J. Deng, B. Huang, Identification of Nonlinear Parameter Varying Systems with Missing Output Data, AIChE J. 58(11) (2012) 3454–3467.

[58] M. Keshavarz, B. Huang, Bayesian and Expectation Methods for Multivariate Change Point Detection, Comput. Chem. Eng. 60 (2014) 339–353.

[59] S. Imtiaz, S. Shah, Treatment of Missing Values in Process Data Analysis, Canad. J. Chem. Eng. 86 (2008) 838–858.

[60] J. L. Schafer, J. W. Grahm, Missing Data: Our View of the State of the Art, Psychol. meth. 7 (2008) 2076–2089.

[61] D. A. Whitley, Genetic Algorithm Tutorial, Computer Science Department. Colorado State University.

[62] C. B. Lucasius, G. Kateman, Understanding and Using Genetic Algorithms Part 1. Concepts, Properties and Context, Chemomet. Intellig. Lab. Sys. 19 (1993) 1–33.

[63] A. M. Martinez, J. Vitria, Learning Mixture Models Using a Genetic Version of the EM Algorithm, Patt. Recogn. Let. 21 (2000) 759–769.

[64] F. Pernkopf, D. Bouchaffra, Genetic-based EM Algorithm for Learning Gaussian Mixture Models, IEEE Trans. Pat. Analy. Mach. Intel. 27(8) (2005) 1344–1348.

[65] R. H. Byrd, J. C. Gilbert, J. Nocedal, A Trust Region Method Based on Interior Point Techniques for Non-linear Programming, Math. Prog. 89(1) (2000) 149–185.

[66] R. H. Byrd, M. E. Hribar, J. Nocedal, An Interior Point Algorithm for Large-scale Non-linear Programming, SIAM J. optimiz. 9(4) (1999) 877–900.

[67] M. A. Henson, D. E. Seborg, Input-output Linearization of General Nonlinear Processes, AIChE J. 36(11) (1990) 1753–1757.

[68] R. Muradore, P. Fiorini, A PLS-based Statistical Approach for Fault Detection and Isolation of Robotic Manipulators, IEEE Trans. Ind. Electron. 59(8) (2012) 3167–3175.

[69] S. Yin, X. Zho, O. Kaynak, Improved PLS Focused on Key-Performance Indicator-Related Fault Diagnosis, IEEE Trans. Ind. Electron. 62(3) (2015) 1651–1658.

[70] D. Bruckner, R. Velik, Behavior Learning in Dwelling Environments with Hidden Markov Models, IEEE Trans. Ind. Electron. 57(11) (2010) 3653–3660.

[71] N. Sammaknejad, B. Huang, A. Fatehi, Y. Miao, F. Xu, A. Espejo, Adaptive Monitoring of the Process Operation Based on Symbolic Episode Representation and Hidden Markov Models with Application Toward an Oil Sand Primary Separation, Comp. Chem. Eng. 71(4) (2014) 281–297.

[72] M. Jager, F. A. Hamprecht, Principal Component Imagery for the Quality Monitoring of Dynamic Laser Welding Processes, IEEE Trans. Ind. Electron. 56(4) (2009) 1307–1313.

[73] A. Soualhi, G. Clerc, H. Razik, Detection and Diagnosis of Faults in Induction Motor Using an Improved Artificial Ant Clustering Technique, IEEE Trans. Ind. Electron. 60(9) (2013) 4053–4062.

[74] T. Boukra, A. Lebaroud, G. Clerc, Statistical and Neural-network Approaches for the Classification of Induction Machine Faults Using the Ambiguity Plane Representation, IEEE Trans. Ind. Electron. 60(9) (2013) 4034–4042.

[75] Y. Gritli, L. Zarri, C. Rossi, F. Filippetti, G. Capolino, D. Casadei, Advanced Diagnosis of Electrical Faults in Wound-rotor Induction Machines, IEEE Trans. Ind. Electron. 60(9) (2013) 4012–4024.

[76] B. Li, M. Y. Chow, Y. Tipsuwan, J. C. Hung, Neural-network-based Motor Rolling Bearing Fault Diagnosis, IEEE Trans. Ind. Electron. 47(5) (2000) 1060–1069.

