Hierarchical Monitoring and Probabilistic Graphical Model Based Fault Detection and Diagnosis

by

Mengqi Fang

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

As the rapid development of modern industry, data based fault detection and diagnosis for industrial processes have become increasingly critical to ensure process safety and product quality. To effectively make use of underlying features of process data, multiple data based fault detection and diagnosis algorithms have been developed, among which the multivariate statistical process monitoring (MSPM) algorithms and the probabilistic graphical model based algorithms have been widely used. Through unsupervised training, the conventional MSPM algorithms have the advantage of simplicity but do not use the labeled fault information in the training phase. On the other hand, the probabilistic discriminative classifiers are supervised models and trained with label information. This thesis starts from solving a practical industrial fault detection and diagnosis problem based on the unsupervised MSPM approaches. Then to fully make use of both process observations and fault information, a supervised probabilistic discriminative framework, namely conditional random field (CRF) model, is introduced and then extended to deal with various practical scenarios and challenges.

Specifically, as a practical study on real-time fault detection and diagnosis, an early flare event prediction for a refinery process is first considered. Different operating conditions and production requirements from different process units result in hybrid data characteristics, therefore a single fault detection and diagnosis algorithm is not sufficient to deal with the problem. In this sense, a hierarchically distributed framework is designed to solve this problem, with two integrated and interactive monitoring layers to detect faults and track the root causes. Based on this layout, the majority of flare events can be successfully predicted with limited false positives.

Additionally, when fault label is available, supervised probabilistic classifiers are further explored. As a discriminative counterpart of the widely used hidden Markov models (HMMs), the linear-chain CRF (LCCRF) is introduced with demonstrated superior fault diagnosis performance to the HMMs. Then three practical challenges are addressed by extending the conventional LCCRF frameworks to variants of CRFs. First, to deal with the missing data problem, a marginalized CRF model is developed with a proposed efficient inference strategy. Second, to solve the feature selection and online adaption problem for operating mode diagnosis, a two-stage hidden CRF (HCRF) structure is proposed by combining the max-margin trained HCRF and LCCRF into a hierarchical framework. Third, to address the fault detection and diagnosis problem for processes with multiple operating conditions, a switching CRF model is proposed to deal with the variations of the process conditions, by extending unitary LCCRF to multiple LCCRFs.

This thesis aims to provide improved solutions to the fault detection and diagnosis problems in practical processes. As shown through multiple case studies of different chapters, the effectiveness of the proposed algorithms is demonstrated.

Preface

This thesis is an original work by Mengqi Fang under the supervision of Dr. Biao Huang and is funded in part by Natural Sciences and Engineering Research Council (NSERC) of Canada. Most of materials have been published by the author in peer-reviewed journals or conference proceeding, listed below:

1. Mengqi Fang, Fadi Ibrahim, Hariprasad Kodamana, Biao Huang, Noel Bell, and Mark Nixon. Hierarchically distributed monitoring for the early prediction of gas flare events. *Industrial & Engineering Chemistry Research*, 58(26):11352–11363, 2019. (*Chapter 3*)

2. Mengqi Fang, Hariprasad Kodamana, Biao Huang, and Nima Sammaknejad. A novel approach to process operating mode diagnosis using conditional random fields in the presence of missing data. *Computers & Chemical Engineering*, 111:149–163, 2018. (*Chapter 4*)

3. Mengqi Fang, Hariprasad Kodamana, and Biao Huang. Real-time mode diagnosis for processes with multiple operating conditions using switching conditional random fields. *IEEE Transactions on Industrial Electronics*, 67(6):5060–5070, 2020. (*Chapter 6 - Complete version*)

4. Mengqi Fang, Hariprasad Kodamana, and Biao Huang. Switching conditional random field approach to process operating mode diagnosis for multi-modal processes. In *2018 IEEE Conference on Decision and Control (CDC)*, pages 5146–5151. IEEE, 2018. (*Chapter 6 - Short version*)

Acknowledgements

First of all, I am bringing tons of gratitude to thank my supervisor, Prof. Biao Huang, for all his supervisions and guidance during my entire Ph.D. life. As an excellent professor, he is always being an example to me for not only his rigorous attitudes and intelligence in researches but also his encouraging and positive views to life and challenges. I would like to show my sincere acknowledgement to his patience when explaining the concepts and ideas that I do not understand completely, and his inspirations and trust when I am lack of confidence and even good progresses. All of these bring me consistent encouragements and make me always keep passion and curiosity to explore and extend my understanding to the fields that I am working on. Frankly, Ph.D. journey is indeed a tough and long journey, and I am so fortunate to be shaped and equipped with the help from my supervisor to be more mature to the future life.

I would like to thank our industrial partners, especially Emerson, Suncor Energy and Spartan Controls Automation Centre, for their highly supports to the process data, process knowledge, site experiences and the feedbacks on all the proposed algorithms. I greatly appreciate Noel Bell and Mark Nixon for their contributions to the flare event prediction problem solution and the revisions of the manuscripts.

I would like to especially thank my former and current group members Dr. Hariprasad Kodamana, Dr. Nima Sammaknejad and Dr. Fadi Ibrahim for their academic supports to make me grow and learn faster and more from almost blank. And I also want to acknowledge my wonderful colleagues and friends in the CPC group to make my Ph.D. life vivid and full of hope. Additionally, many thanks to the students during my teaching assistant period and the co-op students I have ever worked with, I learned a lot from the interactions with them and their expectations.

I would like to thank the Alberta Innovates and National Science and Engineering Research Council of Canada for the financial supports and the University of Alberta to provide great opportunity for my Ph.D. experiences.

Last but not least, I would like to gratefully thank my beloved parents. I know what they sacrificed are much more than that I perceived. The protections and devotions from them are silent but influential, regardless of the distance. Finally, I would like to thank my lovely husband, Xin Zhao, for his selfless and infinite supports, endless tolerances and love, which make me fearless to face all the upcoming unknowns.

Contents

List of Tables

List of Figures

List of Abbreviations and Notations

Abbreviations

BPNN	Back propagation neural network
CMLE	Conditional maximum likelihood estimation
CRF	Conditional random field
CSTR	Continuous stirred tank reactor
cwt	Continuous wavelet transform
EM	Expectation maximization
E-step	Expectation step
FFT	Fast Fourier transform
FGRS	Flare gas recovery system
FT	Fourier transform
HCRF	Hidden conditional random field
HMM	Hidden Markov model
ICA	Independent component analysis
KL divergence	Kullback-Leibler divergence

L-BFGS	Limited memory BFGS
LCCRF	Linear-chain CRF
MLE	Maximum likelihood estimation
MMHCRF	Max-margin HCRF
M-step	Maximization step
MSPM	Multivariate statistical process monitoring
MWMSAPCA	Moving window multi-scale adaptive PCA
PC	Principal component
PCA	Principal component analysis
PLS	Partial least squares
PV	Process variable
QP	Quadratic programming
RBC	Reconstruction-based contribution
RFE	Recursive feature elimination
rRBC	Relative reconstruction-based contribution
SCRF	Switching conditional random field
SFA	Slow feature analysis
SPM	Statistical process monitoring
SVM	Support vector machine
VB	Variational Bayesian
WT	Wavelet transform

Notations

α	Forward variable
$\alpha_{n,h}$	Weights in the dual function
β	Backward variable
γ_t	Intermediate effect variable for $O_t^{(mis)}$
ξ_n	Slack variable in max-margin training
ξ_t	Intermediate forward variable containing $O_t^{(mis)}$
ζ_l	Support variable in Dirichlet distribution
η_t	Intermediate effect variable for all possible missing observations at time t
η_{y_t}	Concentration parameters of Dirichlet distribution
$\boldsymbol{\Lambda}_n$	Eigenvalue matrix of the n^{th} unit
λ_k	Weighting factor of the transition feature function
μ_d	Mean value of the transition duration
μ_m	Weighting factor of the emission feature function
σ_d	Standard deviation of the transition duration
σ_i	Validity width of the i^{th} scheduling variable
$\tau_{\boldsymbol{X}^{(old)}}$	Posterior probability of the variable $\boldsymbol{X}$ in the previous EM iteration
Θ	Unknown parameters
$\phi_{a,b}$	Basis function for wavelet transform
φ_t	Intermediate feature function at time t
$\boldsymbol{\Omega}$	Slowness matrix in SFA

ω_k	Slowness of the k^{th} slow feature
D_c	Complete data
$D_{KL}(\cdot)$	KL divergence
D_m	Missing data
D_o	Observed data
d_l	Small moving window length of the l^{th} sample
d_{tr}	Transition duration
E_m	Emission feature function in CRF
$\boldsymbol{E}_n$	Residual matrix of the n^{th} unit
h_t	Operating mode at time t
I_t	Operating condition at time t
$L(\cdot)$	Lower bound
L_l	Liquid level of the liquid seal unit
L_t	Moving window length at time t
$l(\cdot)$	Log likelihood
n_{s_d}	Length of the $s_d^{(th)}$ stationary period
n_{t_r}	Half length of the $t_r^{(th)}$ transition period
$\boldsymbol{O}_{mis}$	Missing components in discrete observations
$\boldsymbol{O}_{obs}$	Observed components in discrete observations
O_t	Observation at time t
P_l	Pressure of the liquid seal unit

$\boldsymbol{P}_n$	Loading matrix of the n^{th} unit
$Q(\cdot)$	Q function
Q	Q statistic
Q_i	Label of the i^{th} instance
$q(\cdot)$	Variational posteriors
$q_{t,n}$	The weighted log likelihood of forward variable at time t
R_t	The summation of the weighted log likelihood up to time t
S_i	The i^{th} fixed operating point
$\boldsymbol{S}_n$	Score matrix of the n^{th} unit
S_t	Scheduling variable at time t
$\boldsymbol{s}_n$	Extracted PC from new sample of the n^{th} unit
T^2	Hotelling's T^2 statistic
T_f^2	Hotelling's T^2 statistic of the faster features
T_k	Transition feature function in CRF
T_s^2	Hotelling's T^2 statistic of the slower features
$\boldsymbol{V}$	Slow feature matrix
$\boldsymbol{v}_f$	Faster features in SFA
$\boldsymbol{v}_s$	Slower features in SFA
$\boldsymbol{W}$	Coefficient matrix in SFA
$W(\alpha)$	Evaluation criterion for variable selection
$\boldsymbol{W}_f$	Faster feature governing matrix

$\boldsymbol{W}_s$	Slower feature governing matrix
$\boldsymbol{X}_n$	Process variables of the n^{th} unit
$X_t(L_t)$	First-stage HCRF output features at time t
$\boldsymbol{x}_n$	New sample of the n^{th} unit
$\boldsymbol{Y}_{PC}$	Integrated PC matrix of the bottom layer
$\boldsymbol{Y}_t$	Observation sequence used in CRF at time t
$\boldsymbol{Y}_t^{(mis)}$	Missing components in $\boldsymbol{Y}_t$
$\boldsymbol{Y}_t^{(obs)}$	Observed components in $\boldsymbol{Y}_t$
y_t	Auxiliary labels of the two-stage CRF
$Z(\cdot)$	Partition function of CRFs

Chapter 1

Introduction

1.1 Motivation

Modern industries are composed of large-scale facilities and involve highly complicated networks, with thousands of control loops and process variables (PVs). To ensure smooth process operations, increase production safety and minimize maintenance costs, employing effective and accurate process monitoring techniques is essential and has inspired many relevant academic and practical researches over the last decades. Generally, process monitoring techniques are utilized to detect, diagnose and remove faults occurring in the processes [1], where a fault is defined as an unpermitted deviation of at least one characteristic property or variable of the system [2].

Typically, process monitoring is composed of four components, namely, fault detection, fault identification, fault diagnosis and fault recovery [1]. Fault detection aims to determine whether a process has a fault. Early fault detection can be used to generate anticipated warnings for process operators to take preventive actions. Fault identification is to identify the PVs most relevant to diagnosing the fault, which can also be treated as a preliminary procedure of fault diagnosis. Fault diagnosis is to diagnose the causes of the fault. Finally, based on the above fault analysis results, interventions can be performed to eliminate the fault, called fault recovery [3]. This thesis mainly focuses on the fault detection and diagnosis problems.

A complete fault diagnosis system can be considered as a mapping from process measurements to fault classes, with intermediate transformations from the measurement space to the feature space, and then the feature space to the fault class space [4]. Compared with the raw measurement space, the transformed feature space has enhanced discriminative capacity and therefore is normally able to improve fault diagnosis performance. According to the available process information and process characteristics, there have emerged various algorithms to perform feature transformation from process knowledge, based on which the transformation between feature space and fault class space can be established. Fig. 1.1 summarizes the categories of the main stream fault detection and diagnosis algorithms. As shown in Fig 1.1, the existing approaches may be classified into three different categories, namely, knowledge-based, analytic-based and data-based algorithms [1]. Knowledge-based approaches are established based on qualitative models, such as causal analysis or expert systems, and analytic-based algorithms usually employ mathematical models built from first principles, while the data-based algorithms are directly constructed from the process measurements [1]. However, considering practical industrial processes, the complicated mechanisms and interactions among all the existing PVs make the first principle modeling difficult, and the lack of a complete understanding of a process also degrades the performance of knowledge-based models. On the other hand, owing to the availability of a large amount of stored process data, data-based approaches offer a potential effective alternative solution to the other two types of approaches.

Among the existing data based fault detection and diagnosis approaches, multivariate statistical process monitoring (MSPM) approaches play an important role due to their capability to handle high dimensional process observations, by compressing the high dimensional raw measurements into lower dimensional latent features. However, the effectiveness of the MSPM algorithms can be achieved only when the corresponding assumptions are satisfied.

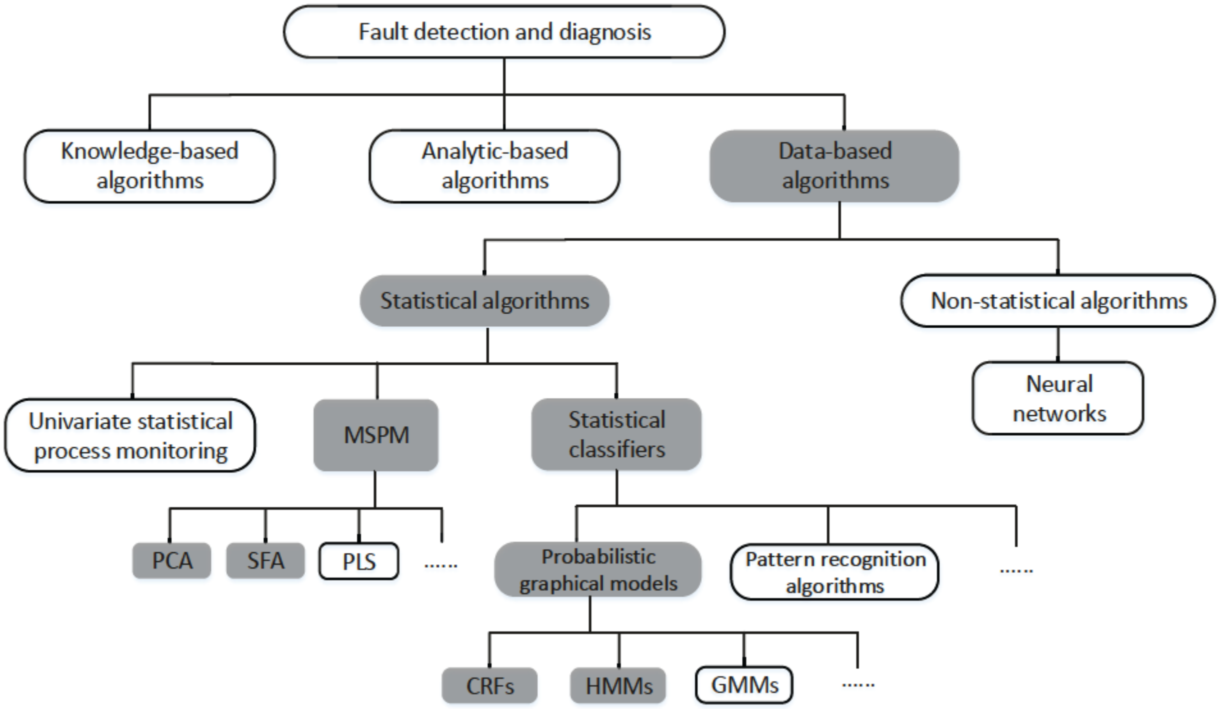

Fault detection and diagnosis
Knowledge-based algorithms
Analytic-based algorithms
Data-based algorithms
Statistical algorithms
Non-statistical algorithms
Neural networks
Univariate statistical process monitoring
MSPM
Statistical classifiers
PCA
SFA
PLS
......
Probabilistic graphical models
Pattern recognition algorithms
......
CRFs
HMMs
GMMs
......

Figure 1.1: Overview of the fault detection and diagnosis algorithms discussed in this thesis, as denoted in the shaded nodes

Generic assumptions of most MSPM algorithms include the unimodal, time-invariant and stationary characteristics of industrial process. In practice, processes are more complicated than the assumptions of the MSPM algorithms, and such processes can exhibit hybrid and time-varying characteristics. In most cases, a single MSPM algorithm is not always able to achieve effective fault detection and diagnosis performance. This thesis starts from solving a real industrial fault detection and diagnosis problem, namely, to provide early flare event detection for a refinery process using MSPM approaches. The refinery process under consideration has a large scale and is composed of different units with a large amount of process observations but with limited process knowledge. The characteristics of this process, such as high dimensionality, time-varying, non-stationary and inconsistency, make the early detection problem more challenging. Various MSPM algorithms are explored and compared, and a systematic hierarchical framework is proposed to integrate different MSPM algorithms to extract informative signatures to predict the faulty events.

Most of the existing MSPM algorithms are restricted to process monitoring when the process is operated under a single operating mode. However, a process can have multiple operating modes switching among each other. The statistical properties of different operating modes can vary significantly, so that the conventional MSPM algorithms may not be able to deal with them. To address the multimodal problem, extensions have been made based on the conventional MSPM approach. One straightforward solution is to use multiple MSPM models to describe more than one operating mode. However, such extension might not be able to well model the process dynamic and uncertainty brought by the operating mode switching. As a result, the probabilistic models are considered as a more expressive mathematical tool for multiple operating mode modeling in this thesis. Meanwhile, for continuous processes, temporal correlations are naturally encoded in the collected process datasets. Such temporal correlations need to be considered when building a model for process monitoring. In recent decades, the hidden Markov models (HMMs) have been widely employed to solve fault detection and diagnosis problems for processes with temporal correlations and multiple operating modes [5]. However, HMMs have two inherent independence assumptions, namely, (i) in first order HMMs, the current state is assumed to be only dependent on the state immediately prior to it and independent of all the other previous information; (ii) the current observation is only dependent on the current state and independent of all the other previous information. Here, the state of HMMs is equivalent to the operating mode. These independence assumptions will degrade the performance of HMMs when they are violated. In this sense, under the probabilistic framework, while preserving the advantages of HMMs, a more flexible modeling structure is deployed to enhance the fault detection and diagnosis performance for complicated processes.

To reduce the restriction of HMMs and facilitate the feature extraction, conditional random field (CRF) model is introduced to solve the fault detection and diagnosis problems for processes with multiple operating modes and complicated temporal correlations. The fault detection and diagnosis performance of CRFs has been demonstrated to be superior to HMMs, and extensions based on the conventional CRF models are proposed to solve the missing data, variable selection and multiple operating condition problems. This forms the second part of this thesis.

1.2 Literature Review

As illustrated in Fig. 1.1, based on the available process information that can be utilized for process monitoring, there are three categories of monitoring strategies, including knowledge-based, analytic-based and data-based algorithms. Data-based algorithms are flourishing over decades and have become a promising means to deal with complicated process monitoring problems without the need to know complete process knowledge. As the core problem for data-based process monitoring is to extract discriminative features from the raw process observations, numerous algorithms have been developed to address different process data characteristics, among which the conventional MSPM and probabilistic graphical models have shown promising potentials and attracted increasing attentions from researchers. In this section, the data-based fault detection and diagnosis algorithms are first reviewed, and then the conventional MSPM and probabilistic graphical model based algorithms are revisited and summarized subsequently.

1.2.1 Process Data Based Fault Detection and Diagnosis

Data based fault detection and diagnosis algorithms are developed on the basis of process historical data, without the requirement of prior process knowledge. According to different mathematical techniques utilized, the existing data based fault detection and diagnosis approaches can be generally classified into statistical and non-statistical categories [6]. In the statistical category, random disturbances are considered and the process is treated as a stochastic process. The normal process operations are considered to follow particular statistical assumptions and modeled by selected probabilistic distributions, where the unknown parameters are identified by historical data. During online monitoring, once a fault occurs, process observations will experience unexpected changes and deviate from the predefined distributions of normal operations. Such deviations can also be used for fault diagnosis. By expanding the existing works, the following statistical process monitoring (SPM) algorithms are explained.

Univariate Statistical Process Monitoring

As one type of the earliest SPM strategies, the univariate control charts are proposed to monitor process production performance online, so that timely correction measures can be done to bring the process back to normal. Shewhart control charts [7] and the cumulative sums charts [8] are two typical examples.

However, limited by the univariate property, such control charts can hardly accommodate the correlations of multidimensional PVs, resulting in misleading monitoring results. Therefore, the MSPM techniques are proposed.

Multivariate Statistical Process Monitoring

Multivariate statistical approaches have the capability to excavate the latent information underlying the high dimensional PVs. The main objective is to transform a number of correlated PVs into a smaller set of uncorrelated components, which are monitored to detect process abnormalities. As typical examples, the principal component analysis (PCA) [9] and partial least squares (PLS) [10] approaches have been first proposed to perform dimension reduction, and then some statistical metrics, such as Hotelling's T^2 statistic [11] and squared prediction error, are computed for real-time process monitoring. However, the traditional MSPM algorithms have inherent statistic assumptions, such as unimode, Gaussian distributions and linearity, etc.. For example, the traditional PCA algorithm assumes that the analyzed data are collected from a linear process under a stationary operating condition [12]. As improvements, various extended works [13, 14, 15] have been proposed.

Statistical Classifiers

Essentially, fault diagnosis can be treated as a classification problem, where the classical statistical pattern recognition framework is adopted as potential solutions [6]. Meanwhile, in order to deal with process uncertainty, the probability theory is introduced. The probabilistic framework provides sufficient flexibility to model the data characteristics under multiple operating modes, and it is also able to model the process dynamics and temporal correlations in the process observations. Gaussian mixture models [16] and HMMs [17] are two typical examples.

Moreover, conventional MSPM strategies have been extended to probabilistic counterparts, such as probabilistic PCA [18]. In this way, by integrating with the above multimodal modeling, probabilistic counterparts of conventional MSPM approaches can be used to solve process monitoring problems in multiple operating modes. Among all the existing probabilistic frameworks, probabilistic graphical models have the simplicity of modeling, generalization and interpretation, and demonstrated effectiveness in process monitoring. Therefore, it is becoming a promising research subject in recent decades. The details of probabilistic graphic model based fault detection and diagnosis will be explained in the following sections.

Non-statistical Classifiers

Different from statistical fault detection and diagnosis algorithms, the non-statistical approaches do not rely on statistical assumptions of process observations. More straightforwardly, parameterized classifiers are established, where a typical example is neural network [6]. Various network architectures are developed to solve different fault diagnosis problems, in both supervised and unsupervised manners. The non-statistical algorithms mainly depend on parameterized models to describe multiple operation conditions, and some feature extraction algorithms, for instance, wavelet analysis, are used for data pre-processing.

1.2.2 Conventional MSPM Approaches Based Fault Detection and Diagnosis

The conventional MSPM algorithms are developed based on statistical theory to extract latent information from process observations with reduced dimensions. In practice, apart from high dimensionality, the collected data from industrial processes usually have multiple characteristics, for example, non-Gaussian distribution, nonlinear correlation, randomness, multimodal and dynamic characteristics, etc.. To address this problem, different MSPM approaches need to be established for accurate discriminative feature extractions.

Due to the limitation of standard PCA with inherent assumptions, several improved algorithms have been proposed. For instance, to address process dynamics, dynamic PCA has been proposed by introducing lags into the observations [19]. Kernel PCA was proposed [20] by introducing a kernel function to make the nonlinear observations tractable with linear approach. For non-Gaussian process observations, the independent component analysis (ICA) algorithm was developed to generate appropriate latent components for further process monitoring [21, 22]. These improvements are motivated by modeling the characteristics of process observations.

Another way of improvement is to adjust criteria of feature extraction for more accurate fault detection and diagnosis. For example, the standard PCA performs dimension reduction while preserving significant variability in the extracted features. Unlike standard PCA, the linear discriminant analysis achieves dimensionality reduction while preserving maximum discriminative information for fault classification [1]. By introducing dynamic autocorrelations into the latent variables during modeling, the slow feature analysis (SFA) approach is formulated as a state space model and then the latent variables are extracted and sorted by the varying velocity [23]. By using SFA model, normal operating condition deviations can be differentiated from the actual faults with dynamic anomalies. Similar works can be found in [24].

Even though there exist lots of MSPM algorithms, it is not possible to find one MSPM approach that is effective to all process monitoring problems. The process complexity and time-varying properties make fault detection and diagnosis problems even more difficult, therefore the development of efficient hybrid process monitoring systems becomes an attractive research subject [6].

1.2.3 Probabilistic Graphical Model Based Fault Detection and Diagnosis

Probabilistic graphical models are formulated by integrating graph theory and probability theory into a unified modeling framework [25]. A probabilistic graphical model can be specified by the model graphical structure and a set of local functions [26]. The graphical model structure qualitatively depicts the correlations among the selected random variables, and the local functions are designed to quantitatively describe the random variable dependences. Effective probabilistic inference and learning strategies can be employed to obtain insights of the process operating status.