[77] S. A. Arogeti, D. Wang, C. B. Low, M. Yu, Fault Detection Isolation and Estimation in a Vehicle Steering System, IEEE Trans. Ind. Electron. 59(12) (2012) 4810–4820.

[78] S. Khatibisepehr, B. Huang, A Bayesian Approach to Robust Process Identification with ARX Models, AIChE J. 59(3) (2013) 845–859.

[79] M. Svensn, C. M. Bishop, Robust Bayesian Mixture Modelling, Neurocomputing. 64 (2005) 235–252.

[80] X. Jin, B. Huang, Robust Identification of Piecewise/Switching Autoregressive Exogenous Process, AIChE J. 56(7) (2010) 1829–1844.

[81] Y. Fang, M. K. Jeong, Robust Probabilistic Multivariate Calibration Model, Technometr. 50(3) (2008) 305–316.

[82] Y. Lu, B. Huang, Robust Multiple LPV Approach to Nonlinear Process Identification Using Mixture t Distributions, J. Process Contr. 24(9) (2014) 1472–1488.

[83] K. Zhang, R Gonzalez, B. Huang, G. Ji, Expectation Maximization Approach to Fault Diagnosis with Missing Data, IEEE Trans. Ind. Electron. 62(2) (2015) 1231–1240.

[84] S. P. Chatzis, D. I. Kosmopoulos, T. A. Varvarigou, Robust Sequential Data Modeling Using an Outlier Tolerant Hidden Markov Model, IEEE Trans. Pat. Anal. Mach. Learn. 31(9) (2009) 1657–1669.

[85] S. P. Chatzis, D. I. Kosmopoulos, A Variational Bayesian Methodology for Hidden Markov Models Utilizing Students-t Mixtures, Patt. Recogn. 44(2) (2011) 295–306.

[86] H. Zhang, Q. M. J. Wu, T. M. Nguyen, Modified Students t - Hidden Markov Model for Pattern Recognition and Classification, IET Signal Process. 7(3) (2013) 219–227.

[87] J. J. Downs, E. F. Vogel, A Plant-wide Industrial Process Control Problem, Comput. Chem. Eng. 17(3) (1993) 245–255.

[88] N. Quijano, A. E. Gil, K. M. Passino, Experiments for Dynamic Resource Allocation, Scheduling, and Control: New Challenges from Information Technology-enabled Feedback Control, IEEE Contr. Sys. Mag. 25(1) (2005) 63–79.

[89] S. Verron, J. Li, T. Tiplica, Fault Detection and Isolation of Faults in a Multivariate Process with Bayesian Network, J. Process Contr. 20(8) (2010) 902–911.

[90] M. F. S. V. DAngelo, R. M. Palhares, R. H. C. Takashi, R. H. Loschi, Fuzzy/Bayesian Change Point Detection Approach to Incipient Fault Detection, Contr. Theo. Appl. 5(4) (2011) 539–551.

[91] J. T. Y. Cheung, G. Stephanopoulos, Representation of Process Trends - Part I. A Formal Representation Framework, Comput. Chem. Eng. 14(4–5) (1990) 495–510.

[92] J. T. Y. Cheung, G. Stephanopoulos, Representation of Process Trends - Part II. The Problem of Scale and Qualitative Scaling, Comput. Chem. Eng. 14(4–5) (1990) 511–539.

[93] B. R. Bakshi, G. Stephanopoulos, Representation of Process Trends - Part III. Multi-scale Extraction of Trends from Process Data, Comput. Chem. Eng. 18(4) (1994) 267-302.

[94] X. Li, M. Parizeau, R. Plamondon, Training Hidden Markov Models with Multiple Observations - A Combinatorial Method, IEEE Trans. Patt. Anal. Mach. Intell. 22(4) (2000) 371–377.

[95] N. Sammaknejad, B. Huang, Process Monitoring Based on Symbolic Episode Representation and Hidden Markov Models - A Moving Window Approach, 5th Intern. Symp. Advan. Contr. Indust. Proc. (May 28-30, 2014, Hiroshima, Japan).