Probabilistic graphical models can be classified into diversified categories from different perspectives [26]. From modeling perspectives, probabilistic graphical models can be separated into generative and discriminative models [27]. Considering temporal correlations in a process, probabilistic graphical models can be classified into static and dynamic models. Extending static probabilistic graphical models to a linear chain or more complicated structures is one way to develop dynamic probabilistic graphical models. In this way, process uncertainty, temporal correlations and multimodal operating situations can be formulated in a unified model for fault detection and diagnosis.

Probabilistic Generative Model Based Fault Diagnosis

Probabilistic generative models formulate the joint probability of selected random variables and encode the correlations among random variables into the obtained models. Based on the learned model, inference can be performed to estimate latent variables from historical process observations, where latent variables can be discrete operating modes or continuous latent features. Bayesian networks include a wide range of probabilistic generative models and have been employed to solve many fault diagnosis problems in complicated industrial processes [28].

Bayesian networks start from modeling the joint probabilistic distributions of random variables and rely on certain independence assumptions to simplify the joint probability. Static Bayesian networks, for example, Bayesian classifiers and probabilistic mixture models, assume samples from different sampling instants independent with each other, which are used to solve multimodal process monitoring problems. Specifically, finite Gaussian mixture models [29, 30], mixture probabilistic PCA models [31] and mixture Bayesian regularization based probabilistic PCA models [32] have been proposed to describe the process multimodal property by formulating the process observations with different probabilistic distributions.

However, such mixture models do not concern about process dynamics, such as operational mode transitions, and most of them commonly assume that process data follow a Gaussian distribution in one operating mode. As an improvement, dynamic Bayesian networks are proposed by introducing temporal correlations into random variables. In this way, linear Gaussian state-space model is obtained by formulating autocorrelations in the latent variables of probabilistic PCA model [33]. Autoregressive dynamic latent variable models have been proposed to capture static and dynamic correlations in raw data simultaneously for process monitoring [34]. The dynamics of continuous latent features are generally represented in a state space form, and the dynamics introduced into the discrete latent variables are usually formulated as a Markov process, such as operating mode transitions.

By such a formulation, a series of switching models have been proposed. For example, multiple autoregressive dynamic latent variable models have been integrated in a switching framework to monitor processes with multiple operating modes [35]. Similarly, the static mixture probabilistic PCA model can also be extended into a dynamic form with switching mechanism among the subcomponents [36]. More generally, these formulations can be assorted into HMM frameworks.

HMM is a probabilistic sequence model for estimating the joint probability distribution of hidden states and observations, where states correspond to different operating modes. HMMs have the advantages of simplicity and extensibility, and perform well at modeling state transitions. Considering the industrial data quality, various feature extraction algorithms are combined with HMMs to achieve a satisfactory process monitoring performance, as mentioned above. Moreover, in dynamic process monitoring, HMMs are effective methods to deal with missing data, outlier issues and time-varying transitions [17, 37]. However, probabilistic generative modeling makes it necessary for HMMs to require explicit probabilistic distributions to model the observations, and to simplify the factorization, two independence assumptions need to be satisfied in HMMs [27]. Even though many extended HMMs are proposed to relax the inherent assumptions of the conventional HMMs, for example, autoregressive HMMs [38] and higher-order HMMs [39], etc., such extensions bring up increased computational loads and make the modeling, training and inference of HMMs more complex. Moreover, even with the extended HMMs, one still needs to find appropriate probabilistic distributions to model process observations, which might degrade the fault diagnosis accuracy once the proposed probabilistic model is not accurate.

As a result, probabilistic discriminative models are proposed to compensate the potential drawbacks of generative models. Instead of modeling the joint probability, probabilistic discriminative models directly formulate and optimize conditional probability. In this sense, no explicit probabilistic distributions are required to model the process observations, and the independence assumptions for joint probability factorization can also be relaxed.

Probabilistic Discriminative Model Based Fault Diagnosis

Compared with probabilistic generative models, there are fewer existing works on the probabilistic discriminative model based fault diagnosis. Similar to Bayesian networks, probabilistic discriminative models can also be classified into static and dynamic categories. A typical static probabilistic discriminative model is logistic regression model, which has been employed to solve the rolling element bearing fault diagnosis problem and shows superior performances compared with artificial neural networks and support vector machine (SVM) based algorithms [40]. On the basis of the logistic regression model, a statistical feature selection approach has been proposed to aid fault diagnosis in the presence of massive historical data [41]. By incorporating temporal correlations, dynamic probabilistic discriminative models are proposed, for example, CRFs. CRFs are a type of probabilistic discriminative counterpart of HMMs, and have demonstrated better performance than HMMs in many application fields, such as natural language processing, image processing and speech recognition, etc. [42, 43, 44]. In bearing fault classification problem, the CRF scheme has a better fault classification performance than HMMs [45]. A neighbourhood hidden CRF (HCRF) model is utilized to address the condition monitoring problem of large scale wireless sensor networks [46] with a demonstrated superior performance to HMMs. However, unlike HMMs, CRFs have been seldom used in the process monitoring domain. This fact motivates the works of this thesis. In the following section, contributions of this thesis will be explained.

1.3 Thesis Outline

In Chapter 2, the main mathematical backgrounds and techniques used in this thesis are illustrated and explained in details. As a major component of this thesis, before the in-depth discussions of the CRFs based fault detection and diagnosis approaches, the modeling, learning and inference procedures of CRFs are formulated. Subsequently, in parallel with the traditional maximum likelihood estimation (MLE), the expectation maximization (EM) and variational Bayesian (VB) strategies are explained as two alternative parameter estimation approaches when latent variables are introduced into CRF models. Both EM and VB algorithms have the capability to handle latent variables, with the unknown model parameters and the posterior probabilities of the latent variables calculated iteratively until convergence. However, EM and VB algorithms have different properties and perform differently when dealing with different problems. Therefore, as preliminaries, EM and VB algorithms will be compared.

In Chapter 3, a hierarchically distributed MSPM approach is proposed and an early flare event prediction and diagnosis problem from a real refinery process with large scale plantwide settings is analyzed and solved. The limited access to process knowledge and availability of large amount process historical data make this practical problem a good real world template to develop and evaluate data based process monitoring strategies. In this work, the challenges of the early flare prediction problem and characteristics of the refinery process are first presented. Furthermore, the performance and limitations of the existing MSPM algorithms are analyzed. Finally, a hierarchical monitoring framework is designed and multiple conventional MSPM algorithms are integrated into this proposed framework for early fault detection and diagnosis. For application purposes, an adaptive online strategy is developed to improve the flare prediction accuracy and reduce the false positives. The early flare event prediction performance of the proposed algorithm is demonstrated through this real industrial application, and fault diagnosis is developed under a hierarchical monitoring structure. Because the flare event prediction problem has a limited number of faulty events available, the proposed unsupervised process monitoring approaches are appropriate solutions. However, in other cases, fault labels are accessible as references, and supervised process monitoring approaches can learn the relationship between the process observations and fault labels directly, potentially contributing to a more efficient fault diagnosis. Therefore, the supervised process monitoring algorithms are further explored in the following chapters.

In Chapter 4, a supervised process monitoring algorithm based on CRF is first proposed with consideration of missing observations. To begin with, CRFs and HMMs are compared, in which the inherent independence assumptions of HMMs are illustrated. Being a discriminative modeling approach, it is proven that CRFs are reduced to HMMs in special cases. Second, by extending from fault diagnosis to mode diagnosis, a LCCRF based process operating mode diagnosis algorithm is proposed. Third, the standard LCCRF structure is extended to solve process monitoring problems in presence of missing observations. The expressive strength of CRFs comes from their ability to extract more complicated features, and in return, such advantage also increases the challenges. Therefore, an efficient training and inference algorithm of the proposed CRF model is developed. The solution is established on the basis of MLE, aiming to directly maximize the conditional probability of the proposed CRF model. Finally, a numerical case study and a pilot scale experiment are conducted to demonstrate performance of the proposed algorithm.

In Chapter 5, information redundancy along with a large amount of PVs in industrial processes is taken into consideration. An outstanding advantage of LCCRF is that it has the capability to model complicated and correlated features without the need to specify their probabilistic distributions. However, fault-irrelevant variables may also be used and treated equivalently to the fault-relevant variables during the CRF modeling. As a result, undesired disturbances captured by the redundant features may have undesired impacts on the final fault diagnosis performance. To solve this problem, a two-stage HCRF model is developed. In the first-stage of HCRF, the max-margin training strategy is employed to discriminate multiple operating modes, and by recursively eliminating fault-irrelevant variables, the most relevant variables can be selected during the first-stage training process. The second-stage HCRF is then followed by adapting the monitor to the dynamic changes of the process with time-varying model structure. Therefore, switchings among process operating modes can be captured timely. To demonstrate the performance of the proposed algorithm, a simulation study is conducted with comparisons to the conventional algorithms and is demonstrated superior performance.

In Chapter 6, the monitor is extended from the process with single operating condition to that with multiple operating conditions. Considering the fact that multiple operating modes can exist at any specific operating condition, single CRF model is not sufficient to handle such more complex scenario. Therefore, an extension from a single CRF to multiple CRFs based monitoring framework is proposed, with one CRF model designed for each operating condition. As operating conditions switch between each other in a continuous process, CRF models switch alongside with switching of the corresponding operating conditions. Second, based on the proposed switching CRF (SCRF) model, an effective training algorithm is developed by using the EM algorithm introduced in Chapter 2. The presence of the unknown operating conditions introduces latent variables in a conditional probabilistic framework and makes the training process more complicated. As a result, a simplified SCRF parameter estimation strategy is developed by introducing another indicative latent variable. Third, for online implementation, an online inference approach based on the proposed SCRF model is formulated. The performance of the developed SCRF framework is evaluated through a simulated continuous stirred tank reactor (CSTR) process and a hybrid tank experimental setup.

In Chapter 7, the entire thesis is summarized and future works are presented based on the complete works and the practical needs for further improvements.

1.4 Main Contributions

As a brief summary, the main contributions of this thesis are listed below:

1. Developed a hierarchically distributed process monitoring framework and solved a practical early flare event prediction and diagnosis problem;

2. Performed a systematic analysis and comparison between the widely used HMMs and CRFs, then proposed a CRF based process monitoring framework as a fundamental structure of this thesis. Extended the standard CRF to a marginalized CRF structure to handle missing observations;

3. By making use of the discriminative modeling advantages of CRFs, proposed a two-stage CRF structure considering both variable selection and time-varying adaptation to process dynamics;

4. Developed a switching CRF model based on the standard CRFs to address the monitoring issues in a process with multiple operating conditions.

Chapter 2

Mathematical Foundations

In this chapter, the main mathematical techniques employed throughout this thesis will be introduced. Starting from the discriminative model formulation, CRFs are explained. Then, the EM and VB algorithms are formulated in details, and finally comparisons between these two algorithms are made.

2.1 Conditional Random Fields

In general, CRFs belong to probabilistic model category that is created to interpret dependencies among different random variables. CRFs are first proposed by *Lafferty et al.* [43] in 2001 for the purpose of sequence labeling and segmentation. CRFs are a type of probabilistic discriminative models with variable dependencies embedded into a graphic structure. On the basis of CRF formulation, various applications can be performed, such as speech recognition, image segmentation and information extraction from an article, etc.. Owing to the discriminative formulation, CRFs have the capability to model complicated dependencies among variables. At the expense of increasing model complexity, efficient training and inference algorithms are necessary to make CRFs tractable. To develop a CRF based algorithm, establishing the graphical model, training and inference are three basic problems to solve, which will be reviewed in the following section.

2.1.1 Formulation of CRFs

As a probabilistic graphical model, a CRF can be specified by two components, namely, a graphic structure $\mathcal{G}(\mathcal{V}, \mathcal{E})$ that qualitatively defines the variable dependencies and a set of local functions that quantify the variable correlations [26]. Here, $\mathcal{V}$ and $\mathcal{E}$ denote the vertex and edge sets of graph $\mathcal{G}$, respectively. More specifically, as shown in Fig. 2.1, let $\boldsymbol{Q}$ and $\boldsymbol{O}$ denote two sets of random variables, where $\boldsymbol{Q}$ represents a finite label sequence and $\boldsymbol{O}$ represents observations. In the graphical structure $\mathcal{G}(\mathcal{V}, \mathcal{E})$, vertex set $\mathcal{V}$ is composed by the label space $\boldsymbol{Q} = (\boldsymbol{Q}_v)_{v \in \mathcal{V}}$, and $\mathcal{E}$ represents the edges, including all the connections among $\boldsymbol{Q}$ and $\boldsymbol{O}$. Then a CRF $(\boldsymbol{Q}, \boldsymbol{O})$ is formulated as a conditional distribution of $\boldsymbol{Q}$ given observation $\boldsymbol{O}$ as follows [27]:

$$P(\boldsymbol{Q}|\boldsymbol{O}; \Theta) = \frac{1}{Z(\boldsymbol{O})} \exp\{F(\boldsymbol{Q}, \boldsymbol{O}; \Theta)\} \tag{2.1}$$

where $F(\boldsymbol{Q}, \boldsymbol{O}; \Theta)$ is composed by the feature functions of CRF, and $Z(\boldsymbol{O})$ is an observation-dependent normalization term, known as the partition function, with the following formulation:

$$Z(\boldsymbol{O}) = \sum_{\boldsymbol{Q'}} \exp\{F(\boldsymbol{Q'}, \boldsymbol{O}; \Theta)\} \tag{2.2}$$

For different applications, the conditional probability in Equation (2.1) can be interpreted in different ways. Considering a practical industrial example of flare gas composition inference from the flare images, the flare gas composition can be first labelled to finite number of classes, namely, $q \in \{1, 2, \cdots, P\}$, and different flare gas compositions can generate different types of flares. Due to the change of process, the compositions of flare gas could vary along with time, which can be identified from the flare image sequence. Here, $\boldsymbol{Q} = [Q_1, Q_2, \cdots, Q_N]$ is a finite sequence of composition labels, and $\boldsymbol{O}$ denotes the pixel data from the flare images. Without explicit probabilistic distribution modeling, the conditional probability of the composition label sequence given all the flare images are formulated directly as a CRF model, which can be used to predict the composition label sequence given a new set of flare images by maximizing the conditional probability $P(\boldsymbol{Q}|\boldsymbol{O})$.

As one of the solutions, a linear-chain graphic structure $\mathcal{G}(\mathcal{V}, \mathcal{E})$ shown in Fig. 2.1 can be used to describe the variable dependencies in this problem. In the linear-chain structure, the first-order Markov dependencies are assigned to the tag sequence $\boldsymbol{Q}$ which enables the adjacent connections between Q_i and Q_{i-1}.

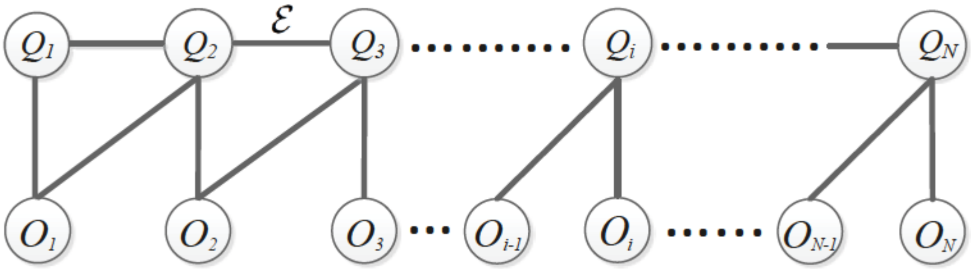


Figure 2.1: A LCCRF structure to solve the flare gas composition classification problem

The graphical structure $\mathcal{G}$ of a CRF model can be constructed from prior knowledge and is problem dependent. In this problem, it is assumed that the flare gas composition changes over time and follows the first-order Markov property, therefore it can be modelled by a linear-chain structured CRF model. After fixing the graphic structure $\mathcal{G}$ of a CRF model, a set of local functions need to be defined to quantify the correlations corresponding to each edge in $\mathcal{E}$, which are called feature functions. Specifically, in this example, three types of feature functions are defined as follows:

$$f_{k_1}(Q_i, Q_{i-1}) = \begin{cases} 1 & \text{if } Q_i = q_1 \text{ and } Q_{i-1} = q_2 \\ 0 & \text{otherwise} \end{cases} \tag{2.3}$$

$$f_{k_2}(Q_i, O_i) = \begin{cases} f(O_i) & \text{if } Q_i = q_1 \\ 0 & \text{otherwise} \end{cases} \tag{2.4}$$

$$f_{k_3}(Q_i, O_{i-1}) = \begin{cases} f(O_{i-1}) & \text{if } Q_i = q_1 \\ 0 & \text{otherwise} \end{cases} \tag{2.5}$$

where $f(\cdot)$ represents user selected functions, for example, in linear or quadratic forms.

The first feature function $f_{k_1}(Q_i, Q_{i-1})$ describes the transitions among different flare gas compositions. Since $Q_i = q \in \{1, 2, \cdots, P\}$, the total number of $f_{k_1}(Q_i, Q_{i-1})$ should be P^2 and all the possible transitions are considered. Correspondingly, each $f_{k_1}(Q_i, Q_{i-1})$ has a weighting factor θ_{k_1} to model the strength of correlation between Q_i and Q_{i-1}. Similarly, the second and third feature functions describe the correlations between the current flare gas composition Q_i and the flare image pixel data $[O_i, O_{i-1}]$. The previous observations O_{i-1} is employed to enhance the classification accuracy of current flare gas composition Q_i.

Explicitly, with the feature functions defined in Equations (2.3) - (2.5), the CRF model in Equation (2.1) can be re-formulated as

$$P(\boldsymbol{Q}|\boldsymbol{O};\Theta) = \frac{1}{Z(\boldsymbol{O})} \exp\{\sum_{i=1}^{N}[\sum_{k_1} \theta_{k_1} f_{k_1}(Q_i, Q_{i-1}) + \sum_{k_2} \theta_{k_2} f_{k_2}(Q_i, O_i) + \sum_{k_3} \theta_{k_3} f_{k_3}(Q_i, O_{i-1})]\} \tag{2.6}$$

where the normalization term $Z(\boldsymbol{O})$ has the following form:

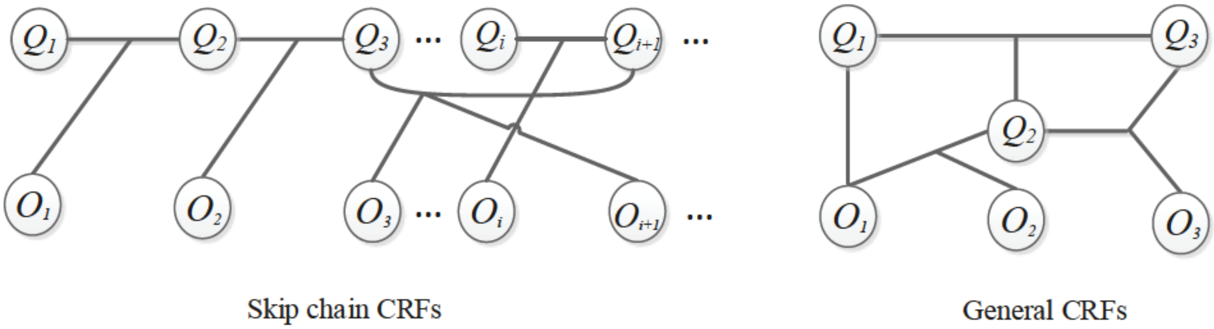
$$Z(\boldsymbol{O}) = \sum_{Q'_{1:N}} \exp\{\sum_{i=1}^{N}[\sum_{k_1} \theta_{k_1} f_{k_1}(Q'_i, Q'_{i-1}) + \sum_{k_2} \theta_{k_2} f_{k_2}(Q'_i, O_i) + \sum_{k_3} \theta_{k_3} f_{k_3}(Q'_i, O_{i-1})]\} \tag{2.7}$$

Comparing with Equations (2.1) - (2.2), the explicit formulation of function $F(\boldsymbol{Q}, \boldsymbol{O}; \Theta)$ is a linear combination of feature functions, shown as follows:

$$F(\boldsymbol{Q}, \boldsymbol{O}; \Theta) = \sum_{i=1}^{N}[\sum_{k_1} \theta_{k_1} f_{k_1}(Q_i, Q_{i-1}) + \sum_{k_2} \theta_{k_2} f_{k_2}(Q_i, O_i) + \sum_{k_3} \theta_{k_3} f_{k_3}(Q_i, O_{i-1})] \tag{2.8}$$

With such formulation, the numerator of $P(\boldsymbol{Q}|\boldsymbol{O};\Theta)$ can be factorized into a series of exponential functions. The positivity and monotonicity of the exponential functions provide higher probability to the features with higher significance to classification. Meanwhile, the exponential formulation also facilitates the log likelihood computation during the CRF training process. In this case, unlike the probabilistic generative models, one does not need to find explicit probabilistic distributions to model the evidence $\boldsymbol{O}$ given label sequence $\boldsymbol{Q}$.

Various CRFs have been proposed by extending the graphic structure $\mathcal{G}$ as shown in Fig. 2.2. Such variants have more complicated dependencies among the vertices $\mathcal{V}$ and result in increased computational loads, which is a cost paid for better modeling and expressive capability. As a result, efficient training and inference strategies need to be developed for the designed CRF models.



Skip chain CRFs

General CRFs

Figure 2.2: Two variants of CRFs [27]

2.1.2 Training of CRFs

In section 2.1.1, the basic formulations of CRFs have been introduced. In this section, by taking the LCCRF in Equation (2.6) as an example, the training procedures will be explained. For the variants of CRF models, similar idea has been employed for model training.

In CRF modeling, the conditional probability $P(Q_{1:N}|O_{1:N};\Theta)$ is formulated in Equation (2.6) when provided with fully labeled training dataset $\{Q_{1:N}, O_{1:N}\}$. The unknown weighting factors Θ need to be estimated based on the training dataset $\{Q_{1:N}, O_{1:N}\}$. The objective function is the log likelihood of the conditional probability $P(Q_{1:N}|O_{1:N};\Theta)$. By maximizing the log likelihood, the optimal estimation of Θ can be obtained, which is called MLE. On the basis of the CRF formulated in Equation (2.6), the objective function for parameter estimation is formulated as below:

$$l(\Theta) = \sum_{i=1}^{N}[\sum_{k_1}\theta_{k_1} f_{k_1}(Q_i, Q_{i-1}) + \sum_{k_2}\theta_{k_2} f_{k_2}(Q_i, O_i) + \sum_{k_3}\theta_{k_3} f_{k_3}(Q_i, O_{i-1})] - \log Z(\boldsymbol{O}) - \frac{||\Theta||_2^2}{2\sigma^2} \tag{2.9}$$

where σ is a regularization parameter in the penalty term to avoid overfitting.

In Equation (2.9), it can be observed that the first three terms of the objective function are a simple summation of weighted feature functions, and the entire computational complexity of the CRFs arises from the log normalization term $\log Z(\boldsymbol{O})$, which also makes the closed-form solution of Θ unavailable. Therefore, the numerical optimization algorithms, such as quasi-Newton algorithms, are employed to get the solution. Taking the unknown parameter θ_{k_1} as an example, the corresponding gradient is derived as

$$\begin{aligned}
\frac{\partial l(\Theta)}{\partial \theta_{k_1}} =& \sum_{i=1}^{N} \frac{\partial \sum_{k_1}\theta_{k_1} f_{k_1}(Q_i, Q_{i-1})}{\partial \theta_{k_1}} - \frac{1}{Z(\boldsymbol{O})} \cdot \frac{\partial Z(\boldsymbol{O})}{\partial \theta_{k_1}} - \frac{\theta_{k_1}}{\sigma^2} \\
=& \sum_{i=1}^{N} f_{k_1}(Q_i, Q_{i-1}) - \frac{1}{Z(\boldsymbol{O})} \sum_{Q'_{1:N}} \sum_{i=1}^{N} \{ f_{k_1}(Q'_i, Q'_{i-1}) \cdot \exp\{ \sum_{i'=1}^{N} [\sum_{k_1}\theta_{k_1} f_{k_1}(Q'_{i'}, Q'_{i'-1}) + \\
& \sum_{k_2}\theta_{k_2} f_{k_2}(Q'_{i'}, O_{i'}) + \sum_{k_3}\theta_{k_3} f_{k_3}(Q'_{i'}, O_{i'-1})]\}\} - \frac{\theta_{k_1}}{\sigma^2} \\
=& \sum_{i=1}^{N} f_{k_1}(Q_i, Q_{i-1}) - \sum_{Q'_{1:N}} \sum_{i=1}^{N} \{ f_{k_1}(Q'_i, Q'_{i-1}) \cdot P(Q'_{1:N}|O_{1:N})\} - \frac{\theta_{k_1}}{\sigma^2} \\
=& \sum_{i=1}^{N} f_{k_1}(Q_i, Q_{i-1}) - \sum_{i=1}^{N} \sum_{Q'_i, Q'_{i-1}} P(Q'_i, Q'_{i-1}|O_{1:N}) \cdot f_{k_1}(Q'_i, Q'_{i-1}) - \frac{\theta_{k_1}}{\sigma^2}
\end{aligned} \tag{2.10}$$

Similarly, the gradients of other unknown parameters are computed as follows [27]:

$$\begin{aligned}\frac{\partial l(\Theta)}{\partial \theta_{k_2}} &= \sum_{i=1}^{N} f_{k_2}(Q_i, O_i) - \sum_{i=1}^{N} \sum_{Q_i'} P(Q_i'|O_{1:N}) \cdot f_{k_2}(Q_i', O_i) - \frac{\theta_{k_2}}{\sigma^2} \\ \frac{\partial l(\Theta)}{\partial \theta_{k_3}} &= \sum_{i=1}^{N} f_{k_3}(Q_i, O_{i-1}) - \sum_{i=1}^{N} \sum_{Q_i'} P(Q_i'|O_{1:N}) \cdot f_{k_3}(Q_i', O_{i-1}) - \frac{\theta_{k_3}}{\sigma^2}\end{aligned} \tag{2.11}$$

The above gradients can be interpreted as the subtraction between the actual activated feature function values and the expectation of the activated feature function values. When the gradients are equal to zero, this means that the marginal probabilities of actual labels are equal to one with the estimated model parameters. Following this search direction, the optimal parameter estimation can be achieved. The convexity of the objective function makes the global solution of optimization achievable [27].