[96] I. Daubechies, The Wavelet Transform, Time Frequency Localization and Signal Analysis, IEEE Trans. Inf. Theo. 36 (1990) 961-1005.

[97] L. A. Zadeh, Fuzzy Sets, Inf Contr. 8(3) (1965) 338-353.

[98] J. C. Bezdec, Pattern Recognition with Fuzzy Objective Function Algorithms, NewYork/London: Plenum Press (1981).

[99] S. E. Levinson, L. R. Rabiner, M. M. Sondhi, An Introduction to the Application of the Theory of Probabilistic Functions of a Markov Process to Automatic Speech Recognition, Bell Syst. Tech. J. 62(4) (1983) 1035-1074.

[100] P. M. Baggenstoss, A Modified BaumWelch Algorithm for Hidden Markov Models with Multiple Observation Spaces, IEEE Trans. Speech. Audio Process. 9(4) (2001) 411-416.

[101] D. Shipping, C. Tao, Z. Xianyin, W. Jian, W. Yuming, Training Second-order Hidden Markov Models with Multiple Observation Sequences, Int. Forum Comput. Sci. Technol. Appl. (2009) 25-29.

[102] W. Sun, A. Palazoglu, J. A. Romangoli, Detecting Abnormal Process Trends by Wavelet-domain Hidden Markov Models, AIChE. J. 49(1) (2003) 140-150.

[103] R. M. Turian, F. L. Hsu, T. W. Ma, Estimation of the Critical Velocity in Pipeline Flow of Slurries, Powder Technol. 51(1) (1987) 35-47.

[104] A. P. Poloski, H. E. Adkins, Deposition Velocities of Newtonian and non-Newtonian Slurries in Pipelines, U.S. Depart. Ener., Pacif. Nort. Nation. Lab. (2009).

[105] M. A. Kokpinar, M. Gogus, Critical Flow Velocity in Slurry Transporting Horizontal Pipelines, J. Hydraul. Eng. 127(9) (2011) 763-771.

[106] S. K. Lahiri, K. C. Ghanta, Artificial Neural Network Model with Parameter Tuning Assisted by Genetic Algorithm Technique: Study of Critical Velocity of Slurry Flow in Pipeline, Asia-Pacif. J. Chem. Eng. 5 (2011) 763-771.

[107] R. Durand, Basic Relationships of the Transportation of Solids in Pipes - Experimental Research, Proceedings: Minnesota Internat. Hydr. Conven., Intern. Assoc. for Hydr. Res. (1953) 89-103.

[108] R. G. Gillies, Pipeline Flow of Coarse Particle Slurries, PhD thesis, University of Saskatchewan (1993).

[109] R. G. Gillies, J. Schaan, R. J. Sumner, M. J. McKibben, C. A. Shook, Deposition Velocities for Newtonian Slurries in Turbulent Flow, Canadian J. Chem. Eng. 78(4) (2000) 704–708.

[110] C. A. Shook, R. G. Gillies, R. S. Sanders, Pipeline Hydrotransport with Applications in Oil Sands Industry, SRC Pipe Flow Technol. Cent. (2002).

[111] F. M. Hepy, Z. Ahmad, M. L. Kansal, Critical Velocity for Slurry Transport Through Pipeline, Dam Eng. 19(3) (2008) 169–184.

[112] A. A. Shahmirzad, The Effect of Fine Flocculating Particles and Fine Inerts on Carrier Fluid Viscosity, MSc. Thesis, University of Alberta (2012).

[113] M. Tanaka, G. Girard, R. Davis, A. Peuto, N. Bignell, Recommended Table for the Density of Water Between $0°C$ and $40°C$ Based on Recent Experimental Reports, Metrologia 38 (2001) 301–309.

[114] J. L. Smith, Measurements of Carrier Fluid Viscosities for Oil Sand Extraction and Tailings Slurries, MSc. Thesis, University of Alberta (2013).

[115] T. Al-Shemmeri, Engineering Fluid Mechanics, Bookboon (2012).

[116] R. B. Brown, Soil Texture, University of Florida, IFAS Extension (2003). 1–8.