In the above gradient calculation, the most complicated part is to derive the marginal probabilities $P(Q_i', Q_{i-1}'|O_{1:N})$ and $P(Q_i'|O_{1:N})$. As the increase of CRF model complexity, the computations of the marginal probabilities get harder. As a result, efficient inference strategies of CRF models need to be further explored.

2.1.3 Inference of CRFs

In CRFs, two common inference problems should be considered, namely, the marginal probability calculation in the training process and the most likely label sequence estimation when given new observations [27]. The optimal solutions of both inference problems need to be searched from an exponential number of possible combinations. Therefore, efficient inference algorithms are proposed to find out the solutions. Still based on the LCCRF defined in Equation (2.6), a forward-backward propagation strategy can be employed to obtain the exact solutions of the inference problems.

The calculation of the marginal probability $P(Q_i, Q_{i-1}|O_{1:N})$ is conducted from the following formulation:

$$P(Q_i, Q_{i-1}|O_{1:N}) = \frac{1}{\sum_{Q_{1:N}} \exp F(Q_{1:N}, O_{1:N}; \Theta)} \cdot \{ \sum_{Q_{1:i-2}} \exp F(Q_{1:i-1}, O_{1:N}; \Theta) \} \cdot \\ \exp F(Q_{i-1:i}, O_{1:N}; \Theta) \cdot \{ \sum_{Q_{i+1:N}} \exp F(Q_{i:N}, O_{1:N}; \Theta) \} \tag{2.12}$$

where the enumeration of all the possible label sequences $Q_{1:N}$ is integrated, which can be solved by forward and backward propagations.

Take the LCCRF defined in Equation (2.6) for example, the forward propagation can be separated into three procedures.

1. **Initialization**

   For initialization, an initial term is created as $\alpha_1(Q_1) = \exp\{\sum_{k_2} \theta_{k_2} f_{k_2}(Q_1, O_1)\}$

2. **Propagation**

   For propagation, the linear-chain structure enables the updated feature functions included into the intermediate forward variable $\alpha^{(i)}(Q_i, Q_{i-1})$ as the chain length increasing from $i = 1$ to $i = 2$ as below:

$$\alpha^{(2)}(Q_2, Q_1) = \alpha_1(Q_1) \cdot \exp\{\sum_{k_1} \theta_{k_1} f_{k_1}(Q_2, Q_1) + \sum_{k_2} \theta_{k_2} f_{k_2}(Q_2, O_2) + \sum_{k_3} \theta_{k_3} f_{k_3}(Q_2, O_1)\} \tag{2.13}$$

   which can be generalized to the cases $i = 3, \cdots, N$, as follows:

$$\alpha^{(i)}(Q_i, Q_{i-1}) = \alpha_{i-1}(Q_{i-1}) \cdot \exp\{\sum_{k_1} \theta_{k_1} f_{k_1}(Q_i, Q_{i-1}) + \sum_{k_2} \theta_{k_2} f_{k_2}(Q_i, O_i) + \sum_{k_3} \theta_{k_3} \\ \cdot f_{k_3}(Q_i, O_{i-1})\} \tag{2.14}$$

   where the variable $\alpha_{i-1}(Q_{i-1})$ is derived from the previous integration result.

3. **Integration**

For integration, the intermediate forward variable $\alpha^{(2)}(Q_2, Q_1)$ will be marginalized as below:

$$\alpha_2(Q_2) = \sum_{Q_1} \alpha^{(2)}(Q_2, Q_1) \tag{2.15}$$

which can also be generalized to the cases with $i = 3, \cdots, N$ as

$$\alpha_i(Q_i) = \sum_{Q_{i-1}} \alpha^{(i)}(Q_i, Q_{i-1}) \tag{2.16}$$

The above propagation and integration procedures will be performed iteratively with i alongside the linear chain.

The forward propagation is able to generate a sequence of forward variables $\{\alpha_i(Q_i)\}_{i=1}^{N}$ with the following formulation:

$$\alpha_i(Q_i) = \sum_{Q_{1:i-1}} \exp\{\sum_{i'=1}^{i}[\sum_{k_1} \theta_{k_1} f_{k_1}(Q_{i'}, Q_{i'-1}) + \sum_{k_2} \theta_{k_2} f_{k_2}(Q_{i'}, O_{i'}) + \sum_{k_3} \theta_{k_3} f_{k_3}(Q_{i'}, O_{i'-1})]\} \tag{2.17}$$

which can be used to compute the denominator and the first summation of the numerator exponential term in Equation (2.12).

The backward propagation has similar procedures to the forward propagation, but with a reversed direction. The backward propagation procedures have been summarized as below.

1. **Initialization**

   The backward propagation initialization starts from the end of the sequence, formulated as $\beta_N(Q_N) = 1$.

2. **Propagation**

   Starting from the end of the sequence, the feature functions are propagated backwards to the beginning of the sequence. With $i = N - 1$ to $i = 1$, the backward intermediate variables $\beta^{(i)}(Q_{i+1}, Q_i)$ are computed as

$$\beta^{(i)}(Q_{i+1}, Q_i) = \beta_{i+1}(Q_{i+1}) \cdot \exp\{\sum_{k_1} \theta_{k_1} f_{k_1}(Q_{i+1}, Q_i) + \sum_{k_2} \theta_{k_2} f_{k_2}(Q_{i+1}, O_{i+1}) + \sum_{k_3} \theta_{k_3} f_{k_3}(Q_{i+1}, O_i)\} \tag{2.18}$$

3. **Integration**

   After information propagation, the intermediate backward variable $\beta^{(i)}(Q_{i+1}, Q_i)$ will be marginalized as below:

$$\beta_i(Q_i) = \sum_{Q_{i+1}} \beta^{(i)}(Q_{i+1}, Q_i) \tag{2.19}$$

Similar to the forward propagation, for $i = N, N-1, \cdots, 1$, the backward propagation enables a marginal sequence as

$$\beta_i(Q_i) = \sum_{Q_{i+1:N}} \exp\{\sum_{i'=i}^{N}[\sum_{k_1} \theta_{k_1} f_{k_1}(Q_{i'}, Q_{i'-1}) + \sum_{k_2} \theta_{k_2} f_{k_2}(Q_{i'}, O_{i'}) + \sum_{k_3} \theta_{k_3} f_{k_3}(Q_{i'}, O_{i'-1})]\} \tag{2.20}$$

which can be used to compute the third exponential summation term of the numerator in Equation (2.12).

As a result, the first inference problem of CRFs can be solved by making use of the forward-backward propagation results as below:

$$\begin{aligned} P(Q_i, Q_{i-1}|O_{1:N}) &= \frac{\alpha_{i-1}(Q_{i-1}) \cdot \exp F(Q_{i-1:i}, O_{1:N}; \Theta) \cdot \beta_i(Q_i)}{\sum_{Q_N} \alpha_N(Q_N)} \\ P(Q_i|O_{1:N}) &= \frac{\alpha_i(Q_i) \cdot \beta_i(Q_i)}{\sum_{Q_N} \alpha_N(Q_N)} \end{aligned} \tag{2.21}$$

The second inference problem of CRFs can be represented to estimate the most likely label sequence, namely, $\boldsymbol{Q}^* = \text{argmax}_{\boldsymbol{Q}} P(\boldsymbol{Q}|\boldsymbol{O}_{new}; \Theta)$. To solve this problem, Viterbi decoding algorithm will be employed by following similar patterns to the backward propagation procedures [27].

In summary, the most fundamental exact inference algorithms of the LCCRFs have been formulated in this section as the preliminary of this thesis. The extended CRF based approaches proposed in the subsequent chapters will need their specifically designed inference algorithms to get the solutions, which are based on the basic forward-backward algorithms.

2.2 Expectation Maximization (EM) and Variational Bayesian (VB) Algorithms

The basic problem of MLE is to search for the optimal solution of the unknown parameters in a probabilistic model, which can maximize the likelihood of the observations. When the variables in the established probabilistic model are all known, the standard MLE algorithms work well. However, when there exist latent variables, such as hidden operating modes, unknown distribution of the parameters or incomplete observations, the application of the standard MLE algorithms turns out to be difficult. In this situation, the EM and VB algorithms are proposed to efficiently solve the MLE problem in an iterative way by involving the latent variables into the solution with reduced computational load. This section briefly introduces the mathematical background of the EM and VB algorithms, and then discusses the consistency and difference between the EM and VB approaches.

2.2.1 EM Algorithm

The EM algorithm is composed of two steps, the expectation step (E-step) and maximization step (M-step), which are performed iteratively until convergence. In the E-step, the posterior distribution of latent variables are estimated by the posterior probabilities calculated with the observed variables and the current estimate of model parameters. In the M-step, with the estimated posterior distribution of latent variables in the E-step, the model parameters are updated to maximize the likelihood function [47].

Assume a complete dataset D_c consists of the observed dataset D_o and the latent dataset D_m, i.e., $D_c = \{D_o, D_m\}$. The probabilistic distribution of D_c is parameterized by an unknown parameter set Θ. An optimal estimation of Θ that maximizes $P(D_o|\Theta)$ is known as an MLE solution. However, in presence of the latent variables D_m, instead of directly maximizing $P(D_o|\Theta)$, the lower bound of the actual log likelihood, known as Q-function, is maximized in the following E-step and M-step [47].

- *E-step: Calculate the posterior probability with respect to the latent variables and formulate the Q-function*

  $$Q(\Theta|\Theta^{(k)}) = E_{p(D_m|D_o;\Theta^{(k)})} \log\{P(D_o, D_m|\Theta)\} \tag{2.22}$$

  where k indicates the current iteration.

- *M-step: Find $\Theta^{(k+1)}$ as any value of $\Theta \in \Omega$ that maximizes the Q-function*

  $$Q(\Theta^{(k+1)}|\Theta^{(k)}) \geq Q(\Theta|\Theta^{(k)}) \tag{2.23}$$

  where Ω is the solution space of Θ.

For the EM algorithm, it is critical to prove that after each EM iteration, the likelihood function of the observed dataset, i.e., $\log P(D_o|\Theta)$, does not decrease, which has been proven in [48] and will be explained briefly as follows.

First, according to Bayes rule, the following relationship between D_o and D_m holds:

$$p(D_c|D_o;\Theta) = \frac{p(D_c;\Theta)}{p(D_o;\Theta)} \tag{2.24}$$

where $p(\cdot)$ represents the probability density function of the target dataset.

Then the log likelihood of the observed dataset can be represented by

$$\log p(D_o;\Theta) = \log p(D_c;\Theta) - \log p(D_c|D_o;\Theta) \tag{2.25}$$

Taking expectations of both sides in the above equation with respect to the conditional probability of D_m given D_o parameterized with $\Theta^{(k)}$, one can get

$$\begin{aligned} E_{p(D_m|D_o;\Theta^{(k)})}\{\log p(D_o;\Theta)\} &= E_{p(D_m|D_o;\Theta^{(k)})}\{\log p(D_c;\Theta)\} - E_{p(D_m|D_o;\Theta^{(k)})}\{\log p(D_c|D_o;\Theta)\} \\ &= Q(\Theta|\Theta^{(k)}) - H(\Theta|\Theta^{(k)}) \end{aligned} \tag{2.26}$$

Since the observed dataset D_o has no correlation with D_m, the above equation can be simplified as

$$\log p(D_o;\Theta) = Q(\Theta|\Theta^{(k)}) - H(\Theta|\Theta^{(k)}) \tag{2.27}$$

Therefore, one can have that

$$\begin{aligned} \log p(D_o;\Theta^{(k+1)}) - \log p(D_o;\Theta^{(k)}) = \{&Q(\Theta^{(k+1)}|\Theta^{(k)}) - Q(\Theta^{(k)}|\Theta^{(k)})\} - \\ &\{H(\Theta^{(k+1)}|\Theta^{(k)}) - H(\Theta^{(k)}|\Theta^{(k)})\} \end{aligned} \tag{2.28}$$

The difference between $H(\Theta^{(k+1)}|\Theta^{(k)})$ and $H(\Theta^{(k)}|\Theta^{(k)})$ is calculated as

$$\begin{aligned} H(\Theta^{(k+1)}|\Theta^{(k)}) - H(\Theta^{(k)}|\Theta^{(k)}) &= E_{p(D_m|D_o;\Theta^{(k)})}\{\log p(D_c|D_o;\Theta^{(k+1)}) - \log p(D_c|D_o;\Theta^{(k)})\} \\ &= E_{p(D_m|D_o;\Theta^{(k)})}\{\log \frac{p(D_c|D_o;\Theta^{(k+1)})}{p(D_c|D_o;\Theta^{(k)})}\} \end{aligned} \tag{2.29}$$

Applying Jensen's inequality, the above equation turns out to be

$$\begin{aligned} H(\Theta^{(k+1)}|\Theta^{(k)}) - H(\Theta^{(k)}|\Theta^{(k)}) &\le \log E_{p(D_m|D_o;\Theta^{(k)})}[\frac{p(D_c|D_o;\Theta^{(k+1)})}{p(D_c|D_o;\Theta^{(k)})}] \\ &= \log \int_{D_m} p(D_m|D_o;\Theta^{(k)}) \cdot \frac{p(D_m|D_o;\Theta^{(k+1)})}{p(D_m|D_o;\Theta^{(k)})} \; dD_m \\ &= \log \int_{D_m} p(D_m|D_o;\Theta^{(k+1)}) \; dD_m \\ &= 0 \end{aligned} \tag{2.30}$$

Therefore, the second difference term in Equation (2.28) is proven to be non-positive. Together with the first nonnegative difference term derived from M-step, one can conclude $\log p(D_o;\Theta^{(k+1)}) \ge \log p(D_o;\Theta^{(k)})$ after each EM iteration.

2.2.2 VB Algorithm

As an alternative to solve the MLE problem with latent variables, the VB algorithm is proposed with more flexible formulation than the EM algorithm. In the following content, the VB inference and VB-EM algorithms will be briefly reviewed.

VB Inference

Still considering a complete dataset $D_c = \{D_o, D_m\}$ following a particular distribution parameterized by Θ, one might be interested in the posterior probability $p(D_m|D_o;\Theta)$ which may have no tractable solutions. Therefore, a predefined probability distribution $q(D_m)$ is determined to approximate the actual posterior. To measure the similarity between $p(D_m|D_o)$ and $q(D_m)$, the nonnegative Kullback-Leibler (KL) divergence is proposed as follows [49]:

$$D_{KL}(q(D_m)||p(D_m|D_o)) = -\int_{D_m} q(D_m) \log \frac{p(D_m|D_o)}{q(D_m)} \; dD_m \tag{2.31}$$

By using the Bayes rule $p(D_m|D_o) = p(D_m, D_o)/p(D_o)$ in Equation (2.31), the KL divergence can be further expanded as below:

$$\begin{aligned} D_{KL}(q(D_m)||p(D_m|D_o)) &= -\int_{D_m} q(D_m)\log\frac{p(D_m, D_o)}{q(D_m)p(D_o)}\ dD_m \\ &= -\int_{D_m} q(D_m)\log\frac{p(D_m, D_o)}{q(D_m)}\ dD_m + \log p(D_o) \end{aligned} \tag{2.32}$$

where the first integral term is known as variational lower bound, which can be denoted as $L(q(D_m))$.

As a result, Equation (2.32) can be rearranged as

$$\log p(D_o) = D_{KL}(q(D_m)||p(D_m|D_o)) + L(q(D_m)) \tag{2.33}$$

The left-hand side of Equation (2.33) is the log likelihood of the observed dataset, which is independent of $q(D_m)$, so the summation of the KL divergence and $L(q(D_m))$ can be treated as a constant with respect to $q(D_m)$. Since the KL divergence is nonnegative, by minimizing the KL divergence, the lower bound $L(q(D_m))$ can be maximized. When the approximated $q(D_m)$ equals to the actual posterior, the lower bound is equal to the log likelihood of the observations. The objective of VB inference is to compute the approximated $q(D_m)$ by either minimizing the KL divergence or maximizing the lower bound, which is equivalent to solving an optimization problem.

VB-EM Algorithm

When taking the model parameter Θ into consideration, the VB inference combined with EM idea becomes a solution to parameter estimation. Under the VB-EM framework, there are also two steps performed iteratively [49].

- *VB-E step: Calculate the approximate posterior $q(D_m)$ by maximizing the lower bound with fixed parameters*

$$\hat{q}(D_m) = \underset{q(D_m)}{\operatorname{argmax}}\, L(q(D_m), \Theta^{(k)}) \tag{2.34}$$

where k indicates the current iteration.

- *VB-M step: Find out Θ by maximizing the lower bound with fixed $q(D_m)$ derived from E-step*

$$\Theta^{(k+1)} = \underset{\Theta}{\operatorname{argmax}}\, L(\hat{q}(D_m), \Theta) \tag{2.35}$$

In summary, the VB-EM framework can be treated as a combination of the VB inference and the standard EM algorithm. The VB-E step tries to make the lower bound as close to $\log p(D_o)$ as possible, and the VB-M step tries to maximize the lower bound, therefore maximizing the target log likelihood $\log p(D_o)$.

2.2.3 The Comparison between EM and VB Algorithms

As reviewed in the above two subsections, the main difference between the standard EM and the VB-EM algorithms lies in the VB inference in the VB-E step. Instead of approximating $q(D_m)$, the exact posterior $p(D_m|D_o)$ is calculated in the E-step of the standard EM algorithm, where the KL divergence naturally becomes zero. With the exact posterior $p(D_m|D_o)$, the lower bound in the VB-M step turns out to be

$$\begin{aligned} L(\Theta) &= \int_{D_m} q(D_m) \log \frac{p(D_m, D_o; \Theta)}{q(D_m)}\; dD_m \\ &= \int_{D_m} p(D_m|D_o; \Theta^{(k)}) \log \frac{p(D_m, D_o; \Theta)}{p(D_m|D_o; \Theta^{(k)})}\; dD_m \\ &= \int_{D_m} p(D_m|D_o; \Theta^{(k)}) \log p(D_m, D_o; \Theta)\; dD_m - \int_{D_m} p(D_m|D_o; \Theta^{(k)}) \\ &\quad \cdot \log p(D_m|D_o; \Theta^{(k)})\; dD_m \\ &= Q(\Theta|\Theta^{(k)}) - C_\Theta \end{aligned} \tag{2.36}$$

where C_Θ represents that the second term can be treated as a constant with respect to the unknown parameter Θ.

As a result, the standard EM algorithm is a special case of the VB-EM algorithm. When the accuracy of $p(D_m|D_o)$ is higher than the approximate $q(D_m)$, the standard EM algorithm might achieve better performance than the VB-EM algorithm. On the other hand, owing to the VB inference procedure, the VB-EM algorithm provides more flexible solutions than the standard EM algorithm. The following two situations are raised as examples:

1. When dealing with certain complicated distributed D_m, the exact posterior $p(D_m|D_o)$ calculation is intractable. Then the mean-field approximation [50] can be employed to derive $q(D_m)$.

2. In more general cases, D_m can not only represent the latent variables, but also the unknown model parameters. Then the posterior distribution of the model parameters $q(\Theta)$ can be included and estimated as solutions, rather than a point estimation solution in the standard EM algorithm.

In summary, both EM and VB algorithms have the capability to efficiently deal with unknown parameter estimation problem with incomplete dataset, which cannot be achieved by the standard MLE approach. Compared with EM algorithm, VB algorithm has more flexibility of modeling the posteriors of latent variables and therefore it is able to deal with parameter estimation of more complicated models than the EM algorithm. Consequently, the computational complexity of VB algorithm turns out to be higher than the EM algorithm. On the other hand, the EM algorithm has a simpler formulation and tends to provide solutions with reduced computational loads than the VB algorithm. However, when the posteriors of latent variables are too complicated to be derived, the EM algorithm will lose its effectiveness.

Chapter 3

Hierarchically Distributed Monitoring

3.1 Introduction

Gas flaring is the controlled burning of waste gases that cannot be processed for sale or further use due to various technical and logistical reasons [51]. Gas flaring also improves process safety, because it protects vessels and pipelines from over-pressuring due to unplanned upsets, thereby, avoiding accidental explosions [52]. However, gas flaring contributes to pollution, and it is also an important source of greenhouse gas emissions [53]. Moreover, burning of the flare gases results in waste of energy that can potentially be reused in industrial processes.

In order to reduce the undesired environmental and economic impacts of flaring, many solutions have been developed, which include timely maintenance of flare systems, modifying start-up and shut-down procedures, etc., and installation of new equipment to recover the waste gases. Such a recovery process is known as a flare gas recovery system (FGRS) that captures flare gases for reuse in the plant or for sale [54, 55]. However, the amount of flare gases can sometimes exceed the capacity of the FGRS and eventually the excess waste gases need to be burnt in the flare stack, resulting in a flare event. Such flare events are undesirable due to their harmful impacts on the environment and the economical losses

[1] Part of this chapter has been published as Mengqi Fang, Fadi Ibrahim, Hariprasad Kodamana, Biao Huang, Noel Bell, and Mark Nixon. Hierarchically distributed monitoring for the early prediction of gas flare events. *Industrial & Engineering Chemistry Research*, 58(26):11352–11363, 2019.

during plant operations. Hence, the flare event predictions are useful as the operators can proactively intervene and reduce the chance of a flare event.

To the best of the authors' knowledge, currently, there are no monitoring strategies reported in the literature that provide early warning of a potential flare event. The case studied in this chapter is based on real refinery process data. The major impediment in the current case is a very limited knowledge of the underlying process. Thus it is a typical case suitable for application of data analytics. While historical process data are usually available, they are highly autocorrelated, high dimensional, and contain outliers and missing data. Preliminary study proposed by Noel and Mark [56] as well as the traditional centralized PCA approach have been attempted. However, both approaches can only predict a small fraction of the flare events. Therefore, in this chapter, we propose a systematic methodology for early flare prediction by making full use of the available process data. The by-product of this study is to create a tool that can help industries to better predict thus reduce the flare events.

Benefiting from the large body of MSPM strategies proposed in literature, there are various solutions available for fault detection and diagnosis of complicated industrial processes, for instance, PCA [57], PLS [58], and dynamic PCA [59], among others. As the process complexity increasing, distributed monitoring strategies emerged and attracted wide attentions from researches. In contrast to centralized monitoring, distributed monitoring is conducted by dividing a large-scale process into several sub-blocks and then monitoring the variations in each sub-block and further the entire process. Several multiblock methods have been developed [60, 61] and employed as part of distributed monitoring strategies [62, 63, 64, 65]. From a probabilistic perspective, a unified probabilistic framework for process monitoring has also been proposed [66]. Recently, SFA was proposed for dynamic process monitoring by separating the temporal correlations from the steady-state process information [67]. Further, frequency domain analysis methods such as the fast Fourier transform (FFT) [68] and the wavelet transform (WT) [69] have also been employed as alternative ways of feature extraction that can be used for MSPM [70]. In this work, motivated by the success of multiblock process monitoring approaches, such as hierarchical PCA [71], we propose a distributed and hierarchical monitoring framework for real-time early warnings of potential flare events by analyzing real-time plant data.

The contributions of this chapter can be summarized from both practical and theoretical perspectives, as follows: (i) The dataset under research is from a real refinery process, and this chapter is the first which focuses on and successfully solved the early flare event prediction problem in refinery by performing the plant-wide process monitoring. The problem is challenging. Multiple data analytic strategies have been attempted but the unique signatures of early flare events are very difficult to extract. To this end, the hierarchical structure designed in this work achieves the optimal prediction performance, which will be beneficial to related industries for a more environmental friendly operation. (ii) There are very few existing works on the distributed SFA and frequency-domain approaches. This chapter proposes the use of SFA for distributed process monitoring. Both frequency and time-domain methods are analyzed and compared in this work under the distributed framework.

The remainder of this chapter is organized as follows: In the process description section, the refinery process and FGRS are reviewed. In the problem statement section, details of the refinery dataset used in this work, challenges of this problem, and two preliminary studies are explained. In the next section, the proposed hierarchically distributed monitoring framework is presented, including both time and frequency domain techniques. Further, we apply the proposed approach to the refinery dataset used in this work. Finally, conclusions are drawn in the last section.

3.2 Process Description

A refinery process is composed of several units designed for crude oil processing, such as a crude desalter, heat exchangers, reaction related units, and separation units. When one or more process units undergo upsets due to abnormal events, such as power blips, equipment failures, or crude composition changes, the resulting excess gases are directed to a FGRS. The flare events will occur when FGRS is not able to recover the excess gases. Fig. 3.1 illustrates a typical FGRS, which is essentially composed of a flare header, flare gas recovery components, and a flare stack. During low volume flare periods, the flare gas flowrate does not exceed the FGRS capacity, and the waste gas recovery process, which includes the flare gas compression and separation, is activated. Finally, the compressed waste gases are recycled, usually as fuel gas. However, during abnormal operational scenarios, the flare gas flowrate in the flare header might exceed the processing capacity of the FGRS, and consequently the excess flare gases will pass through the liquid seal to the flare stack, where the gases will be burnt. Once the flare gas flowrate increases rapidly and passes the liquid seal, the pressure in the flare stack is also increased. The ratio of pressure and liquid level in the flare stack, called flare ratio, is an indicator of a flare event, once it surpasses a pre-determined tolerant value. Flare ratio is defined as:

$$Flare\ ratio = \frac{P_l}{L_l} \tag{3.1}$$

where P_l and L_l represent the pressure and liquid level in unit 1 of the refinery dataset used in this work, respectively. Generally, the operational range of the flare ratio during flare events is around $0.85 - 0.97$.

The work presented here attempts to predict these flare events at an early stage by identifying the signatures that are latent in the routine operational data. Specifically, an early prediction can be defined as at least 15 minute early warning before the actual flare event occurrence.

3.3 Problem Statement

In this section, the refinery process under consideration and the challenges involved in solving this problem are presented. The available dataset is for a one year period and contains data

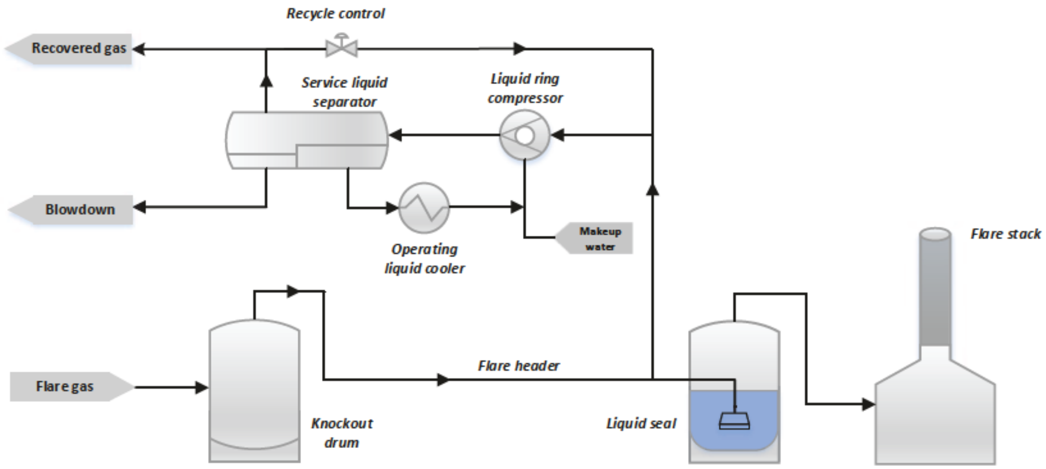


Figure 3.1: A general schematic of the FGRS [55]

from the refinery's units that are connected to the FGRS. Salient properties of the dataset are summarized in Table 3.1 and 3.2, where tag names and unit identities of PVs are known, but no further process knowledge is provided.

Table 3.1: Details regarding the refinery process dataset

Number of PVs					Number of available units	Sampling time interval	Number of samples
Pressure	Flowrate	Temperature	Level	Others			
113	9	4	1	5	18	1 minute	502498

Table 3.2: Available refinery units' names, identities and their related PV numbers

Unit number	Unit description	PV number	Unit number	Unit description	PV number
1	Flare system # 1	4	10	Aromatics unit # 1	10
2	Flare gas recovery unit	13	11	Aromatics unit # 2	6
3	Flare system # 2	1	12	Alkylation unit # 1	6
4	Hydrocracking unit	12	13	Alkylation unit # 2	2
5	Saturated gas unit # 1	10	14	Amine regen unit	21
6	Saturated gas unit # 2	10	15	Naphtha unit	2
7	Crude unit # 1	9	16	Utilities	4
8	Crude unit # 2	10	17	SRU # 1	4
9	Hydrogen desulfurization	4	18	SRU # 2	4

There were 14 flare events in total throughout the one year period, during which the flare ratio increased above 0.85. The durations of the flare events varied widely as shown in Table

3.3.

Table 3.3: The durations of all the 14 flare events

Flare event number	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Duration (min)	93	128	4	52	2	1	79	256	35	5	8	6	8	114

Moreover, within the one year investigation period, the entire process operation status is time-varying and shows non-stationary characteristics. As an example, a tag of unit 14 is shown in Fig. 3.2. The operating conditions change over time and changes are not consistent from unit to unit. In the absence of process knowledge, such changes are difficult to capture based on experience. Because the available dataset contains only 14 flare events, the abnormal data are too few to conduct supervised learning. Additionally, the abnormal patterns of the 14 flare events vary from one to the other, because they are not all caused by the same process upset.

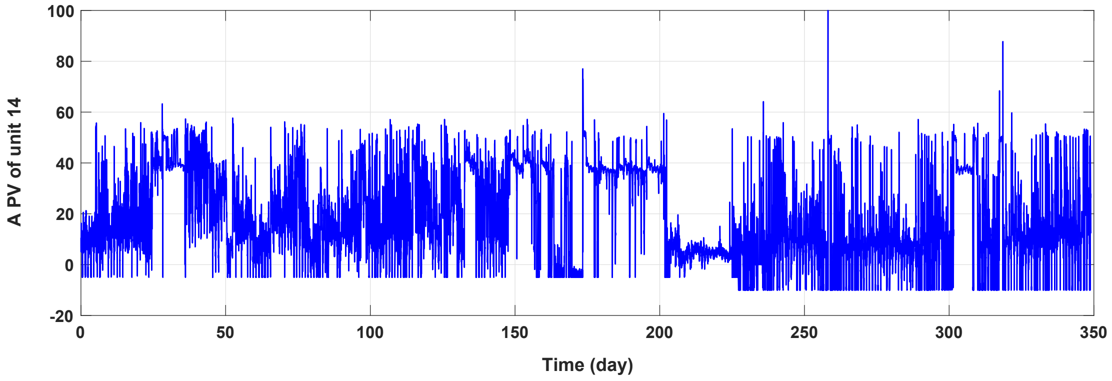


Figure 3.2: An illustration of one pressure tag in unit 14 showing process variability within one year investigation period

The challenges in solving this problem are summarized as follows: (1) limited process knowledge; (2) high dimensionality of the dataset with 132 tags from 18 different units; (3) non-stationarity and time-varying process; (4) various patterns of different PVs from different units; (5) high correlations among PVs; (6) small number of flare event samples; (7) non-universal signatures associated with the flare occurrences.

A preliminary method proposed by Noel and Mark [56] has been attempted, wherein four available flowrates of units 3, 4 and 8 are employed in a moving window FFT framework to extract features, which is further utilized to develop a PCA model for early flare monitoring. Even though the selected PVs are highly informative, only 5 out of 14 flare events get early predictions. The traditional centralized PCA method has also been attempted using all the available PVs, with the prediction results and false rate shown in Table 3.4. The false rate here is calculated as the percentage of false alarms during the entire year, as follows:

$$False\ rate = \frac{Total\ number\ of\ false\ alarms}{Total\ number\ of\ sampling\ points} \tag{3.2}$$

Table 3.4: Early flare event prediction performance of traditional PCA approach on all the PVs

	T^2 statistic	
	Potential predictions	False rate
Traditional PCA approach	5/14	4.10%

We can see that the above early attempts predicted less than half of the total flares that occurred. To improve the flare prediction performance, a hierarchically distributed monitoring framework with an online adaptive strategy is proposed in this work and will be presented in the next section.

3.4 Hierarchically Distributed Monitoring Approaches for Early Flare Event Prediction

In this section, we present a systematic monitoring methodology based on the available PV dataset for the early flare event predictions when there is limited knowledge of the underlying process. Two realistic assumptions are made before proceeding to the data analysis: (1) The PVs within the same unit have higher correlations with each other compared to the PVs from other units; (2) One or more units may contribute to the flare events. On the basis of the above assumptions, all the available PVs are grouped based on their corresponding units and multivariate feature extraction methods are applied to each group. The rationale behind grouping PVs by unit is that process changes or drifts in a particular unit would affect the PVs in that unit in a more consistent fashion. This allows us to extract the most meaningful information from every unit. Then, integrating this unit-wise representative information by employing a hierarchical layer can better capture the signatures latent in the whole plant in contrast to the case when we treat the PVs in the whole plant as one group.

The schematic of the proposed strategy is presented in Fig. 3.3, wherein two monitoring layers are constructed hierarchically at unit-wise and plant-wide levels, respectively. A few dominant features are selected from PV data to represent the characteristics of a unit in the lower layer, which are then passed on to the upper process monitoring layer to synthesize monitoring statistics.

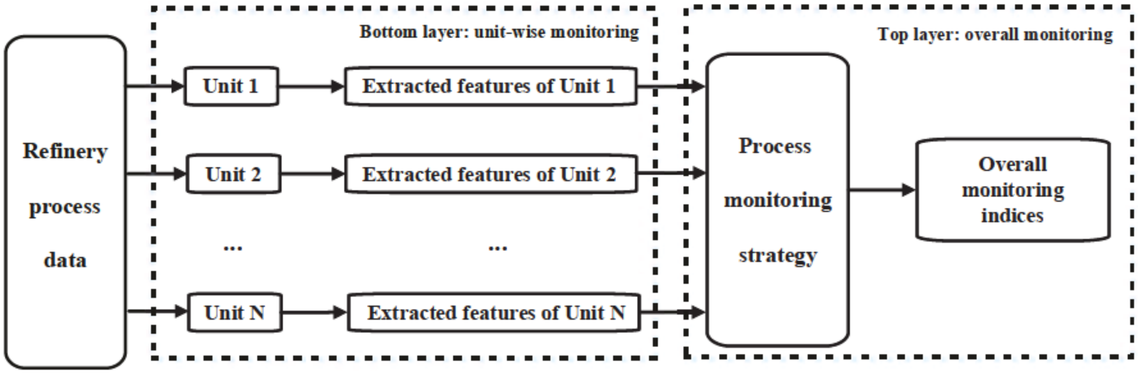

Bottom layer: unit-wise monitoring
Top layer: overall monitoring
Refinery process data
Unit 1
Unit 2
...
Unit N
Extracted features of Unit 1
Extracted features of Unit 2
...
Extracted features of Unit N
Process monitoring strategy
Overall monitoring indices

Figure 3.3: Schematic diagram of the proposed hierarchically distributed process monitoring framework for early flare event prediction

It is expected that process abnormalities are reflected as changes in the PV behaviours either in their time series evolution and/or in their frequency evolution. Therefore, the proposed hierarchically distributed monitoring strategy is applied in both time-domain and frequency-domain to fully capture the hidden signatures. Two hierarchical monitoring approaches are proposed in this work:

(1) The hierarchical time-domain approach, using PCA for unit-wise feature extraction and SFA for overall process monitoring;

(2) The hierarchical frequency-domain approach, using WT for unit-wise feature extraction and PCA for overall process monitoring.

In the proposed hierarchical time-domain approach, PCA is first employed at the unit-wise level and then SFA is used in the plant-wide level. Such a configuration is capable of extracting both slower and faster changing features of PVs in an individual unit using PCA, which are then fed to the SFA-based hierarchical layer for overall process monitoring. In the proposed frequency-domain method, the WT approach is first employed in the unit-wise layer to extract frequency features from raw PVs, subsequent to that PCA is employed for hierarchical monitoring. Since in the frequency domain approach, the frequency scales selection have been conducted in the unit-wise level, the slower and faster changing variables do not need to be separately monitored again. Therefore, in the hierarchical frequency domain approach, PCA is sufficient for overall monitoring.

3.4.1 The Time-domain Hierarchical Monitoring Approach

In this section, the hierarchical PCA-SFA approach and online model update strategy will be illustrated and explained.

3.4.1.1 PCA Process Monitoring Approach

Suppose that the PVs are grouped according to the unit as $\{\boldsymbol{X}_1, \cdots, \boldsymbol{X}_n, \cdots, \boldsymbol{X}_N\}$, where N represents the total number of the analyzed units. Given a PV group $\boldsymbol{X}_n \in \Re^{T\times m_n}$ with T data samples and m_n PVs, the following PCA model is developed by decomposing each $\boldsymbol{X}_n$ matrix into a principal component (PC) space and a residual space [72]:

$$\boldsymbol{X}_n = \boldsymbol{S}_n\boldsymbol{P}_n^T + \boldsymbol{E}_n \tag{3.3}$$

where $\boldsymbol{S}_n \in \Re^{T \times k_n}$ is the score matrix, $\boldsymbol{P}_n \in \Re^{m_n \times k_n}$ is the loading matrix with k_n PCs, and $\boldsymbol{E}_n$ denotes the residual matrix.

For the purpose of process monitoring, when a new sample $\boldsymbol{x}_n$ of the n^{th} unit's PV group appears, it is projected on the PC space to predict the scores online as follows:

$$\boldsymbol{s}_n = \boldsymbol{P}_n^T \boldsymbol{x}_n \tag{3.4}$$

where $\boldsymbol{s}_n \in \Re^{k_n}$ indicates the extracted PCs from $\boldsymbol{x}_n$. Furthermore, by means of the established PCA model, two classical monitoring statistics, namely, Hotelling's T^2 and Q statistic are constructed to monitor the PC and residual spaces respectively, as below [59]:

$$\begin{aligned} T_n^2 &= \boldsymbol{s}_n^T \boldsymbol{\Lambda}_n^{-1} \boldsymbol{s}_n \\ Q &= \| (\boldsymbol{I} - \boldsymbol{P}_n \boldsymbol{P}_n^T) \boldsymbol{x}_n \|^2 \end{aligned} \tag{3.5}$$

where the diagonal matrix $\boldsymbol{\Lambda}_n$ is composed by the first k_n leading eigenvalues of the sample covariance matrix $\dfrac{1}{T-1} \boldsymbol{X}_n^T \boldsymbol{X}_n$.

Further, by integrating all the selected PCs from units, an integrated PC matrix is formulated as $\boldsymbol{Y}_{PC} = [\boldsymbol{S}_1, \cdots, \boldsymbol{S}_N] \in \Re^{T \times K}$, which will be analyzed and monitored by the hierarchical layer.

3.4.1.2 The Hierarchical PCA-SFA Monitoring Approach

The extracted latent features of SFA reflect the slowness of the process and therefore are referred to as slow features which are sorted from the slowest to the fastest. In real processes, some abnormal situations are reflected in the slower changing PVs and some are reflected in the faster ones. Therefore, for better monitoring performance, the slower and faster features are monitored separately in the hierarchical layer. Here, the PCA method is employed first to ensure that the extracted features of each unit could preserve the information reflecting both slower and faster process changes, so that the slower and faster process changes can be monitored by SFA [73] at the hierarchical layer for a complete analysis. Otherwise, some of the process change information related to potential flare events might get lost in the unit-wise analysis.

Given the integrated PC matrix $\boldsymbol{Y}_{PC}$, the slow features can be obtained by the following mapping relation:

$$\boldsymbol{V} = \boldsymbol{Y}_{PC}\boldsymbol{W}^T \tag{3.6}$$

where $\boldsymbol{V} = [\boldsymbol{v}^1, \cdots, \boldsymbol{v}^K] \in \Re^{T\times K}$ is the slow feature matrix, and $\boldsymbol{W} = [\boldsymbol{w}_1, \cdots, \boldsymbol{w}_K]^T$ denotes the coefficient matrix obtained by conducting the following decomposition [67]:

$$\boldsymbol{A}\boldsymbol{W} = \boldsymbol{B}\boldsymbol{W}\boldsymbol{\Omega} \tag{3.7}$$

In Equation (3.7), $\boldsymbol{A}$ and $\boldsymbol{B}$ denote the covariance matrices of the first order derivative of $\boldsymbol{Y}_{PC}$ and raw input $\boldsymbol{Y}_{PC}$, respectively, and $\boldsymbol{\Omega} = diag\{\omega_1, \cdots, \omega_K\}$ is a diagonal matrix composed of the slowness of individual slow features, where $\omega_{1:K}$ are sorted in an increasing order. Such a decomposition ensures that the individual slow feature has zero mean and unit variance, while being uncorrelated with each other.

For process monitoring, based on the slowness $\omega_{1:K}$, the extracted latent features can be clustered into slower and faster changing groups, respectively. Given a new sample $\boldsymbol{y}$, slower and faster feature components can be calculated as below:

$$\begin{aligned} \boldsymbol{v}_s &= \boldsymbol{W}_s\boldsymbol{y} \\ \boldsymbol{v}_f &= \boldsymbol{W}_f\boldsymbol{y} \end{aligned} \tag{3.8}$$

where $\boldsymbol{W}_s$ and $\boldsymbol{W}_f$ are the corresponding rows of coefficient matrix $\boldsymbol{W}$ governing slower and faster dynamics, respectively. Then, the Hotelling's T^2 statistics are computed for the two feature groups as shown in Equation (3.9).

$$
\begin{aligned}
T_s^2 &= \boldsymbol{v}_s{}^T \boldsymbol{v}_s \\
T_f^2 &= \boldsymbol{v}_f{}^T \boldsymbol{v}_f
\end{aligned}
\tag{3.9}
$$

This enables us to monitor the features with higher and lower velocities, separately.

3.4.1.3 Online Model Update Strategy for PCA-SFA Approach

In the time-domain monitoring approach, to get adapted to the normal process drifts, both PCA and SFA models need to be updated during online monitoring process. In normal operating period, all the units are running in a relatively stable status without drastic oscillations. The latter is a potential indication of flares. Therefore, the model update strategy is designed to adapt to the process drifts within a normal range.

For a reliable model update, instead of using the most recent data, a lag is selected between the data used for model update and the current time instant. Then the historical data in a specified window length are extracted to determine if the update is necessary.

3.4.2 The Frequency-domain Hierarchical Monitoring Approach

In this section, the second proposed approach, namely hierarchical frequency-domain process monitoring approach, will be introduced. Variability in a multivariate dataset obtained from complex processes is expected to change not only in time domain but also in frequency domain. Such processes are multi-scale processes and in order to detect changes in such processes, it is important to analyze them in both time and frequency domains [74]. To extract frequency information from a signal, the Fourier transform (FT) is usually used to decompose the signal into its frequency constituents over the entire time period. However, FT does not provide any indication when the changes in frequencies occur with time. Alternatively, the WT can be used as a tool for providing the time-frequency multi-resolution information of a signal [75] in such cases.

The frequency information provided by WT is referred to as scales and has the advantages of being decorrelated compared with the original time-domain signals that are auto-correlated [76]. Consequently, PCA can be applied on the non-auto-correlated scales for the purposes of process monitoring. This is another advantage of using scales in SPM in addition to the time localization property.

For online application, a moving window approach is employed first to extract frequency information, i.e., scales, from the PVs at every sampling instant. Subsequently, an adaptive PCA model is then built based on the scales and is used to predict the monitoring statistics, namely, T^2 and the Q statistic. The proposed approach is referred to as moving window multi-scale adaptive PCA (MWMSAPCA) and the details of each component in this approach are presented next.

3.4.2.1 Wavelet Transform

WT is a spectral decomposition method of a time-dependent signal that provides a set of frequency bands known as scales that represent time-dependent frequencies [75]. Given the original signal $X(t)$ and an analyzing function $\phi_{a,b}(t)$, the scales are computed as follows:

$$Scales(a,b) = \int_{-\infty}^{+\infty} X(t)\phi_{a,b}(t)dt \tag{3.10}$$

The analyzing function $\phi_{a,b}(t)$ represents a family of wavelets scaled by a parameter b and translated by another parameter a as follows:

$$\phi_{a,b}(t) = b^{-\frac{1}{2}}\phi(\frac{t-a}{b}) \tag{3.11}$$

As a result, the original time-domain signal $X(t)$ is transferred to a multi-resolution frequency-time scale denoted by $Scales(a,b)$. The obtained scales are ranked from the smallest to the largest, where the smaller scales represent the evolution of high frequency signals with time, while the larger scales represent the evolution of low frequency signals with time.

3.4.2.2 Moving Window Wavelet Transform for Online Application

For online applications, WT is applied on a selected group of PVs over a moving window of length M to extract online frequency features, i.e., the scales, at every sampling instant. Analyzing a window of PVs with dimension $l \times M$ at every sampling instant t results in wavelet scales of dimension $(l_s \times l) \times M$ as shown in Equation (3.12), where l_s is the number of scales obtained from each PV.

$$Scales(t - M : t) = WT(PVs(t - M : t)) \tag{3.12}$$

where the function WT represents the wavelet transfer analysis on the selected PVs, and $Scales(t - M : t) \in \Re^{(l_s \times l) \times M}$ denote the obtained scales over the moving window of length M at the sampling instant t.

The maximum value of every moving window scale, i.e., $\max(Scales(t - M : t))$, is considered as the representative frequency feature at the current time t of the moving window PVs, i.e., $PVs(t - M : t)$, and is denoted by $Scales_{max}(t) \in \Re^{(l_s \times l) \times 1}$. The reason for selecting the maximum point of scales, other than the middle or the end point from the moving window, is that it results in clearer and more persistent features related to flare events.

3.4.2.3 Online Adaptive PCA Based Process Monitoring Using Frequency Domain Information

After using moving windows of $PVs(t - M : t)$ to obtain a sequence of wavelet features $Scales_{max}(t)$ up to the current time t, a PCA model is trained on these scales and is used to monitor the process at time $t + 1$ by using $PVs(t - M + 1 : t + 1)$ as a new sample. The relevant statistics, thus computed, will be integrated later in a hierarchical layer.

To avoid any potential issue arising due to the non-stationary behaviour of the data, we also propose a model update strategy as presented below.

The update rules can be derived based on the variability of the process, that is, if the variability is considered as within a pre-determined tolerance level, a model update is calculated. If the variability is beyond a threshold, it may indicate the beginning of a potential flare event, and therefore the model update needs to be avoided.

In this work, the process variability is quantified by the difference between two successive maxima of standard deviations of two successive scales, as shown in Algorithm 1.

Algorithm 1 The PCA model update strategy

```
if $\max(std(Scales(t - M : t))) - \max(std(Scales(t - M - 1 : t - 1))) < thr$ then
    Update the PC model
else
    Keep using the previous PC model
end if
```

The model updating threshold thr, the moving window length M, and the update rate are tuning parameters and can be selected empirically. For hierarchical monitoring, an additional layer of moving window adaptive PCA is added to integrate the monitoring of the selected PCs as shown in Fig. 3.4, to obtain a single monitoring Q_{stat} for overall decision making.

3.4.3 Fault Isolation under the Hierarchically Distributed Framework

Having proposed both time-domain and frequency-domain hierarchical approaches, we turn to isolation problem in this section. Under the proposed hierarchically distributed framework, fault isolation can be conducted to pinpoint the affected units. The most affected units can be first determined based on the decomposed input variable contributions to the overall monitoring statistics. Then the decomposed input variable contribution to each affected unit's monitoring statistics can be used to isolate the responsible PVs. From various existing fault isolation indexes that are reported in literature [77], the reconstruction-based contribution (RBC) index has been selected for this purpose because of the demonstrated

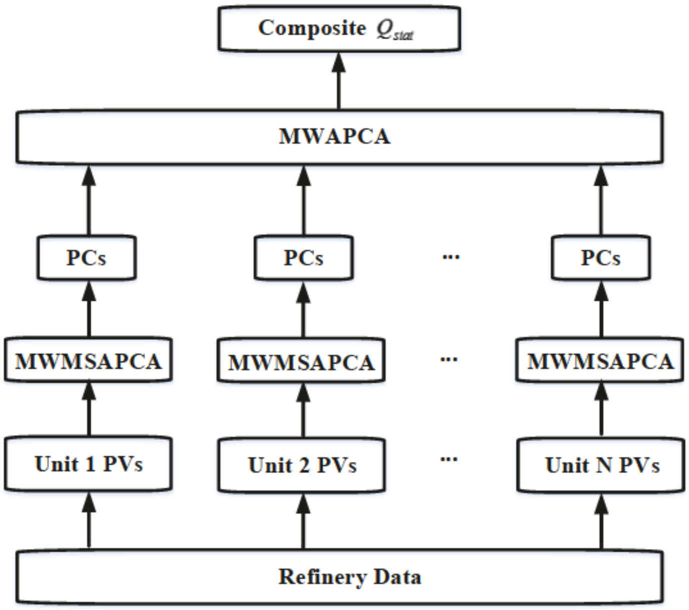


Figure 3.4: Refinery data are grouped by unit, and the MWMSAPCA strategy is applied on every unit for monitoring. All the obtained PCs are then integrated into a moving window adaptive PCA modeling strategy to yield a composite Q_{stat} index that monitors the whole process for early flare event prediction

diagnosability for various fault types [78]. And the relative RBC (rRBC) index is finally selected for more accurate diagnosis results [77]. The rRBC calculations of PCA and SFA approaches can be found in Table 3.5, where the matrices W_s, W_f and P can be selected by referring Equation (3.4) and (3.8).

Table 3.5: rRBC for fault isolation [77, 73]

rRBC	SFA		PCA
	T_s^2	T_f^2	Q *statistic*
$rRBC_i = \frac{(\xi_i^T \boldsymbol{M x})^2}{\xi_i^T \boldsymbol{MDM} \xi_i}$	$\boldsymbol{M} = \boldsymbol{W}_s^T \boldsymbol{W}_s$	$\boldsymbol{M} = \boldsymbol{W}_f^T \boldsymbol{W}_f$	$\boldsymbol{M} = \boldsymbol{I} - \boldsymbol{P}\boldsymbol{P}^T$
	$\boldsymbol{D} = \frac{1}{T-1} \boldsymbol{Y}_{PC}^T \boldsymbol{Y}_{PC}$		$\boldsymbol{D} = \frac{1}{T-1} \boldsymbol{PC}^T \boldsymbol{PC}$

Here $\xi_i = [0, \cdots, 1, \cdots, 0]^T$ represents the activation of the i^{th} evaluated variable.

3.5 Application: Early Gas Flare Event Prediction in A Refinery

In this section, the performance of both hierarchical monitoring approaches are evaluated for their ability to predict flare events before they happen. The hierarchical time-domain approach is first employed for flare prediction, followed by the hierarchical frequency-domain approach. Then the flare prediction performances are summarized. Finally, by taking the first flare event as an example, the fault isolation performance is illustrated based on the rRBC index, and the most likely source unit is identified.