[117] R. S. Sanders, A. L. Ferre, W. B. Maciejwski, R. G. Gillies, C. A. Shook, Bitumen Effects on Pipeline Hydraulics During Oil Sand Hydrotransport, The Canadian J. Chem. Eng. 78(4) (2000) 731–742.

[118] W. Li, L. Xing, L. Fang, J. Wang, H. Qu, Application of Near Infrared Spectroscopy for Rapid Analysis of Intermediates of Tanreqing Injection, J. Pharmac. Biomed. Anal. 53(3) (2010) 350–358.

[119] O. Haavisto, H. Hyötyniemi, Recursive Multimodel Partial Least Squares Estimation of Mineral Flotation Slurry Contents Using Optical Reflectance Spectra, Analytic. Chimic. Act. 642(1) (2009) 102–109.

[120] B. S. Dayal, J. F. MacGregor, Improved PLS Algorithms, J. Chemomet. 11(1) (1997) 73–85.

[121] B. S. Dayal, J. F. MacGregor, Recursive Exponentially Weighted PLS and Its Applications to Adaptive Control and Prediction, J. Process Contr. 7(3) (1997) 169–179.

[122] M. Chen, Data-driven Methods for Near Infrared Spectroscopy Modelling, MSc. Thesis, University of Alberta (2013).

[123] S. Mu, Y. Zeng, R. Liu, P. Wu, H. Su, J. Chu, Online Dual Updating with Recursive PLS Model and Its Application in Predicting Crystal Size of Purified Terephthalic Acid (PTA) Process, J. Process Contr. 16(6) (2006) 557–566.

[124] N. Sammaknejad, B. Huang, R. S. Sanders, Y. Miao, F. Xu, A. Espejo, Adaptive Soft Sensing and On-line Estimation of the Critical Minimum Velocity with Application to an Oil Sand Primary Separation Vessel, IFAC 9^{th} Intern. Symp. Advan. Cont. Chem. Proc. (June 7-10, 2015, Whistler, Canada).

[125] L. Chen, A. Tulsyan, B. Huang, F. Liu, Multiple Model Approach to Nonlinear System Identification with an Uncertain Scheduling Variable Using EM Algorithm, J. Proc. Contr. 23 (2013) 1480-1496.

[126] E. Thorsn, Multinomial and Dirichlet-Multinomial Modeling of Categorical Time Series, (2014).

[127] C. Sutton, A. McCallum, An Introduction to Conditional Random Fields, Mach. Learn. 4(4) (2011) 267–373.

Appendix A

Details of the Derivations in Chapter 3

Details of the derivations for Equation (3.21):

$$\int P(Y_k \mid \theta^{old}, C_{obs}, I_k = i) log f(Y_k \mid I_k = i, \mu_i, \Sigma_i) dY_k$$

$$= \int P(Y_k \mid I_k = i, \mu_i^{old}, \sigma_i^{old}) log((2\pi)^{-P/2} |\Sigma_i|^{-1/2} exp(-\frac{1}{2}(Y_k - \mu_i)^T \Sigma_i^{-1}(Y_k - \mu_i)) dY_k$$

$$= -\frac{1}{2} log((2\pi)^P |\Sigma_i|) - \frac{1}{2} \int P(Y_k \mid I_k = i, \mu_i^{old}, \Sigma_i^{old}) \times (Y_k - \mu_i)^T \Sigma_i^{-1}(Y_k - \mu_i) dY_k$$

Using the properties of the expected value of the quadratic form, the integral can be calculated as,

$$= -\frac{1}{2} log((2\pi)^P |\Sigma_i|) - \frac{1}{2}(tr(\Sigma_i^{-1}\Sigma_i^{old}) + (\mu_i^{old} - \mu_i)^T \Sigma_i^{-1}(\mu_i^{old} - \mu_i)) \qquad \text{(A.1)}$$