3.5.1 Hierarchical Time-domain Early Flare Event Prediction

In the time-domain hierarchical monitoring approach, the PCA algorithm is used for preliminary feature extraction in the bottom layer. For obtaining a trade-off between useful information and noise, only the first two dominant PCs of each unit are aggregated for overall monitoring decision making. For each unit, the first five days data are selected to build the initial model and the first 10 days PCs collected from the bottom layer are employed for top layer model building. To make this approach adapt effectively to the process variations throughout the entire year, all the models are checked and updated online everyday if the mean of standard deviation is within the normal range. In this work, the slower and faster features in the SFA model are grouped by inspecting the increment between two adjacent slowness values of the extracted slow features. The transition between slower and faster feature groups is triggered when the increment of two adjacent slowness values changes significantly and consistently. In this case, the control limits on both slower and faster monitoring statistics updated once the statistical monitoring models are updated. To achieve a more reliable fault detection performance, the control limit is selected at the 99% confidence level.

Furthermore, in order to avoid false positives caused by random jumps of the monitoring statistics with small magnitudes, a comprehensive alarm invoking strategy is proposed. Instead of activating alarms simply based on the currently generated monitoring statics, a sequence of most recent monitoring statistics is investigated comprehensively to invoke the alarms. Here, based on the deviations of the monitoring statistic and the calculated control limit, two levels of alarms are defined. Low alarms are generated when the analyzed monitoring statistic jumps over the control limit but no larger than twice the control limit, and lasts for 300 minutes. High alarms are generated when the analyzed statistics are larger than twice the control limits, with a duration of 200 minutes.

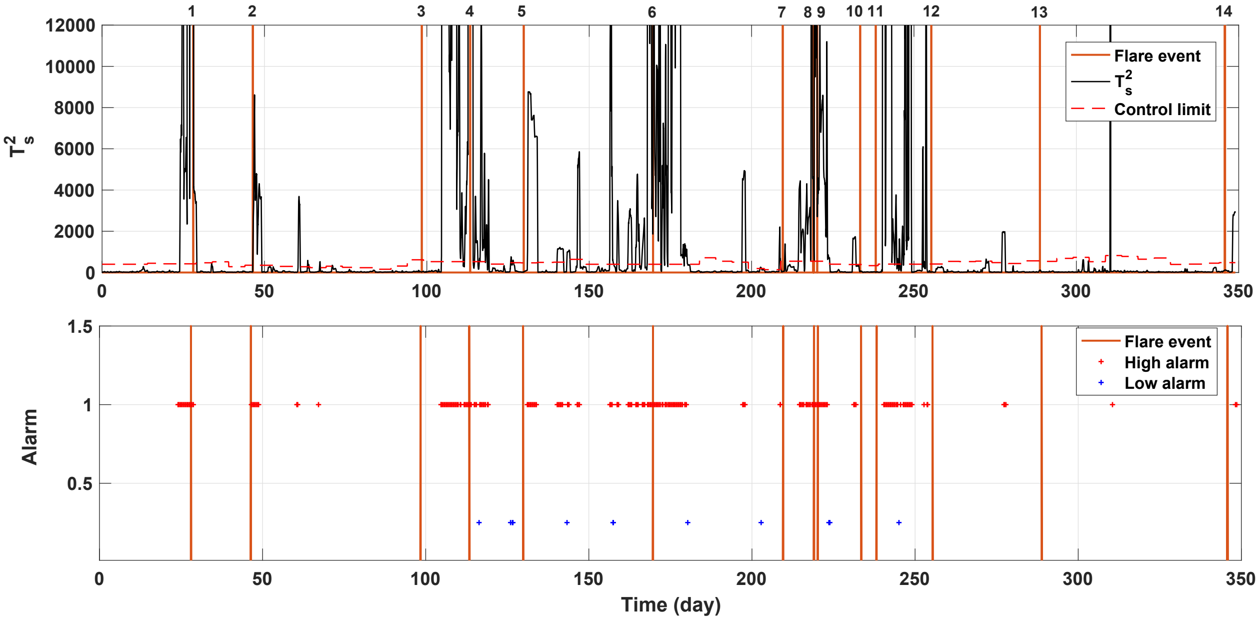


Figure 3.5: The hierarchical PCA-SFA slower feature group monitoring result of all the flare events

The overall monitoring and alarming results of the slower and faster clusters are listed in Fig. 3.5 and 3.6, respectively. By evaluating low and high alarm percentages over the entire year period, the results of the early flare prediction performance are reported in Table 3.6.

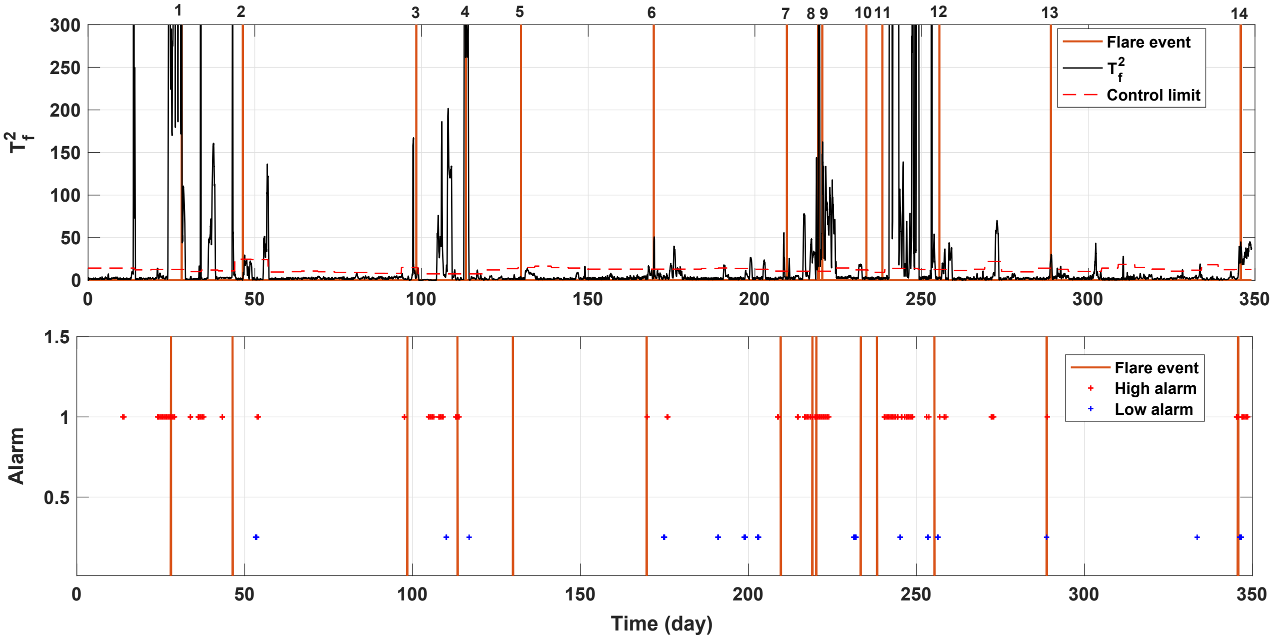


Figure 3.6: The hierarchical PCA-SFA faster feature group monitoring result of all the flare events

Table 3.6: Alarm evaluation of PCA-SFA.

	T_s^2		T_f^2	
	Potential predictions	False rate	Potential predictions	False rate
High alarm	7/14	4.508%	8/14	4.229%
Low alarm	0/14	0.404%	1/14	1.029%

3.5.2 Hierarchical Frequency-domain Early Flare Event Prediction

The hierarchical frequency-domain approach is applied on the process data from the bottom layer to the top layer. We have used the embedded continuous wavelet transform function (cwt) to obtain the wavelet scales, and the analytic Morse wavelet is selected by using the WAVELET toolbox in MATLAB. The minimum and maximum scale numbers are determined automatically based on the energy spread of the wavelet in frequency and time. The length of the moving window M in Equation (3.12) is selected to be two weeks. The unit-wise model update frequency is tuned to be six hours as long as the process behaviour is classified as normal. The model updating threshold *thr* described in Algorithm 1 is tuned to obtain optimal results across all the units. For the hierarchical modeling, the first three PCs are selected from each unit and passed on to the hierarchical layer, and the model update frequency and threshold *thr* are kept the same as those for unit-wise monitoring. The control limit is selected based on the training dataset at the 95% confidence level. The extracted frequency features show smoothly changing patterns during normal operation and there are few occasionally occurred small spikes, so oscillations in the frequency features are more likely to be correlated with flare events and should be paid much attention to. Therefore, compared with the time-domain approach, in the frequency-domain approach, more PCs from each unit can be selected to include more information without introducing too much noise, and the control limit is selected at 95% confidence level to increase detection sensitivity owing to relative smoothness of the monitoring statistics. The low alarms are flagged when the Q statistic jumps beyond the control limit but with magnitude less than twice the control limit, and the high alarms are flagged when the Q statistic magnitude is larger than twice the control limit. The hierarchical monitoring performance results can be found in Fig. 3.7, where the Q statistic and the corresponding alarms are presented in the first and second subfigures, respectively.

Based on the frequency-domain hierarchical monitoring performance, the potential flare predictions and false alarm rate are calculated and summarized in Table 3.7.

Table 3.7: Alarm evaluation of hierarchical MWMSAPCA

	Q statistic	
	Potential predictions	False rate
High alarm	5/14	4.37%
Low alarm	7/14	8.65%

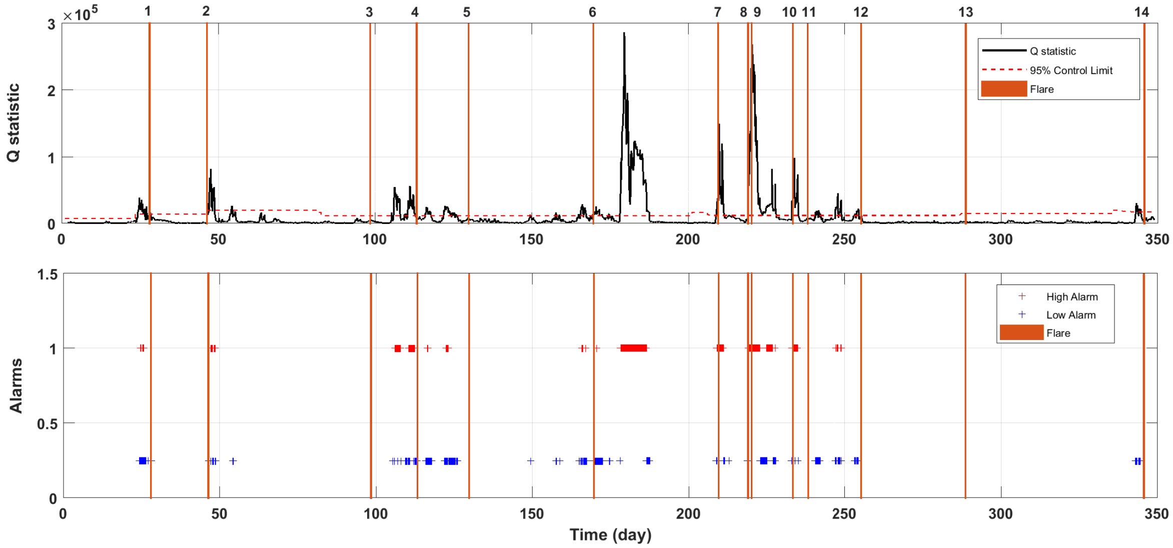


Figure 3.7: The hierarchical MWMSAPCA monitoring result of all the flare events

3.5.3 Discussion on the Early Flare Event Prediction Performance

In this section, we zoom into the results of the early flare predictions considering both the frequency and time-domain monitoring approaches, and also compare with the traditional centralized PCA method.

In the traditional PCA approach, all the available PVs are utilized and a PCA model is trained by using first five days data. The control limit at the 99% confidence level is selected for an optimal trade-off between flare predictions and false positives, and this follows the same alarm logic strategy as in hierarchical time-domain approach. The prediction performance can be found in Fig. 3.8, where the first subfigure shows the T^2 statistic of the PCA model and the second subfigure represents alarms generated according to the control limit.

Here, the flares 1, 4, 6, 9, 10 can be clearly detected by this method, but 9 out of 14 flares did not get detected. The traditional centralized PCA approach treat all the PVs as one group, and as a result some useful information in some PVs might be overwhelmed by the dominant variations reflected by a majority of the PVs. On the other hand, from the hierarchical monitoring performance as shown before, some of the flares can be predicted by

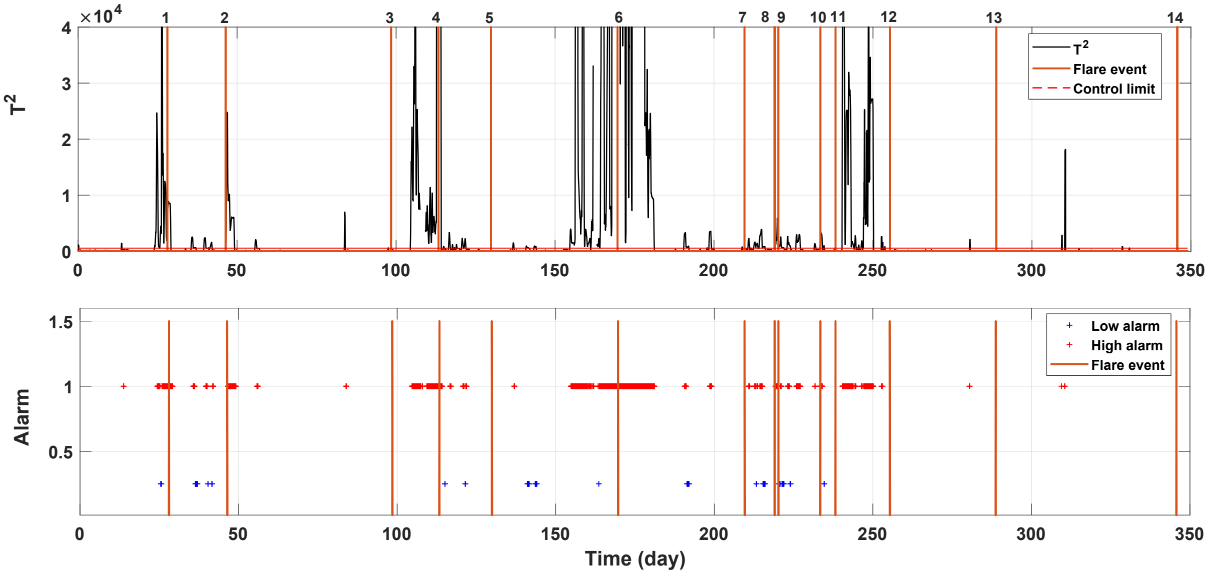


Figure 3.8: The early flare prediction performance of the traditional PCA approach

both of the time and frequency-domain approaches, while some can be predicted by only one of the two approaches, indicating that the two proposed approaches can complement to each other. In Table 3.8, all the flare prediction results are compiled for comparative evaluation.

Table 3.8: Predicted flares by all the available approaches

Predicted approaches	Traditional PCA approach	MWMSAPCA	H-PCA-SFA
Identity of predicted flare	1, 4, 6, 9, 10	1, 4, 8, 9 10, 12, 14	1, 3, 4, 6, 7, 8 9, 10, 13, 14
Total number	5	7	10

Compared with the traditional PCA approach, the hierarchical frequency-domain approach provides three extra early predictions, i.e., flares 8, 12 and 14. The hierarchical time-domain approach has five more early predictions. However, compared with the time-domain approach, the frequency-domain approach is able to predict flare 12 clearly, while it is missed by the time-domain method. The time-domain approach is able to provide early flare predictions of flare 3, 6, 7 and 13, which are missed by the frequency-domain method. A total of 11 out of 14 flare events could be predicted by our proposed approaches, if both of the proposed hierarchical time-domain and frequency-domain approaches are used simultaneously.

Therefore, two proposed approaches can be used to support each other for a better prediction performance, resulting in an obvious improvement compared with the traditional PCA based monitoring as well as the existing approach attempted [56]. The reasons for missed detections from both approaches could be various. For instance, from the data inspection it was clear that some sensors were bad during the period of flare 5, resulting in missing data. Moreover, the available data had only a limited number of units from the refinery, and a limited number of PVs for each unit. Hence, addition of data from other units and PVs could be worthwhile for gaining more information related to the missed flare prediction.

Additionally, the false rates of all the employed approaches have been calculated and compared, which are all restricted in a small range. As time-domain monitoring algorithms, the false rates of both traditional PCA and hierarchical PCA-SFA approaches are listed in Table 3.4 and 3.6, from which one can conclude that the T^2 statistics of PCA and PCA-SFA provide comparable false rates within the investigation time. By observing Figure 3.5 and 3.6, the alarms from T_s^2 and T_f^2 in PCA-SFA approach are visualized and most of the false rates of the two metrics are overlapped, so the final false rates of PCA-SFA approach is less than the addition of the false rates from T_s^2 and T_f^2. As a comparison, Figure 3.8 shows the false rate of PCA, which is sparse but meanwhile PCA approach lost effective predictions of majority flare events. Therefore, during actual online implementation, if one treats every alarm seriously, the PCA-SFA approach will provide more effective alarming information than the traditional PCA approach.

3.5.4 Faulty Unit Isolation at the Hierarchical Level

In this section, the fault isolation at the hierarchical layer will be investigated. The faulty units can be tracked through the top layer by evaluating the contribution of each PC when an alarm occurred. Taking the first flare event for instance, the rRBC plots corresponding to the time and frequency domain approaches are presented in Fig. 3.9 and 3.10, respectively.

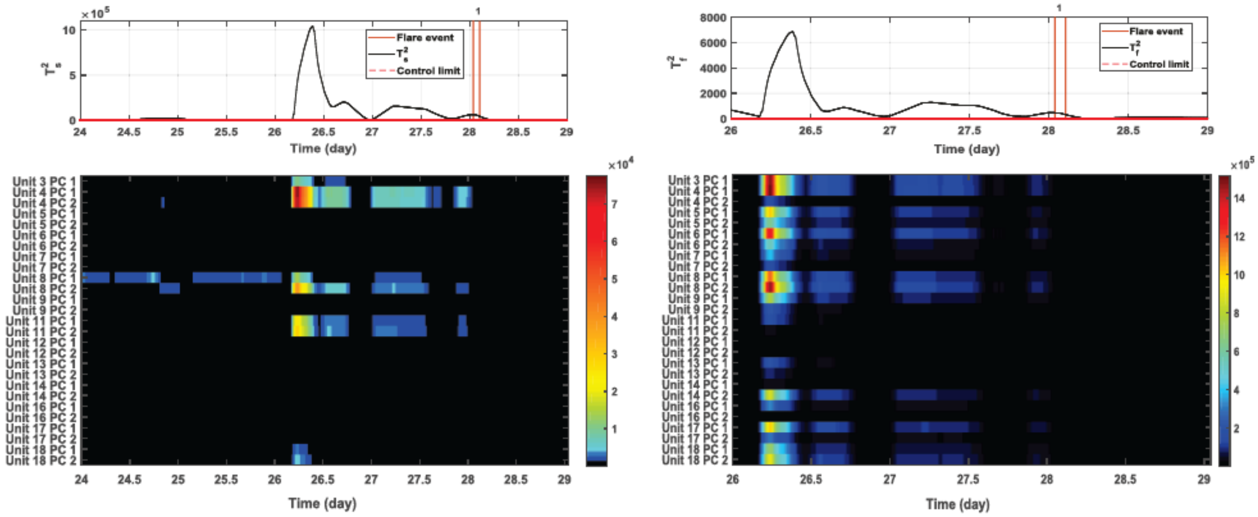


Figure 3.9: The unit contributions to the first flare prediction of the time-domain PCA-SFA approach. The left two subfigures indicate the T_s^2 statistic and the corresponding unit-wise contribution, respectively, and the right two subfigures represent the T_f^2 statistic and the corresponding unit-wise contribution, respectively

The above figures indicate the contributions of different units in different approaches along with time when the first flare event occurred. The color indicates the magnitude of the rRBC index, where the red color indicates high contributions of individual units. From the above figure, one can see that the slower and faster feature groups indicate that the units 3, 4, 6 and 8 are the most likely source units contributing to the flare events based on time-domain analysis. From the frequency domain analysis result, the units 4, 5 and 10 should be considered as possible causes. Based on the analysis of faulty unit isolation over all the flare events, it has become evident that the unit 4, i.e., hydrocracking unit, is the main contributor of multiple flare events.

Furthermore, the robustness of rRBC to noise and outliers is evaluated in this work. Since the calculation of rRBC index is based on the MSPM models, namely, SFA and PCA, the robustness of rRBC to noise and outliers is closely related to its corresponding MSPM algorithms. In SFA, the noise is absorbed by the fast features and in PCA, the noise is

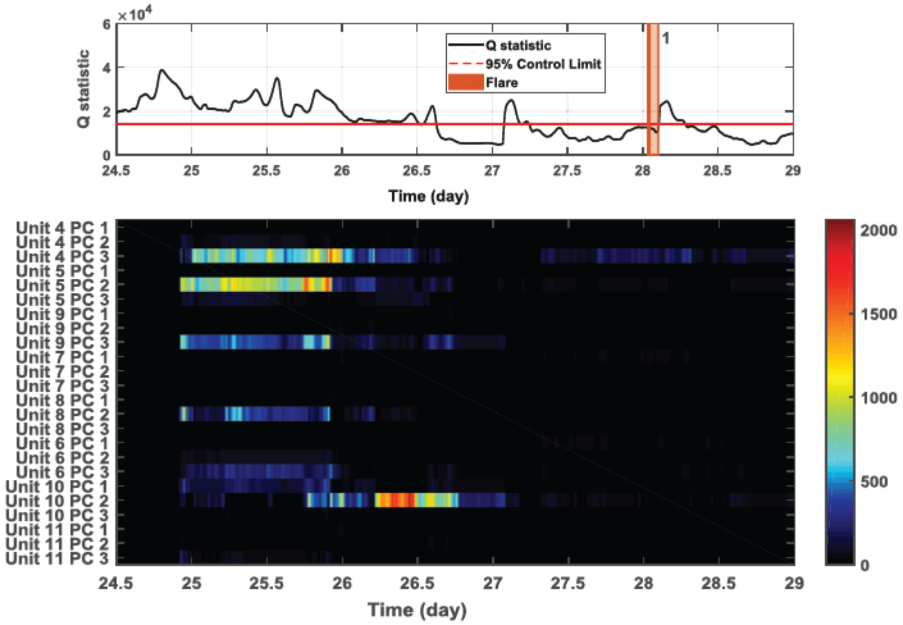


Figure 3.10: The unit contributions to the first flare prediction of the frequency-domain MWMSAPCA approach. The first and second subfigures are the Q-statistic of PCA and the corresponding unit contribution plot, respectively

absorbed by the minor PCs. Therefore, the slow features or main PCs are not sensitive to noises. Since at the bottom layer of the hierarchical structure, the first few main PCs of individual unit are selected for further analysis, the noise does not provide a significant impact on the extracted features and the rRBC indices. In terms of outliers, even though the conventional SFA and PCA approaches are not specifically designed to resist the outliers, by making use of the hierarchical structure, the influence of outliers can also get reduced. In addition, data preprocessing before applying the monitoring algorithms also prevent outliers from influencing the final detection results. In order to evaluate the sensitivity of rRBC index to the outliers in a general case, different percentages of the outliers are simulated and directly added onto the extracted features, and the rRBC indices of slower feature group around the first flare event are taken as an example to perform the investigation. We have simulated various percentages of outliers (1%, 2% and 5%) to test their effects as shown in the following figure. As it can be seen that the algorithm can indeed resist the outliers. Particularly the dominant contribution plot is quite robust to the outliers.

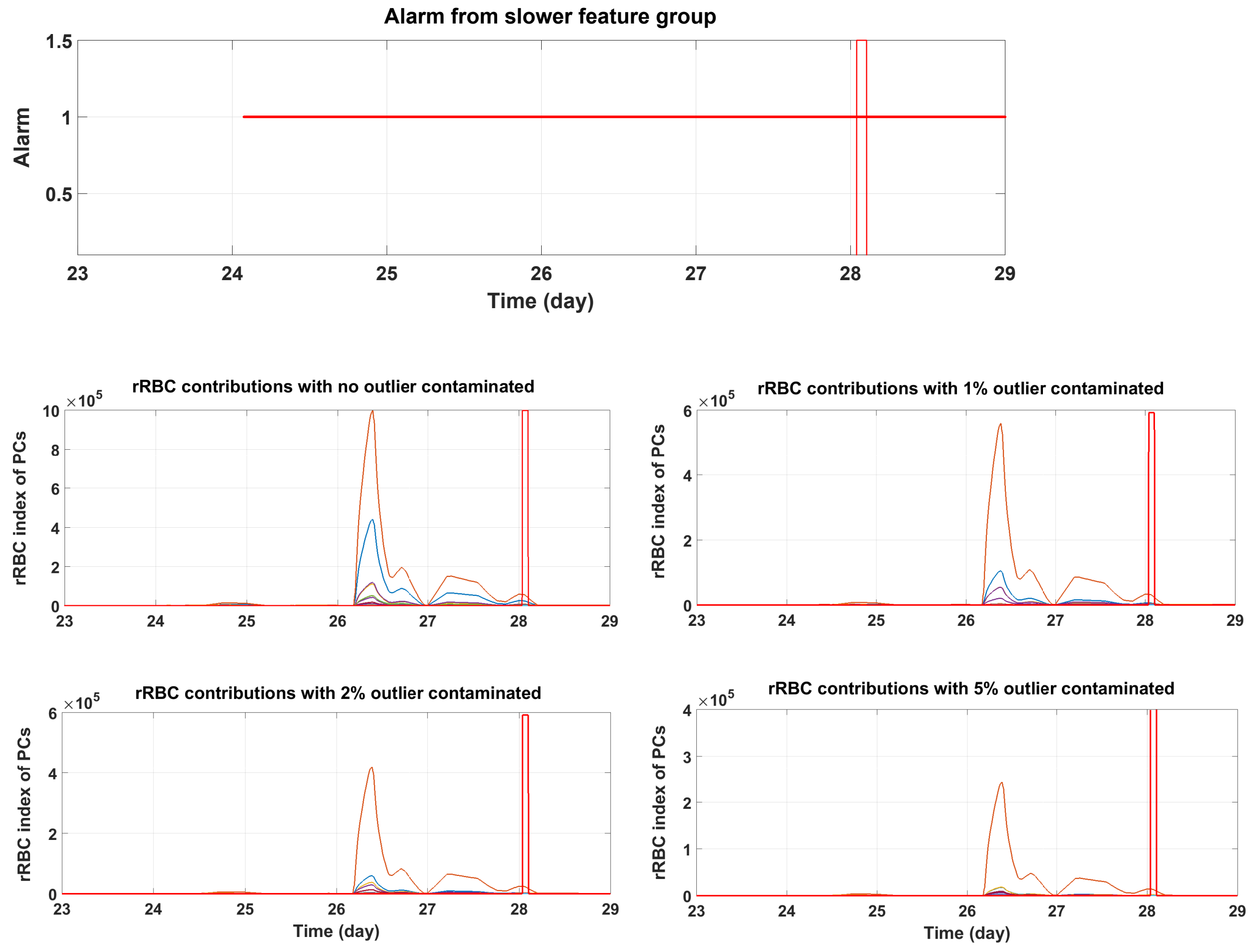


Figure 3.11: The comparison of the rRBC indices of PCs with different percentage outlier contaminations, with respect to the first flare event in the slower feature group

As a result, the proposed hierarchical structure can reduce the impact of noise and outliers. But if there exist a large portion of outliers on the extracted features, to get more accurate fault detection and isolation results, the outlier removal procedures can be used in the preprocessing stage.

3.6 Conclusions

In this chapter, we have proposed various strategies for early flare event prediction and tested them successfully on a set of industrial refinery data. A hierarchically distributed process monitoring framework was developed as an effective solution for flare prediction, even though we had limited access to process information. Based on this framework, the available PVs were first grouped according to their respective process units, and the features extracted from the units were integrated in a hierarchical layer for overall monitoring and decision making. Two hierarchical approaches, namely, hierarchical PCA-SFA and the hierarchical MWMSAPCA approaches, were developed for the early flare event predictions. The results show that 11 out of 14 flare events could be predicted in advance by the two approaches with an acceptable rate of false positives, and 10 out of 14 flare events could also be predicted based only on the single hierarchical time-domain approach. Finally, similar to the other unsupervised process monitoring algorithms, the proposed algorithm is also restricted from not using any fault information. To make full use of both process data and the fault information, the supervised process monitoring approaches are worth to be attempted. Given the potential advantages of the supervised approaches, this thesis goes on to explore and develop novel supervised approaches.

Chapter 4

A Novel Approach to Process Operating Mode Diagnosis Using Conditional Random Fields in the Presence of Missing Data

4.1 Introduction

In industrial processes, the two most critical requirements are the safety and consistently high product quality. The adoption of flexible process designs and operational strategies in process industries demands for advanced process monitoring approaches in order to reduce operational risks and potential safety hazards. The task of process monitoring includes operating mode diagnosis, detection and diagnosis of faults [79, 64], and determination of their root causes. The existing process monitoring algorithms can be classified into methods that employ empirical knowledge, first principles based models [80] and data based models

[2] Part of this chapter has been published as Mengqi Fang, Hariprasad Kodamana, Biao Huang, and Nima Sammaknejad. A novel approach to process operating mode diagnosis using conditional random fields in the presence of missing data. *Computers & Chemical Engineering*, 111:149–163, 2018.

[81]. As models based on first principles are complex under most of the industrial scenarios, it is difficult to obtain comprehensive physical and mathematical process models. As a result, data based models have emerged as effective alternatives to the first-principles models [12], among which some simplified state-space models [82, 83] and statistical models [84] have been employed for process monitoring.

Owing to external environmental fluctuations and process uncertainties, industrial process data usually exhibit various characteristics such as nonlinearity, non-Gaussianity, multimode [85] and temporal correlations [86]. Even though many MSPM approaches such as PCA, PLS and ICA, etc. [87] have been developed to deal with various challenges mentioned above, an inherent assumption for most of them is the unimodality of the data which is not easily satisfied in reality. As a result, the HMM has become a popular framework and has been widely employed for the diagnosis of the multimodal dynamic processes owing to its ability to model operating mode transitions. *Sammaknejad et al.* [88] proposed an HMM based adaptive monitoring strategy for fault detection of the primary separation unit, a key process unit in the oil sands extraction process in Northern Canada. Also, there were several studies to combine the unimodal MSPM techniques with HMM for online process monitoring, such as the PCA based HMM approach proposed in [89], the adaptive ICA based blended HMM approach developed in [90], HMM based statistics pattern analysis algorithm proposed in [91], and the references therein. Although the HMM approach models the switching mechanism of process operating modes [92], it is limited by two inherent conditional independence assumptions, namely, (i) in first order HMMs, the current state is considered to be dependent on the state immediately prior to it and independent of all other previous information given the state prior to it, (ii) the current observation is only dependent on the current state and independent of the other past states given the current state.

As a probabilistic generative model, the HMM framework employs the above mentioned conditional independent probability assumptions to factorize the joint distribution of sequential observations. The diagnosis performance could be adversely affected if these assumptions are broken down by reality. In order to address this lacunae, we propose to employ the CRF, a probabilistic discriminative model, initially proposed by [43]. While retaining the properties of HMMs that enable state transition descriptions, CRFs avoid intricate computation of the prior distribution of observations and have a more flexible framework to comprehensively describe the temporal autocorrelations among the observations [93]. CRFs use feature functions to model the dependency relations among the variables. By choosing appropriate feature functions as a special case, the CRF model has been proven to be equivalent to the HMM [94]. Moreover, unlike HMMs, the training of CRFs involves only the convex objective functions, which helps to obtain global optimal parameter estimation [27]. Recently, the CRF framework has been used extensively in the fields of natural language processing, image processing and speech recognition, etc., and has shown to perform superiorly to HMMs [95, 96, 97]. Even so, CRFs have not been employed for the online process monitoring of industrial processes where the data shows complex temporal dependencies. Employment of CRFs in online process monitoring is expected to solve some outstanding issues which cannot be effectively solved by HMMs.

Another significant issue while dealing with industrial data is missing measurements. The missing measurement problem is usually due to different sources such as sensor failures and other data collection errors. In order to deal with operating mode diagnosis problems that include missing measurements, an HMM based approach was proposed [37] based on the EM algorithm. Similarly, *Zhang et al.* [98] also proposed an EM approach to fault diagnosis with missing data. Moreover, *Koushanfar et al.* [99] designed a semi-Markov model to estimate the statistical patterns of missing measurements for fault detection and diagnosis. Even though issues such as missing labels in the context of CRFs have been studied previously [100], hardly any of the published work is concerned with the implications of missing measurements in the modeling of CRFs. In a related work, *Dietterich et al.* [101] used an imputation based approach to fill missing measurements with non-missing measurements in CRF modeling. However, this might lead to a biased estimation of parameters. In order to solve these problems, in our proposed work, an enumeration based approach to account for all possible missing measurement combinations is considered and included in the CRF model, and the relative importance of different possible combinations of missing measurements is attributed by means of the weights of CRF. Moreover, to reduce the large computational load occurring due to these steps, an efficient propagation algorithm will be developed for this marginalization framework.

The remainder of this chapter is organized as follows: Section 2 briefly reviews the general formulation of LCCRF and the corresponding framework for industrial process operating mode diagnosis. In section 3, a new marginalized CRF framework is proposed for process operating mode diagnosis when there exist missing measurements; based on this framework, a new propagation algorithm is developed to reduce the computational load. In section 4, the simulated CSTR system and the experimental hybrid tank system are employed for process monitoring performance validation. Finally, conclusions of the study are presented in section 5.

4.2 LCCRF Model for Process Operating Mode Diagnosis

4.2.1 Preliminaries

In this subsection, a general formulation of the LCCRF model will be illustrated. As shown in Fig. 4.1, the general LCCRF model is an undirected graphic model that describes the connections among a set of labels $\boldsymbol{h}$, which are the operating modes in our case, and a set of observations $\boldsymbol{O}$. Let the observation sequence be $\boldsymbol{O} = \{O_1, O_2, ..., O_T\}$ and the system mode sequence be $\boldsymbol{h} = \{h_1, h_2, ..., h_T\}$. Then, CRF models the conditional probability $P(\boldsymbol{h}|\boldsymbol{O})$ as follows [27]:

$$P(\boldsymbol{h}|\boldsymbol{O}) = \frac{1}{Z(\boldsymbol{O})} \exp \sum_{t=1}^{T} \{ \sum_{k=1}^{K} \lambda_k T_k(h_t, h_{t-1}) + \sum_{m=1}^{M} \mu_m E_m(h_t, \boldsymbol{Y}_t) \} \tag{4.1}$$

where $\boldsymbol{Y}_t \subseteq \{O_1, O_2, ..., O_T\}$ is a vector containing all the observations that are needed to model $P(\boldsymbol{h}|\boldsymbol{O})$ at time t, and T represents all the time instances considered. $Z(\boldsymbol{O})$ is an instance-specific normalization factor obtained by marginalizing the numerator of Equation (4.1) over all possible labels, as given by Equation (4.2) below:

$$Z(\boldsymbol{O}) = \sum_{\boldsymbol{h'}} \exp \sum_{t=1}^{T} \{ \sum_{k=1}^{K} \lambda_k T_k(h'_t, h'_{t-1}) + \sum_{m=1}^{M} \mu_m E_m(h'_t, \boldsymbol{Y}_t) \} \tag{4.2}$$

Here, the notation $\boldsymbol{h'}$ denotes all possible combinations of the labels. The function sets $\{T_k(h_t, h_{t-1})\}_{k=1}^{K}$ and $\{E_m(h_t, \boldsymbol{Y}_t)\}_{m=1}^{M}$ are binary or real-valued functions, called as feature functions, and are typically selected based on the nature of the problem [102]. The number of feature functions, K and M, need to be chosen adequately to model the dynamics and are problem specific parameters. The parameter vectors $\boldsymbol{\Lambda} = \{\lambda_k\}_{k=1}^{K}$ and $\boldsymbol{\mathcal{M}} = \{\mu_m\}_{m=1}^{M}$ contain the corresponding weight factors of the feature functions, which are used to differentiate the importance of individual feature functions and need to be estimated for a specific observed dataset.

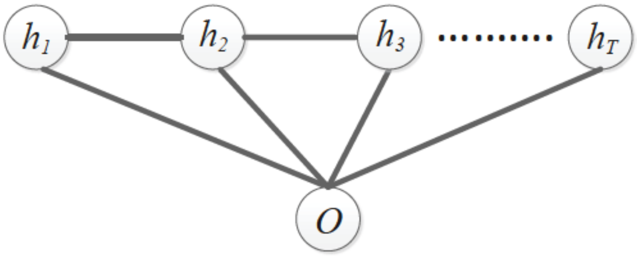


Figure 4.1: The general graphical structure of LCCRFs [103]

Compared to the first order HMMs illustrated in Fig. 4.2, the LCCRF models use multiple observations to model the label at a specific time point, and therefore, are better suited to model rich and complicatedly correlated data attributes, thus resulting in a more general modeling formalism [93].

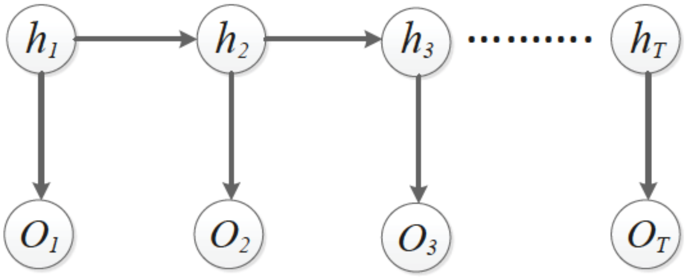


Figure 4.2: The general graphical structure of HMMs [104]

Remark 1 *By choosing $\mathbf{Y}_t = \{O_t\}$ and appropriate feature functions, the HMM and LCCRF can be shown to be equivalent from a modeling perspective. In contrast to HMM, since the weight parameters of CRF have no probabilistic interpretation, their summation is not required to be equal to unity. Details regarding the same can be found in the Appendix A.*

Now that we have introduced LCCRFs, we present our strategy for operating mode diagnosis using CRFs.

4.2.2 Operating Mode Diagnosis Using LCCRFs

In literature, there are few references related to LCCRF based process monitoring. Even though *Wang et al.* [45] proposed a LCCRF based fault classification framework on a bearing system, the decoding algorithm they used still restricts the extension to online process monitoring. This problem is addressed in this work.

For the process operating mode diagnosis problem, we assume that the process system operates in various operating modes such as $Normal$, *Abnormal*, $Faulty$, etc.. At each time point t, let the operating mode be represented as $h_t = i \in \{1, 2, ..., N\}$, where N is the total number of possible process operating modes. In this work, we assume that both the modes $\boldsymbol{h}$ and observations $\boldsymbol{O}$ are discrete. Since T samples of observations are considered for the analysis, the observation sequence becomes $\boldsymbol{O} = \{O_1, O_2, ..., O_T\}$, and our objective is to extract the operating mode sequence $\boldsymbol{h}$, given the observation sequence $\boldsymbol{O}$.

Considering the predictability of the process operating modes, there exist several constraints on the operating mode conditions and the selected observations. For a purely data-based operating mode diagnosis problem as in our case, the process operating modes are assumed to be observable from the process measurements, which is an inherent premise to solve this problem. For example, if some of the abnormal operating modes are not observable from the available process data, then highly likely that, it is not possible to detect such abnormality from the observations. In order to have accurate and timely operating mode diagnosis, the PVs have to be selected in such a way that it would have relatively quick and distinctive signatures of different operating modes and mode changes are manifested through the selected candidate variables' observations. If the PVs are chosen improperly or haphazardly, for example variables that bear with a huge reaction delay or uncorrelated response, it may eventually result in incorrect diagnosis results.

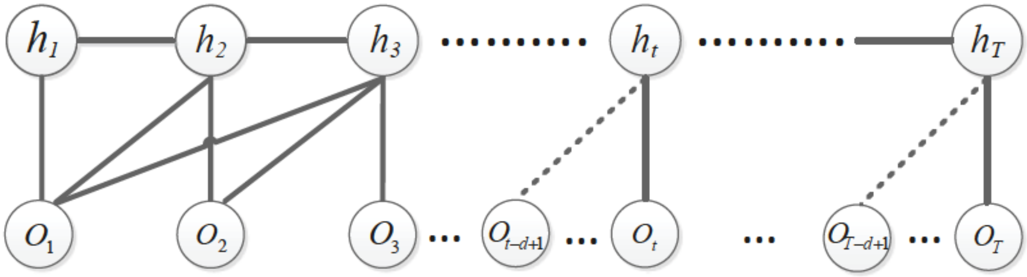


Figure 4.3: The LCCRF structure designed for process operating mode diagnosis problem

As shown in Fig. 4.3, we assume that the operating mode at sampling time t, i.e. h_t, is dependent on both the operating mode at previous sampling time h_{t-1} and a sequence of observations, i.e. $\boldsymbol{Y}_t = \{O_t, O_{t-1}, ..., O_{t-d+1}\}$. Compared with the HMMs, this framework allows us to model the state transitions with the Markov property, and meanwhile, it is more flexible and therefore can model the observation autocorrelations which are introduced by factors that have no explicit relationship with the states, such as those out of the external environment change.

The feature functions for this problem need to be defined as next. We define two sets of feature functions: (i) a function that relates the operating modes h_t and h_{t-1}, i.e. $T_k(h_t, h_{t-1})$, and (ii) a function that relates the operating mode h_t and a sequence of observations $\boldsymbol{Y}_t = \{O_t, O_{t-1}, ..., O_{t-d+1}\}$, i.e. $E_m(h_t, \boldsymbol{Y}_t)$. Since the sets $\boldsymbol{h}$ and $\boldsymbol{O}$ contain a finite number of elements which are discrete, the feature functions can be selected as Boolean functions [27]. Since the relative significance of each feature can be reflected by its corresponding weight factor, a unit valued feature function is sufficient to represent features. Hence, for convenience, we choose binary valued feature functions to model mode transition as given below [43]:

$$T_k(h_t, h_{t-1}) = \begin{cases} 1 & \text{if } h_{t-1} = i \quad \text{and} \quad h_t = j \\ 0 & \qquad \text{otherwise} \end{cases} \tag{4.3}$$

where $i, j = 1, 2, ..., N$ and $k = 1, 2, ..., K$. K indicates all possible scenarios of mode transitions, which is equal to N^2 in this case. The corresponding weight factor λ_k has the same role as the transition probability in HMM and needs to be estimated under the CRF framework. Each observation belongs to the finite observation set $\boldsymbol{\mathcal{B}}$, and the dependency of the observation sequence $\boldsymbol{Y}_t$ on the current mode h_t is formulated using the following feature function:

$$E_m(h_t, \boldsymbol{Y}_t) = E_m(h_t, O_t, O_{t-1}, ..., O_{t-d+1}) = \begin{bmatrix} E_{m_1}(h_t, O_t) \\ E_{m_2}(h_t, O_{t-1}) \\ ... \\ E_{m_d}(h_t, O_{t-d+1}) \end{bmatrix} \tag{4.4}$$

Here, $m = 1, 2, ..., M$, where M is the total number of feature functions that relate observations to the current operating mode. The elements in the above equation are all assumed to be indicator functions with the following form:

$$E_{m_l}(h_t, O_{t-l+1}) = \begin{cases} 1 & \text{if } h_t = i \text{ and } O_{t-l+1} \in \boldsymbol{\mathcal{B}} \\ 0 & \text{otherwise} \end{cases} \tag{4.5}$$

where $l = 1, 2, ..., d$, $i = 1, 2, ..., N$ and the corresponding weight factors $\mu_m = [\mu_{m_1}\ \mu_{m_2}\ ...\ \mu_{m_d}]$ form a vector, which has the same role as an emission probability matrix under the HMM scheme.

As a result, in the CRF model, the unknown parameters are the weight factors of all the feature functions, i.e., $\mathbf{\Lambda} = \{\lambda_k\}_{k=1}^{K}$ and $\boldsymbol{\mathcal{M}} = \{\mu_m\}_{m=1}^{M}$, which can be calculated using the conditional maximum likelihood estimation (CMLE). By setting the gradient of the conditional maximum likelihood function to zero, we obtain a set of coupled nonlinear equations. Since there are no explicit analytic solutions to this problem, numerical algorithms have been employed in literature for parameter estimation. In this work, the limited memory BFGS(L-BFGS) algorithm [105] has been used to solve the CMLE problem. After training, we employ the maximal posterior probability assessment [93] for process operating mode diagnosis.

4.3 Operating Mode Diagnosis Using Marginalized CRFs in the Presence of Missing Measurements

In this section, we propose a novel marginalized CRF framework for operating mode diagnosis in the presence of missing measurements. The modeling, parameter estimation, and related inference problems of this marginalized CRF are illustrated in detail in the following subsections.

4.3.1 Problem Formulation

In this case, the observation dataset $\boldsymbol{O}$ is assumed to be partially observed and can be partitioned as $\boldsymbol{O} = \{\boldsymbol{O}_{obs}, \boldsymbol{O}_{mis}\}$. Here, $\boldsymbol{O}_{obs}$ and $\boldsymbol{O}_{mis}$ represent the observed and missing components, respectively. For the missing component $\boldsymbol{O}_{mis}$, simply ignoring it will cause loss of information, and directly replacing it with a known value will cause bias in some cases [106]. As a result, we consider the marginalization over all missing components over all possible combinations of the missing measurements, to provide an overall estimation of current operating mode by accounting missing components. As given below, the conventional LCCRF model in Equation (4.1) is marginalized over the missing measurements:

$$P(\boldsymbol{h}|\boldsymbol{O}_{obs}) = \frac{\overbrace{\sum_{\boldsymbol{O}_{mis}} \exp \sum_{t=1}^{T} \{ \sum_{k=1}^{K} \lambda_k T_k(h_t, h_{t-1}) + \sum_{m=1}^{M} \mu_m E_m(h_t, \boldsymbol{Y}_t^{(obs)}, \boldsymbol{Y}_t^{(mis)}) \}}^{Z(\boldsymbol{h}, \boldsymbol{O}_{obs})}}{\underbrace{\sum_{\boldsymbol{h}'} \sum_{\boldsymbol{O}_{mis}} \exp \sum_{t=1}^{T} \{ \sum_{k=1}^{K} \lambda_k T_k(h_t^{'}, h_{t-1}^{'}) + \sum_{m=1}^{M} \mu_m E_m(h_t^{'}, \boldsymbol{Y}_t^{(obs)}, \boldsymbol{Y}_t^{(mis)}) \}}_{Z(\boldsymbol{O}_{obs})}} \tag{4.6}$$

which can be compactly represented as $P(\boldsymbol{h}|\boldsymbol{O}_{obs}) = \dfrac{Z(\boldsymbol{h}, \boldsymbol{O}_{obs})}{Z(\boldsymbol{O}_{obs})}$. Here, the notations $\boldsymbol{Y}_t^{(obs)}$ and $\boldsymbol{Y}_t^{(mis)}$ represent the observed and missing components in $\boldsymbol{Y}_t$, respectively.

Since the missing measurements occur randomly and have an impact on a certain range of operating modes, it is not possible to directly perform the marginalization of the Equation (4.6) by a simple local summation at each missing measurement instance. Hence, in order to solve this problem, some efficient algorithms need to be sought as explained below.

4.3.2 Parameter Estimation: A Maximum Likelihood Approach

Based on the modeling framework in Equation (4.6), in the model training stage, the unknown parameters will be estimated by maximizing the following log likelihood function, considering all possibilities of missing measurements:

$$l(\boldsymbol{\Theta}) = \log P(\boldsymbol{h}|\boldsymbol{O}_{obs}) - \frac{\|\boldsymbol{\Theta}\|_2^2}{2\sigma^2} = \log Z(\boldsymbol{h}, \boldsymbol{O}_{obs}) - \log Z(\boldsymbol{O}_{obs}) - \frac{\|\boldsymbol{\Theta}\|_2^2}{2\sigma^2} \tag{4.7}$$

where the notation $'\log'$ means the natural logarithm operation. By denoting the missing dataset as $\boldsymbol{O}_{mis} = [O_{m_1}, O_{m_2}, ..., O_{m_\alpha}]$, the first logarithmic term of Equation (4.7) can be factorized by individual missing measurements as shown in the following equation:

$$\begin{aligned}
\log Z(\boldsymbol{h}, \boldsymbol{O}_{obs}) =& \log \sum_{\boldsymbol{O}_{mis}} \exp \sum_{t=1}^{T} \{ \sum_{k=1}^{K} \lambda_k T_k(h_t, h_{t-1}) + \sum_{m=1}^{M} \mu_m E_m(h_t, \boldsymbol{Y}_t^{(obs)}, \boldsymbol{Y}_t^{(mis)}) \} \\
=& \log\{ \exp \sum_{t=1}^{T} \sum_{k=1}^{K} \lambda_k T_k(h_t, h_{t-1}) \cdot \exp \sum_{t=1}^{T} \sum_{m=1}^{M} \mu_m E_m(h_t, \boldsymbol{Y}_t^{(obs)}) \cdot \\
& \sum_{O_{m_1}} \exp \sum_{t=1}^{T} \sum_{m=1}^{M} \mu_m E_m(h_t, O_{m_1}) \cdots \sum_{O_{m_\alpha}} \exp \sum_{t=1}^{T} \sum_{m=1}^{M} \mu_m E_m(h_t, O_{m_\alpha}) \} \\
=& \sum_{t=1}^{T} \sum_{k=1}^{K} \lambda_k T_k(h_t, h_{t-1}) + \sum_{t=1}^{T} \sum_{m=1}^{M} \mu_m E_m(h_t, \boldsymbol{Y}_t^{(obs)}) + \sum_{i=1}^{\alpha} \log\{ \\
& \sum_{O_{m_i}} \exp \sum_{t=1}^{T} \sum_{m=1}^{M} \mu_m E_m(h_t, O_{m_i}) \}
\end{aligned} \tag{4.8}$$

Here, considering the observation sequence $\boldsymbol{Y}_t$, $\boldsymbol{Y}_t^{(obs)}$ and O_{m_i}, the feature function E_m is calculated as in Equation (4.4), by assigning the elements correlated with the observation as ones while treating the rest as zeros. For example, considering the observation sequence $\boldsymbol{Y}_t = \{O_t, O_{t-1}, O_{t-2}\}$, the feature functions $E_m(h_t, \boldsymbol{Y}_t)$ and $E_m(h_t, O_{t-1})$ can be calculated as follows:

$$
\begin{aligned}
E_m(h_t, \boldsymbol{Y}_t) = E_m(h_t, O_t, O_{t-1}, O_{t-2}) &= \begin{bmatrix} E_{m_1}(h_t, O_t) \\ E_{m_2}(h_t, O_{t-1}) \\ E_{m_3}(h_t, O_{t-2}) \end{bmatrix} \\
E_m(h_t, O_{t-1}) &= \begin{bmatrix} 0 \\ E_{m_2}(h_t, O_{t-1}) \\ 0 \end{bmatrix}
\end{aligned}
\tag{4.9}
$$

In Equation (4.8), for the terms with observed components, the solution is obtained in the same way as the regular CRF, where the term "regular CRF" represents the CRF framework derived with complete measurements. For the terms with missing components, all possible missing values are enumerated. The relative effect of different missing measurement combinations on the operating modes are determined through the corresponding weight factors.

The second logarithmic term $\log Z(\boldsymbol{O}_{obs})$ in Equation (4.7) is computed by the summation over both state sequence $\boldsymbol{h'}$ and missing measurements $\boldsymbol{O}_{mis}$ as given below:

$$
\log Z(\boldsymbol{O}_{obs}) = \log \sum_{h'_{1:T}} \sum_{\boldsymbol{O}_{mis}} \exp \sum_{t=1}^{T} \{ \sum_{k=1}^{K} \lambda_k T_k(h'_t, h'_{t-1}) + \sum_{m=1}^{M} \mu_m E_m(h'_t, \boldsymbol{Y}_t^{(obs)}, \boldsymbol{Y}_t^{(mis)}) \}
\tag{4.10}
$$

Due to the complex interplay between operating modes and missing measurements, Equation (4.10) needs to be solved in an efficient way. Based on the forward-backward algorithm, a propagation scheme is proposed and illustrated in detail in the next subsection.