Details of the derivations for Equation (3.23):

$$\frac{\partial(\sum_{i=1}^{M} \sum_{k=t_1}^{t_\alpha} P(I_k = i \mid \theta^{old}, C_{obs}) \times (-\frac{1}{2} log((2\pi)^P |\Sigma_i|) - \frac{1}{2}(Y_k - \mu_i)^T \Sigma_i{}^{1}(Y_k - \mu_i)))}{\partial \mu_i}$$

$$+ \frac{\partial(\sum_{i=1}^{M} \sum_{k=m_1}^{m_\beta} P(I_k = i \mid \theta^{old}, C_{obs}) \times (-\frac{1}{2} log((2\pi)^P |\Sigma_i|) - \frac{1}{2}(tr(\Sigma_i{}^{1}\Sigma_i^{old}) + (\mu_i^{old} - \mu_i)^T \Sigma_i{}^{1}(\mu_i^{old} - \mu_i))))}{\partial \mu_i} = 0$$

Using the derivative properties of the quadratic form we obtain:

$$\Rightarrow \sum_{k=t_1}^{t_\alpha} P(I_k = i | \theta^{old}, C_{obs})(Y_k - \mu_i) + \sum_{k=m_1}^{m_\beta} P(I_k = i | \theta^{old}, C_{obs})(\mu_i^{old} - \mu_i) = 0$$

$$\Rightarrow \mu_i^{new} = \frac{\sum_{k=t_1}^{t_\alpha} Y_k P(I_k = i \mid \theta^{old}, C_{obs}) + \sum_{k=m_1}^{m_\beta} \mu_i^{old} P(I_k = i \mid \theta^{old}, C_{obs})}{\sum_{k=1}^{N} P(I_k = i \mid \theta^{old}, C_{obs})} \quad \text{(A.2)}$$

Details of the derivations for Equation (3.24):

$$\frac{\partial(\sum_{i=1}^{M}\sum_{k=t_1}^{t_\alpha} P(I_k = i \mid \theta^{old}, C_{obs}) \times (-\frac{1}{2}log((2\pi)^P|\Sigma_i|) - \frac{1}{2}(Y_k - \mu_i)^T {\Sigma_i}^{-1}(Y_k - \mu_i)))}{\partial \Sigma_i}$$

$$+\frac{\partial(\sum_{i=1}^{M}\sum_{k=m_1}^{m_\beta} P(I_k = i \mid \theta^{old}, C_{obs}) \times (-\frac{1}{2}log((2\pi)^P|\Sigma_i|) - \frac{1}{2}(tr({\Sigma_i}^{-1}\Sigma_i^{old}) + (\mu_i^{old} - \mu_i)^T {\Sigma_i}^{-1}(\mu_i^{old} - \mu_i))))}{\partial \Sigma_i} = 0$$

Using the derivative properties of the trace, determinant and inverse we obtain:

$$\Rightarrow \sum_{k=t_1}^{t_\alpha} P(I_k = i|\theta^{old}, C_{obs})(-\Sigma_i^{-1} + \Sigma_i^{-1}(Y_k - \mu_i^{new})(Y_k - \mu_i^{new})^T\Sigma_i^{-1})$$

$$+\sum_{k=m_1}^{m_\beta} P(I_k = i|\theta^{old}, C_{obs})(-\Sigma_i^{-1} + \Sigma_i^{-1}\Sigma_i^{old}\Sigma_i^{-1} + \Sigma_i^{-1}(\mu_i^{old} - \mu_i^{new})(\mu_i^{old} - \mu_i^{new})^T\Sigma_i^{-1}) = 0$$

$$\Rightarrow (\Sigma_i)^{new} = \frac{\sum_{k=t_1}^{t_\alpha}(Y_k - \mu_i^{new})(Y_k - \mu_i^{new})^T P(I_k = i \mid \theta^{old}, C_{obs})}{\sum_{k=1}^{N} P(I_k = i \mid \theta^{old}, C_{obs})}$$

$$+\frac{\sum_{k=m_1}^{m_\beta}(\Sigma_i^{old} + (\mu_i^{old} - \mu_i^{new})(\mu_i^{old} - \mu_i^{new})^T)P(I_k = i \mid \theta^{old}, C_{obs})}{\sum_{k=1}^{N} P(I_k = i \mid \theta^{old}, C_{obs})} \quad \text{(A.3)}$$