After marginalization, the loss function $l(\boldsymbol{\Theta})$ may no longer be convex and therefore might lead to local optima, so the numerical optimization algorithms need to be employed for calculating the parameters. Below, we provide the gradients of $l(\boldsymbol{\Theta})$ with respect to the unknown parameters for calculating the parameter values:

$$
\frac{\partial l(\boldsymbol{\Theta})}{\partial \lambda_k} = \sum_{t=1}^{T} T_k(h_t, h_{t-1}) - \sum_{t=1}^{T} \sum_{h'_t, h'_{t-1}} P(h'_t, h'_{t-1} | \boldsymbol{O}_{obs}) T_k(h'_t, h'_{t-1}) - \frac{\lambda_k}{\sigma^2} = 0
\tag{4.11}
$$

$$
\begin{aligned}
\frac{\partial l(\boldsymbol{\Theta})}{\partial \mu_{m_l}} =& \sum_{t=1}^{T} E_{m_l}(h_t, O_{t-l+1}^{(obs)}) + \sum_{t=1}^{T} \sum_{O_{t-l+1}^{(mis)}} w(O_{t-l+1}^{(mis)}) E_{m_l}(h_t, O_{t-l+1}^{(mis)}) \\
&- \sum_{t=1}^{T} \sum_{h'_t} \sum_{O_{t-l+1}^{(mis)}} P(h'_t, O_{t-l+1}^{(mis)} | \boldsymbol{O}_{obs}) E_{m_l}(h'_t, O_{t-l+1}) - \frac{\mu_{m_l}}{\sigma^2} = 0
\end{aligned} \tag{4.12}
$$

where $O_{t-l+1}^{(obs)}$ and $O_{t-l+1}^{(mis)}$ represent the cases when the measurement O_{t-l+1} is observed or missed, respectively, and

$$
w(O_{mi}) = \frac{\exp \sum_{t=1}^{T} \sum_{m=1}^{M} \mu_m E_m(h_t, O_{m_i})}{\sum_{O_{m_i}} \exp \sum_{t=1}^{T} \sum_{m=1}^{M} \mu_m E_m(h_t, O_{m_i})} \tag{4.13}
$$

Similar to $\log Z(\boldsymbol{O}_{obs})$, the marginalized probabilities $P(h'_t, h'_{t-1} | \boldsymbol{O}_{obs})$ and $P(h'_t, O_{t-l+1}^{(mis)} | \boldsymbol{O}_{obs})$ in Equation (4.12) need to be calculated through propagations.

4.3.3 Inference of CRF with Missing Measurements

In this subsection, we solve two inference problems, namely, the marginal probability computation and the optimal mode estimation, while performing offline training and online validation, respectively. This is generally practiced using the forward-backward algorithm. However, due to the high correlations between the observations and operating modes, as well as the presence of missing measurements, the ordinary forward-backward propagation algorithm is unsuitable for our application. As a result, we propose a new propagation algorithm by suitably modifying the existing forward-backward algorithm to solve the inference problem in an efficient way. The details are illustrated in the following subsections.

4.3.3.1 Propagation Algorithm

The forward-backward algorithm is a dynamic programming algorithm which has important applications in both HMM and CRF problems [107]. It is used to simplify the enumeration over operating mode sequence for inference. However, in our case, not only the operating mode sequence, but also the missing measurements need to be enumerated, which causes correlated interactions between operating modes and missing measurements, as shown in Fig. 4.4. Therefore, in order to solve this problem, a new propagation algorithm is developed based on the forward-backward conception by performing a summation over the operating modes and missing measurements.

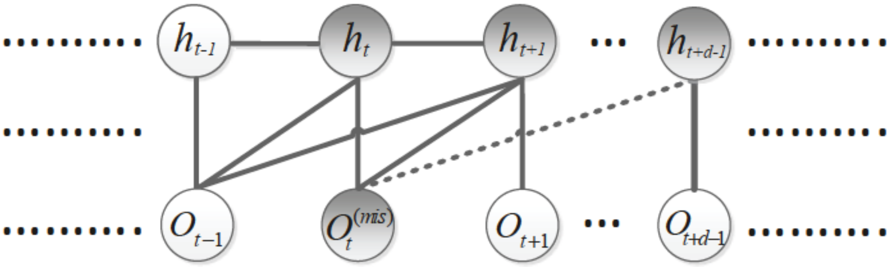


Figure 4.4: The figure illustrates the correlated interactions at time point t by the missing measurement O_t and the corresponding operating mode sequence. The shaded nodes denote the operating mode sequence which is affected by missing measurement O_t

First, for the sake of simplicity, we define the following intermediate terms for our propagation algorithm:

$$\begin{aligned} \varphi_t(h_t, h_{t-1}, \mathbf{Y}_t^{(obs)}) &\overset{def}{=} \exp\{\sum_{k=1}^{K} \lambda_k T_k(h_t, h_{t-1}) + \sum_{m=1}^{M} \mu_m E_m(h_t, \mathbf{Y}_t^{(obs)})\} \\ \varphi_t(h_t, O_{t-l}) &\overset{def}{=} \exp\{\sum_{m=1}^{M} \mu_m E_m(h_t, O_{t-l})\} \end{aligned} \tag{4.14}$$

In order to count the effect of a missing measurement on as many as d adjacent operating mode labels, we formulate an intermediate variable below:

$$\gamma_t(h_t, h_{t+1}, ..., h_{t+d-1}, O_t^{(mis)}) \overset{def}{=} \begin{cases} \prod_{j=0}^{d-1} \varphi_{t+j}(h_{t+j}, O_t^{(mis)}) & \text{if } O_t \text{ is missing} \\ 1 & \text{otherwise} \end{cases} \tag{4.15}$$

where $O_t^{(mis)}$ indicates the missing measurements at time point t. In this formulation, the contribution of an individual missing value to process operating mode transitions is modeled. By integrating all the possible values of $O_t^{(mis)}$, the effect of $O_t^{(mis)}$ can be transferred into the transitions among $h_{t:t-d+1}$. Hereafter, we define a variable η by integrating Equation (4.15) as follows:

$$\eta_t(h_t, h_{t+1}, ..., h_{t+d-1}) = \begin{cases} \sum_{O_t^{(mis)}} \gamma_t(h_t, h_{t+1}, ..., h_{t+d-1}, O_t^{(mis)}) & \text{if } O_t \text{ is missing} \\ 1 & \text{otherwise} \end{cases} \tag{4.16}$$

However, in reality, since a missing measurement generally affects more on the operating modes closer to it, for computational tractability, one can always approximate Equation (4.16) by choosing the length of the missing measurement to be d_s, where $d_s < d$.

Based on the definitions listed above, a set of intermediate variables α_t is defined by considering the enumerations over operating mode sequence $h_{1:t-1}$ and all the missing measurements by time point t for forward propagation:

$$\alpha_t(h_t, \boldsymbol{h}_{t,mis}^{(f)}) \overset{def}{=} \sum_{h_{1:t-1}} \sum_{O_{1:t}^{(mis)}} \prod_{t'=1}^{t} \varphi_{t'}(h_{t'}, h_{t'-1}, \boldsymbol{Y}_{t'}^{(obs)}) \cdot \gamma_{t'}(h_{t'}, h_{t'+1}, ..., h_{t'+d-1}, O_{t'}^{(mis)}) \tag{4.17}$$

where $\boldsymbol{h}_{t,mis}^{(f)}$ represents the operating mode sequence impacted by the missing measurements before time point t. If there are no missing measurements from O_{t-d+2} to O_t, then $\boldsymbol{h}_{t,mis}^{(f)}$ will be $\varnothing$. If O_t is missing, then $O_t^{(mis)}$ will denote the missing variable; otherwise O_t will be included in $\boldsymbol{Y}_t^{(obs)}$ and its corresponding γ_t term will be computed as in Equation (4.15).

As a result, α_t can be calculated iteratively as below:

$$\alpha_{t+1}(h_{t+1}, \boldsymbol{h}_{t+1,mis}^{(f)}) = \sum_{h_t} \varphi_{t+1}(h_{t+1}, h_t, \boldsymbol{Y}_{t+1}^{(obs)}) \eta_{t+1}(h_{t+1}, ..., h_{t+d}) \cdot \alpha_t(h_t, \boldsymbol{h}_{t,mis}^{(f)}) \tag{4.18}$$

Similarly, for the propagation from backward direction, a set of intermediate variables β_t

is defined as below:

$$\beta_t(h_{t+d-2}, \boldsymbol{h}_{t,mis}^{(b)}) \overset{def}{=} \sum_{O_{t:T}^{(mis)}} \sum_{h_{t+d-1:T}} \prod_{t'=t}^{T} \gamma_{t'}(h_{t'}, ..., h_{t'+d-1}, O_{t'}^{(mis)}) \prod_{t'=t+d-1}^{T} \varphi_{t'}(h_{t'}, h_{t'-1}, \boldsymbol{Y}_{t'}^{(obs)}) \tag{4.19}$$

where $\boldsymbol{h}_{t,mis}^{(b)}$ represents the operating mode sequence affected by the potential missing measurements before time point t, which is the subset of $h_{t:t+d-3}$. If there is no missing data from O_t to O_{t+d-3}, then $\boldsymbol{h}_{t,mis}^{(b)}$ will be $\varnothing$. By means of recursion, the intermediate variables $\boldsymbol{\beta}$ for backward propagation can be calculated iteratively as follows:

$$\begin{aligned}\beta_{t-1}(h_{t+d-3}, \boldsymbol{h}_{t-1,mis}^{(b)}) = \sum_{h_{t+d-2}} & \varphi_{t+d-2}(h_{t+d-2}, h_{t+d-3}, \boldsymbol{Y}_{t+d-2}^{(obs)}) \\ & \eta_{t-1}(h_{t-1}, ..., h_{t+d-2})\beta_t(h_{t+d-2}, \boldsymbol{h}_{t,mis}^{(b)})\end{aligned} \tag{4.20}$$

The steps detailing the proposed forward-backward propagation algorithm can be found in B.1. After the forward propagation, the normalization term $Z(\boldsymbol{O}_{obs})$ in Equation (4.10) can be computed by means of the derived results:

$$Z(\boldsymbol{O}_{obs}) = \sum_{h_{1:T}} \sum_{O_{1:T}^{(mis)}} \prod_{t=1}^{T} \varphi_t(h_t, h_{t-1}, \boldsymbol{Y}_t^{(obs)}) \cdot \gamma_t(h_t, ..., h_{t+d-1}, O_t^{(mis)}) = \sum_{h_T} \alpha_T(h_T) \tag{4.21}$$

The marginal probabilities in equations (4.11) and (4.12) can be derived by the intermediate terms $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, which will be illustrated in the next subsection.

4.3.3.2 Marginal Probability Derivation

In Equations (4.11) - (4.12), the gradient calculation is based on the marginal distributions $P(h_t, h_{t-1}|\boldsymbol{O}_{obs})$ and $P(h_t, O_{t-l+1}^{(mis)}|\boldsymbol{O}_{obs})$. Using the definition of $\varphi_t(h_t, h_{t-1}, \boldsymbol{Y}_t^{(obs)})$, we marginalize the relevant terms to obtain $P(h_t, h_{t-1}|\boldsymbol{O}_{obs})$ as follows:

$$\begin{aligned}
P(h_t, h_{t-1}|\boldsymbol{O}_{obs}) =& \frac{1}{Z(\boldsymbol{O}_{obs})} \sum_{h_{1:t-2}, h_{t+1:T}} \sum_{O_{1:T}^{(mis)}} \prod_{t'=1}^{T} \varphi_{t'}(h_{t'}, h_{t'-1}, \boldsymbol{Y}_{t'}^{(obs)}) \gamma_{t'}(h_{t'}, ..., h_{t'+d-1}, O_{t'}^{(mis)}) \\
=& \frac{1}{Z(\boldsymbol{O}_{obs})} \varphi_t(h_t, h_{t-1}, \boldsymbol{Y}_t^{(obs)}) \sum_{h_{t+1:t+d-2}} \{ \sum_{h_{1:t-2}} \sum_{O_{1:t-1}^{(mis)}} \prod_{t'=1}^{t-1} \varphi_{t'}(h_{t'}, h_{t'-1}, \boldsymbol{Y}_{t'}^{(obs)}) \\
& \gamma_{t'}(h_{t'}, ..., h_{t'+d-1}, O_{t'}^{(mis)}) \} \{ \sum_{h_{t+d-1:T}} \sum_{O_{t:T}^{(mis)}} \prod_{t'=t+1}^{T} \gamma_{t'}(h_{t'}, ..., h_{t'+d-1}, O_{t'}^{(mis)}) \\
& \prod_{t'=t+d-1}^{T} \varphi_{t'}(h_{t'}, h_{t'-1}, \boldsymbol{Y}_{t'}^{(obs)}) \} \prod_{t'=t+1}^{t+d-2} \varphi_{t'}(h_{t'}, h_{t'-1}, \boldsymbol{Y}_{t'}^{(obs)}) \\
=& \frac{1}{Z(\boldsymbol{O}_{obs})} \varphi_t(h_t, h_{t-1}, \boldsymbol{Y}_t^{(obs)}) \sum_{h_{t+1:t+d-2}} \alpha_{t-1}(h_{t-1}, \boldsymbol{h}_{t-1,mis}^{(f)}) \beta_t(h_{t+d-2}, \boldsymbol{h}_{t,mis}^{(b)}) \\
& \prod_{t'=t+1}^{t+d-2} \varphi_{t'}(h_{t'}, h_{t'-1}, \boldsymbol{Y}_{t'}^{(obs)})
\end{aligned} \tag{4.22}$$

where the operating mode sequences $\boldsymbol{h}_{t-1,mis}^{(f)}$ and $\boldsymbol{h}_{t,mis}^{(b)}$ are the subsets of $\{h_{t+1}, \cdots, h_{t+d-2}\}$, and the summation over $h_{t+1:t+d-2}$ is calculated by similar forward propagation as illustrated in the previous subsection.

When O_{t-l+1} is missing, by defining $\alpha_t(h_t, \boldsymbol{h}_{t,mis}^{(f)}, O_{t-l+1}^{(mis)})$ as below, the marginal probability $P(h_t, O_{t-l+1}^{(mis)}|\boldsymbol{O}_{obs})$ will be obtained as shown in Equation (4.24).

$$\begin{aligned}
\alpha_t(h_t, \boldsymbol{h}_{t,mis}^{(f)}, O_{t-l+1}^{(mis)}) \overset{def}{=} \sum_{h_{1:t-1}} \sum_{O_{1:t}^{(mis)} \backslash O_{t-l+1}^{(mis)}} \prod_{t'=1}^{t} \varphi_{t'}(h_{t'}, h_{t'-1}, \boldsymbol{Y}_{t'}^{(obs)}) \cdot \\
\gamma_{t'}(h_{t'}, h_{t'+1}, ..., h_{t'+d-1}, O_{t'}^{(mis)})
\end{aligned} \tag{4.23}$$

$$
\begin{aligned}
P(h_t, O^{(mis)}_{t-l+1} | \boldsymbol{O}_{obs}) =& \frac{1}{Z(\boldsymbol{O}_{obs})} \sum_{h_{1:t-1}, h_{t+1:T}} \sum_{O^{(mis)}_{1:T} \backslash O^{(mis)}_{t-l+1}} \prod_{t'=1}^{T} \varphi_{t'}(h_{t'}, h_{t'-1}, \boldsymbol{Y}^{(obs)}_{t'}) \\
& \gamma_{t'}(h_{t'}, ..., h_{t'+d-1}, O^{(mis)}_{t'}) \\
=& \frac{1}{Z(\boldsymbol{O}_{obs})} \sum_{h_{t+1:t+d-1}} \{ \sum_{h_{1:t-1}} \sum_{O^{(mis)}_{1:t} \backslash O^{(mis)}_{t-l+1}} \prod_{t'=1}^{t} \varphi_{t'}(h_{t'}, h_{t'-1}, \boldsymbol{Y}^{(obs)}_{t'}) \\
& \gamma_{t'}(h_{t'}, ..., h_{t'+d-1}, O^{(mis)}_{t'}) \} \{ \sum_{O^{(mis)}_{t+1:T}} \sum_{h_{t+d:T}} \prod_{t'=t+1}^{T} \gamma_{t'}(h_{t'}, ..., h_{t'+d-1}, O^{(mis)}_{t'}) \\
& \prod_{t'=t+d}^{T} \varphi_{t'}(h_{t'}, h_{t'-1}, \boldsymbol{Y}^{(obs)}_{t'}) \} \prod_{t'=t+1}^{t+d-1} \varphi_{t'}(h_{t'}, h_{t'-1}, \boldsymbol{Y}^{(obs)}_{t'}) \\
=& \frac{1}{Z(\boldsymbol{O}_{obs})} \sum_{h_{t+1:t+d-1}} \alpha_t(h_t, \boldsymbol{h}^{(f)}_{t,mis}, O^{(mis)}_{t-l+1}) \cdot \beta_{t+1}(h_{t+d-1}, \boldsymbol{h}^{(b)}_{t+1,mis}) \\
& \prod_{t'=t+1}^{t+d-1} \varphi_{t'}(h_{t'}, h_{t'-1}, \boldsymbol{Y}^{(obs)}_{t'})
\end{aligned}
\tag{4.24}
$$

Here, the operating mode sequences $\boldsymbol{h}^{(f)}_{t-1,mis}$ and $\boldsymbol{h}^{(b)}_{t+1,mis}$ are also the subsets of $\{h_{t+1}, ..., h_{t+d-1}\}$, and the summation over $h_{t+1:t+d-1}$ is computed iteratively.

4.3.3.3 Online Operating Mode Diagnosis

Once the offline parameter estimation step is completed, in order to identify the optimal operating mode sequence online, the probability of the current operating mode given all the previous observations is calculated as:

$$
h^*_t = \operatorname*{argmax}_{h_t} P(h_t | O^{(obs)}_1, ..., O^{(obs)}_t) \tag{4.25}
$$

Solving the above problem is equivalent to a CRF based filtering problem for selecting optimal operating modes. In order to estimate the current optimal operating mode in the presence of previously missing measurements, we define a set of intermediate variables ξ_t as follows:

$$\xi_t(h_t) \overset{def}{=} \sum_{h_{t-d+1:t-1}} \sum_{O^{(mis)}_{t-d+2:t}} \prod_{t'=t-d+2}^{t} \varphi_{t'}(h_{t'}, h_{t'-1}, \boldsymbol{Y}^{(obs)}_{t'}) \cdot \gamma_{t'}(h_{t'}, ..., h_t, O^{(mis)}_{t'}) \cdot \alpha_{t-d+1}(h_{t-d+1}, \boldsymbol{h}^{(f)}_{t-d+1,mis}) \tag{4.26}$$

Here, the forward propagation structure can be employed to deal with the d step ahead calculation from α_{t-d+1} to ξ_t, and the conditional probability $P(h_t|O^{(obs)}_1, ..., O^{(obs)}_t)$ in Equation (4.25) is derived by means of the intermediate variables ξ_t, as below:

$$P(h_t|O^{(obs)}_1, ..., O^{(obs)}_t) = \frac{\xi_t(h_t)}{\sum_{h_t} \xi_t(h_t)} \tag{4.27}$$

As a summary, the main components of this proposed marginalized CRFs include offline parameter estimation, related inference problems and the proposed forward-backward propagation. The pseudocodes for the same are presented in B.2.

4.4 Case Studies

In this section, the proposed approach for operating mode diagnosis is tested on two case studies, (i) CSTR simulation system, and (ii) hybrid tank experimental system. We consider both the complete observations and missing measurement scenarios for evaluating the process monitoring performance on these two systems.

4.4.1 Continuous Stirred Tank Reactor System

4.4.1.1 Process Description

In this study, the CSTR system containing two reactors in series [108] is considered and the schematic of the system is illustrated in Fig. 4.5. This process comprises of two irreversible exothermic reactions occurring in both tanks simultaneously, and product of the first tank acts as the feed to the second reactor. The coolant q_c flows through both reactors and can be treated as the input of the whole system, and the feed flow-rate q_f is the disturbance. During the simulation, the CSTR operates under the closed-loop condition and the control objective is to keep the effluent concentration C_{A2} at a certain reference value by manipulating the coolant flow-rate q_c consistently. In order to obtain an instant and direct operating mode diagnosis result, the product concentration C_{A2} and temperature T_2 in the second reactor are chosen as observation variables for process operating mode diagnosis.

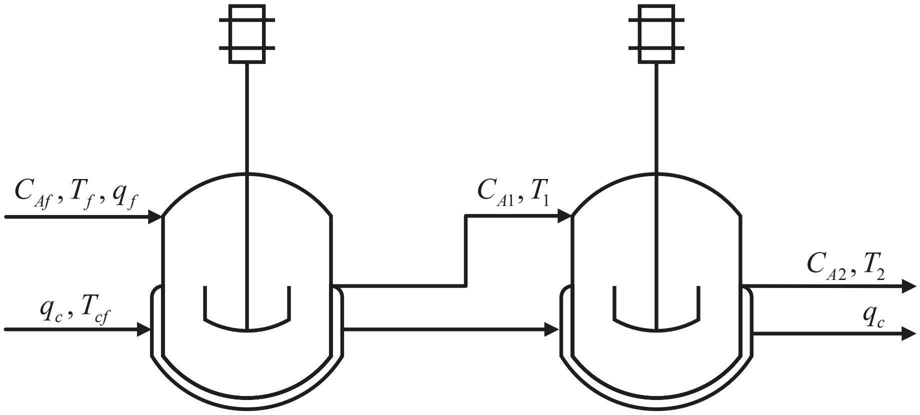


Figure 4.5: The schematic of CSTR in series [108]

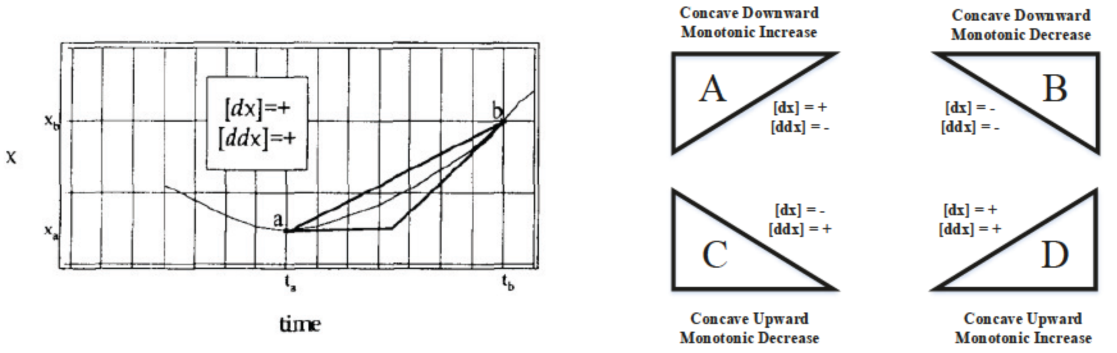
For the CSTR system, the setpoint of C_{A2} is set as $0.0075mol/L$ and a PI controller is designed for closed loop control with parameters $K_c = 350L^2/mol \cdot min$ and $\tau_I = 0.25min$ as shown in [109]. For other process parameters, the readers are referred to the above mentioned reference. Here, the feed flow-rate q_f is contaminated by a white noise disturbance with variance 0.3 under normal operating condition. For simulating the abnormal operating scenarios, a random impulse signal with mean $10L/min$ and variance 3, and a random ramp disturbance with maximum output value $2.5L/min$ and variance 1, are introduced in the simulation.

After data collection, some feature extraction algorithms are employed for data preprocessing, and the proposed algorithm is then used to conduct model training and online process operating mode diagnosis.

4.4.1.2 Discrete Feature Extraction

For many practical applications, in order to make the patterns more recognizable as well as more robust in the presence of noises, the original observed variables are generally preprocessed and transformed into new features, which is called feature extraction [110]. In this case, the wavelet analysis technique [111] is first used to reduce the impact of noise, and then triangular representation is employed to convert the continuous signal to a discrete symbolic sequence based on a finite number of triangles, which captures the trend of observations and also reduces the sensitivity to noise [88]. Details of triangular representation can be found elsewhere [112] and is omitted here for brevity. Further, it brings in the robustness to the proposed algorithm against missing measurements, since a lower percentage of missing measurements may not impact the overall process trend.

After triangular representation, the original continuous signal is represented by a discrete sequence with finite possible values. The discretization rule used in this chapter can be found in Fig. 4.6 and Table 4.1. Here, the whole duration or magnitude range is equally divided into three segments, and the notations “small”, “medium” and “large” indicate the data falling into the segments with lowest, middle and highest ranges, respectively. Based on the continuous dataset illustrated in Fig. 4.7, a segment of the corresponding triangular sequences of the observations are shown in Fig. 4.8. In Fig. 4.8, the three operating modes, namely, $Normal$, *Abnormal* 1 (impulse) and *Abnormal* 2 (ramp), are converted into triangular sequences and are demonstrated in black, red and blue colors, respectively. According to the discretization rule, different operating modes will have different triangle patterns; for example, in the normal data range, the number of triangles with indices corresponding to small duration and small magnitude is more than those of the other two modes.



(a) An illustration example for the triangular episode based process trend description.

(b) Four primitive types of triangles employed in this paper.

Figure 4.6: The illustration of triangular representation [113]

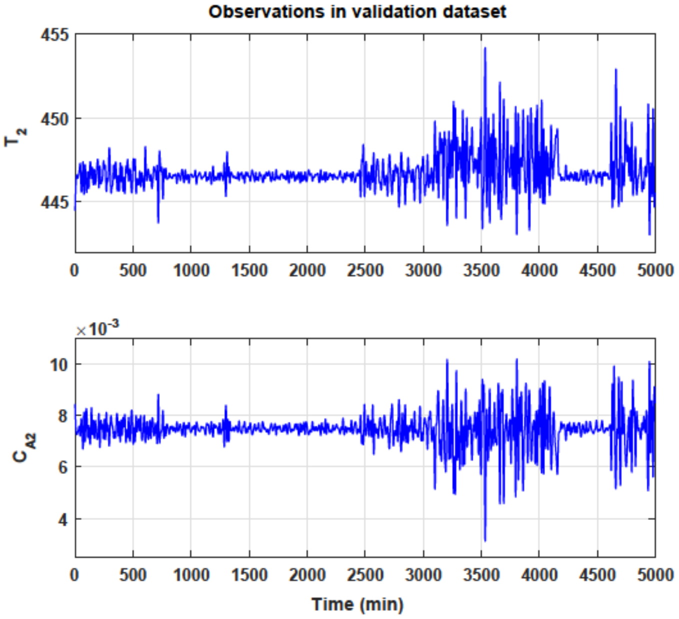


Figure 4.7: The observation changing trends for different operating modes of closed-loop CSTR system in the validation dataset. The upper and lower subfigures indicate the temperature and concentration of the second tank, respectively

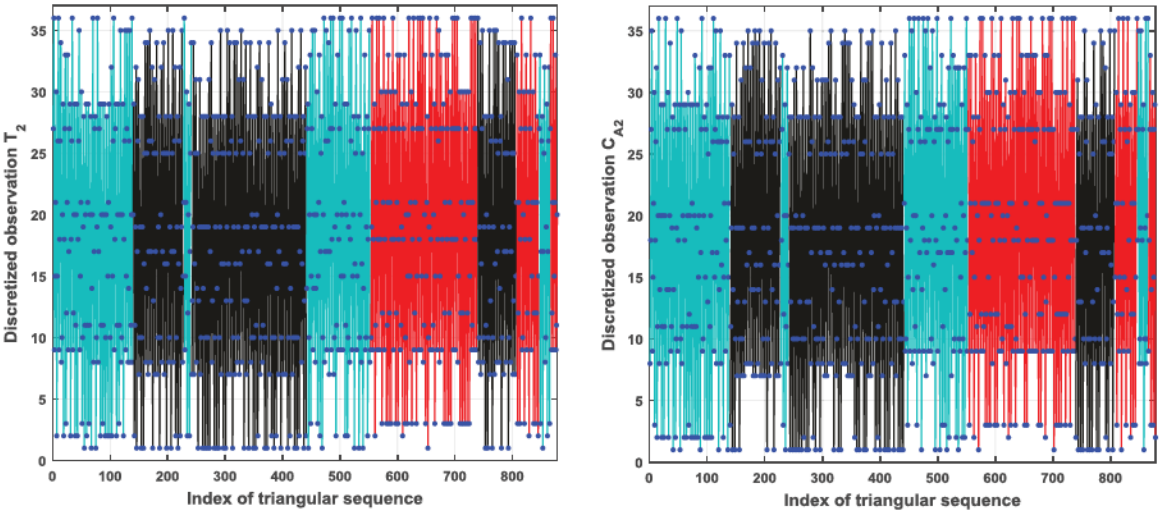


Figure 4.8: The triangular discretization result of the CSTR validation dataset. Temperature and product concentration of the second tank have been discretized in the left and right figures, respectively. The black, red and blue lines indicate the normal, impulse and ramp abnormal cases, respectively

4.4.1.3 Process Operating Mode Diagnosis Performance

In this simulation study, since the CSTR operates under the closed-loop condition and the observations generally exhibit high and complicated correlations, the conditional independence assumptions of HMMs will not be the best fit. For a comparison, two other process operating mode diagnosis algorithms are also considered in this work, i.e., HMM and back propagation neural network (BPNN). For the BPNN algorithm, an HMM is first trained for each observation variable, and then the probabilities of different operating modes given different observations are employed as input to the BPNN for the overall operating mode estimation. For the validation dataset illustrated in Fig. 4.8, the probability estimations of different operating modes provided by three algorithms and the online monitoring performances are shown in Fig. 4.9.

In the CRF diagnosis case, the window length d is selected as two. In the BPNN case, by parameter tuning, a hidden layer with ten nodes is designed for an optimal overall operating

Table 4.1: Look-up table of triangular discretization

Triangle shape ***Triangle size***		Concave downward		Concave upward	
		Monotonic increase	Monotonic decrease	Monotonic decrease	Monotonic increase
Small duration	Small magnitude	1	10	19	28
	Medium magnitude	2	11	20	29
	Large magnitude	3	12	21	30
Medium duration	Small magnitude	4	13	22	31
	Medium magnitude	5	14	23	32
	Large magnitude	6	15	24	33
Large duration	Small magnitude	7	16	25	34
	Medium magnitude	8	17	26	35
	Large magnitude	9	18	27	36

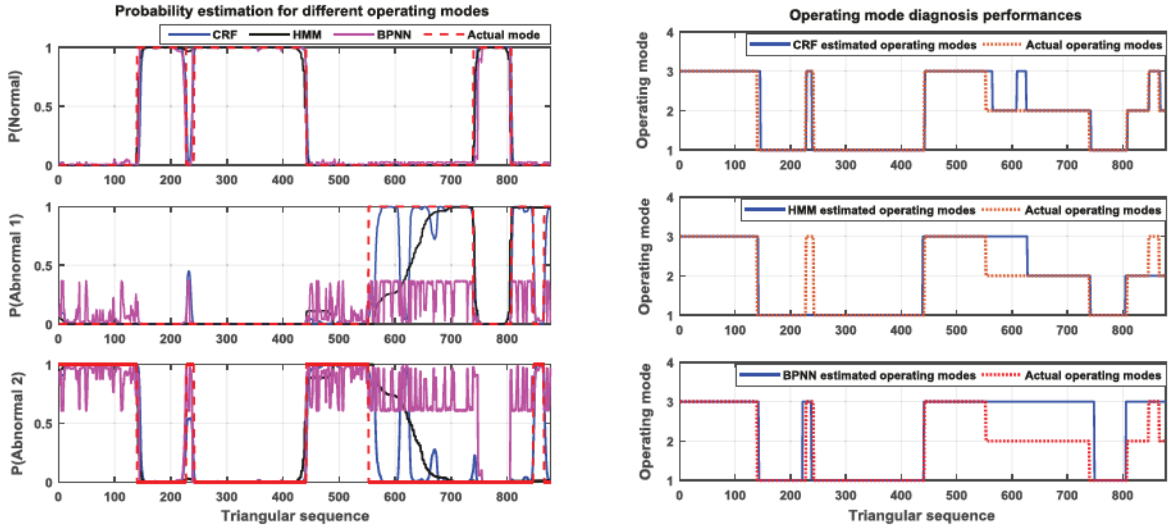


Figure 4.9: Probability of CSTR process operating modes estimated by the CRF (d = 2), HMM and BPNN based algorithms and the corresponding operating mode diagnosis performances. The operating mode numbers 1, 2 and 3 represent normal, impulse and ramp disturbance contaminated cases, respectively

mode decision. By modeling the correlations among the observations, we observe that the CRF gives the highest diagnosis accuracy, and the HMM and BPNN based monitoring algorithms have a lower diagnosis accuracy when the two abnormal operating modes occur sequentially. Moreover, during the time of the short acting ramp disturbance around $t = 840s$, the CRF provides an exact detection, while the other two algorithms are not able to detect the corresponding fault. Referring to the operating mode probability estimation result, the BPNN based algorithm always provides ambiguous estimations around 0.5, which is not reliable compared to the other algorithms.

Additionally, due to the convexity of the loss function in the CRF framework, the initial value of CRF does not impact the training performance as much as the other two algorithms, which is another important advantage of CRF. We also define a metric called diagnosis accuracy, which represents the ratio of correct estimation over all the estimations, and the comparison results using the same metric can be found in Table 4.2.

Table 4.2: CSTR process operating mode diagnosis accuracy with complete dataset

Algorithm	CRF (d=4)	CRF (d=3)	CRF (d=2)	CRF (d=1)	HMM	BPNN
Diagnosis accuracy (%)	94.75	94.30	**95.21**	92.82	86.30	70.62

Due to various practical reasons, missing data is common during an industrial process operation. In order to validate the performance of the algorithm under missing data scenarios, in this simulation, we consider 12 percent of process measurements as missing randomly in the training dataset, and the corresponding discrete representation triangles are also considered to be missing if more than 20 percent of the data points are lost within one triangle. Based on the incomplete training dataset, the marginalized CRF is used for operating mode diagnosis. In order to assess the marginalization performance, the regular CRF is trained with the same training dataset by simply ignoring the missing measurements and is used as the initial guess of the marginalized CRF. The comparison result can be found in Fig. 4.10, which shows that the marginalized CRF exhibits better performance than the regular version.

Additionally, to deal with the missing measurements, the marginalized HMM [114] is employed for operating mode diagnosis. Similar to the complete data case study, the BPNN is also combined with the marginalized HMM for comparison purposes. The comparison results of marginalized CRF, HMM and BPNN methods can be found in Fig. 4.11. Evidently, the performance of the marginalized CRF is much better than the other two. By increasing the percentage of missing measurements, the robustness of all the algorithms to missing measurements is tested. The corresponding diagnosis results are presented in Table 4.3,

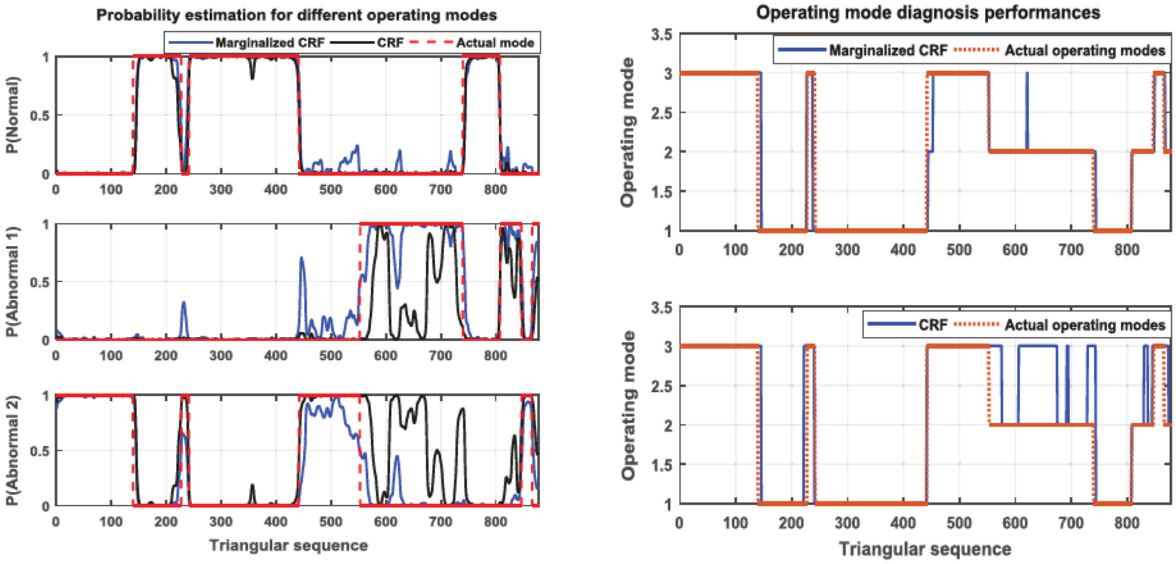


Figure 4.10: The operating mode diagnosis performance comparison between marginalized and regular CRFs in the presence of 12% missing measurements. The operating mode numbers 1, 2 and 3 represent normal, impulse and ramp disturbance contaminated cases, respectively

which indicates that the marginalized CRF shows not only the robustness to the increasing percentage of missing data, but also provides the most accurate result compared with the other two algorithms.

Table 4.3: CSTR process operating mode diagnosis accuracy in the presence of different missing percentages

Data missing percentage	5 %	10 %	15 %
Marginalized CRF diagnosis accuracy (%)	93.16	92.36	91.79
Marginalized HMM diagnosis accuracy (%)	79.79	79.11	76.37
Marginalized BPNN diagnosis accuracy (%)	76.71	75.34	72.95

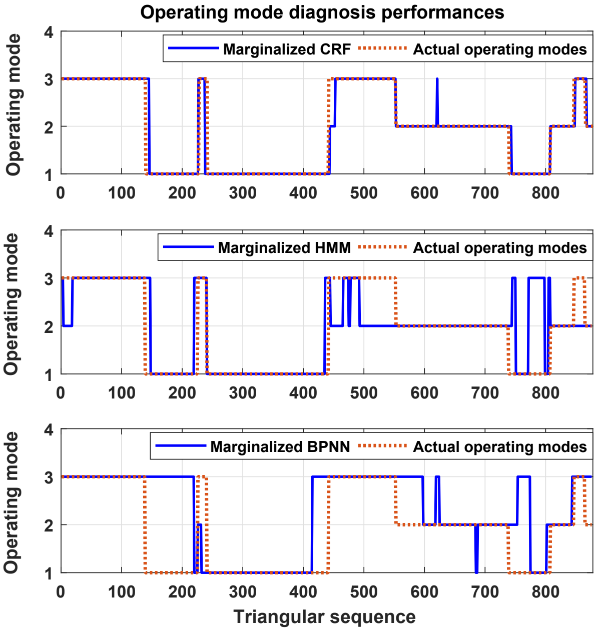


Figure 4.11: The operating mode diagnosis performances compared among marginalized CRF, HMM and BPNN in the presence of 12% missing measurements. The operating mode numbers 1, 2 and 3 represent normal, impulse and ramp disturbance contaminated cases, respectively. The operating mode diagnosis accuracies of the marginalized CRF, HMM and BPNN approaches in this particular case are 96.47%, 79.45% and 73.29%, correspondingly

4.4.2 An Experimental Validation on Hybrid Tank System

4.4.2.1 Process Description

In order to validate practicality of the proposed algorithm on real data, a hybrid tank experimental system is considered. The corresponding equipment schematic is presented in Fig. 4.12. As shown in Fig. 4.12, the hybrid tank system is composed of three connected horizontal tanks and the inlet water flow is pumped into the two tanks on both sides. Manipulation of the on-off valves V_1, V_2, V_3, V_4 will increase the tank levels suddenly and cause overflow. Consequently, closure of the two lower connecting valves V_3 and V_4 will lead to two abnormal cases. In this experimental validation, when all of the connecting valves V_1, V_2, V_3, V_4 are open, it is considered as *Normal* operating mode, while when V_3 or V_4 gets closed, the system is considered to be operating under two different abnormal modes, named

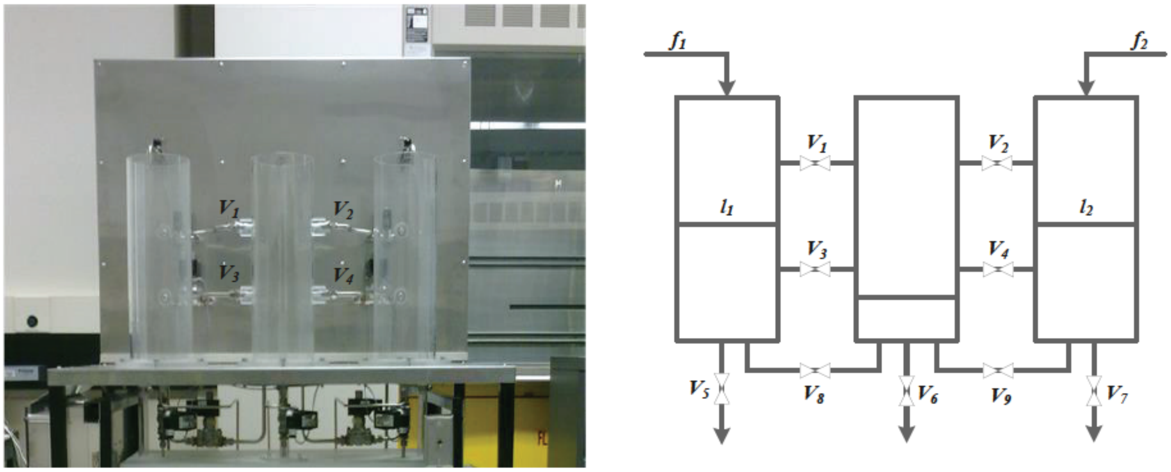


Figure 4.12: The configuration and diagram of the experimental hybrid tank system

Abnormal 1 and *Abnormal* 2. To reflect the real-time operating conditions and provide a reliable diagnosis result, the left and right tank levels, l_1 and l_2, respectively, are chosen as monitoring observations for operating mode diagnosis. Valves V_1, V_2 and $V_5 - V_9$ are kept open during the entire experiment. We used 80% of the data collected from the experiment for training and the remaining for validation.

4.4.2.2 Discrete Feature Extraction

Similar to the CSTR process described in the previous subsection, the triangular discretization algorithm is used for the generation of discrete features in this experiment study. Once one of V_3 and V_4 is closed, the tank level increases faster, changing the normal operating conditions. As a result, the data trends are different between normal and abnormal situations. Fig. 4.13 shows the triangular representation results for both l_1 and l_2 under different operating modes.

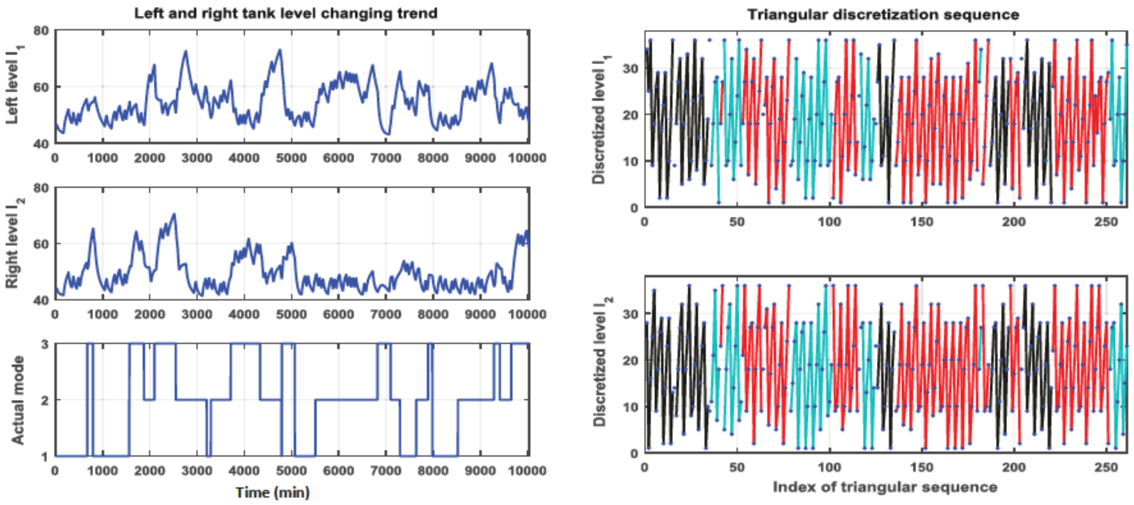


Figure 4.13: The triangular discretization results of observations l_1 and l_2. The left and right figures illustrate the continuous and discretized results of tank levels l_1 and l_2, respectively. In the left figure, the actual operating mode sequence can be found in the third subfigure, where number 1 to 3 denote the $Normal$, *Abnormal* 1 and *Abnormal* 2 modes, separately. In the right figure, the black, red and blue lines correspond to the $Normal$, *Abnormal* 1 and *Abnormal* 2 cases, respectively

4.4.2.3 Process Operating Mode Diagnosis Performance

For operating mode diagnosis, the remaining 20 percent of data excluded from model training is used as the validation dataset to test the diagnosis performance. The comparison results of operating mode diagnosis in Fig. 4.14 successfully demonstrate that the CRF based algorithm shows an obvious advantage over the HMM and BPNN ones. Similar to the simulated CSTR case, the HMM and BPNN based algorithms incorrectly diagnose two sequential abnormal modes, especially when the two abnormal modes share certain common features. For example, consider the last mode switching that occurs around $t = 9600$ *min*, where a large jump of l_2 indicates the start of mode *Abnormal* 2, while the steady states of modes *Abnormal* 1 and 2 also have some similar features. By changing the window length d of the CRF, the CRF based algorithm exhibits different diagnosis accuracy levels as shown in Table 4.4. From the numerical results listed, it is seen that the window length d should

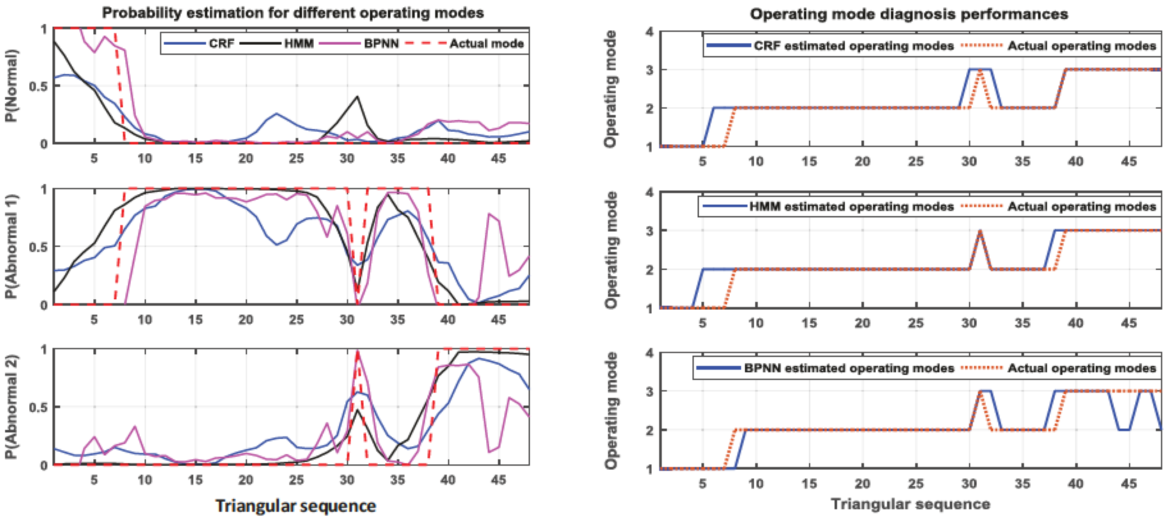


Figure 4.14: Probability of hybrid tank system operating modes estimated by the CRF (d = 10), HMM and BPNN algorithms and the corresponding operating mode diagnosis performances. Here, the operating mode numbers 1, 2 and 3 indicate the *Normal*, *Abnormal* 1 and *Abnormal* 2 operating modes, respectively

be selected properly for satisfactory diagnosis accuracy. A smaller d might not be able to fully describe the autocorrelation among observations, while a larger d might incorporate too much past information which has no impact on the current process but increases the computation. Consequently, the moving window length d should be in a reasonable range for good diagnosis performance.

Table 4.4: Hybrid tank system operating mode diagnosis accuracy with complete dataset

Algorithm	CRF (d=12)	CRF (d=10)	CRF (d=3)	CRF (d=1)	HMM	BPNN
Diagnosis accuracy (%)	75.00	**95.83**	85.42	81.25	91.67	87.50

In order to validate the proposed algorithm in the missing measurement scenario, it is assumed that 12 percent of data is missing at random in both of the tank level measurements l_1 and l_2 from the training dataset. Based on the same training and validation datasets, the proposed algorithm is also used for diagnosis performance evaluation and compared with the regular CRF, marginalized HMM and BPNN algorithms. The online monitoring results are presented in Fig. 4.15 and 4.16. It can be seen that compared to the other algorithms, the proposed marginalized CRF demonstrates more accurate diagnosis performance.

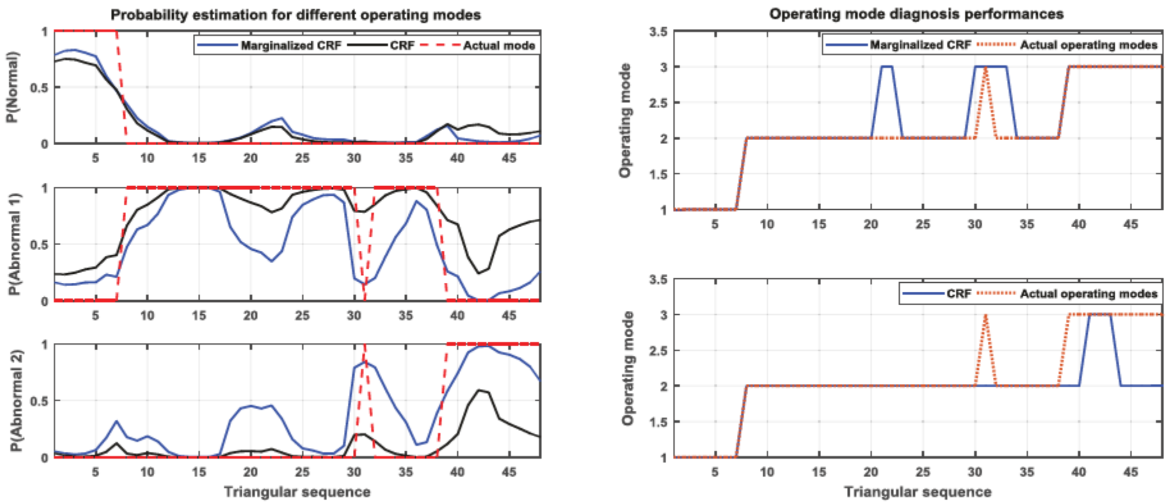


Figure 4.15: The hybrid tank system operating mode diagnosis performance comparison between marginalized and regular CRFs with 12% missing measurements. Here, the operating mode numbers 1, 2 and 3 indicate the *Normal*, *Abnormal* 1 and *Abnormal* 2 operating modes, respectively

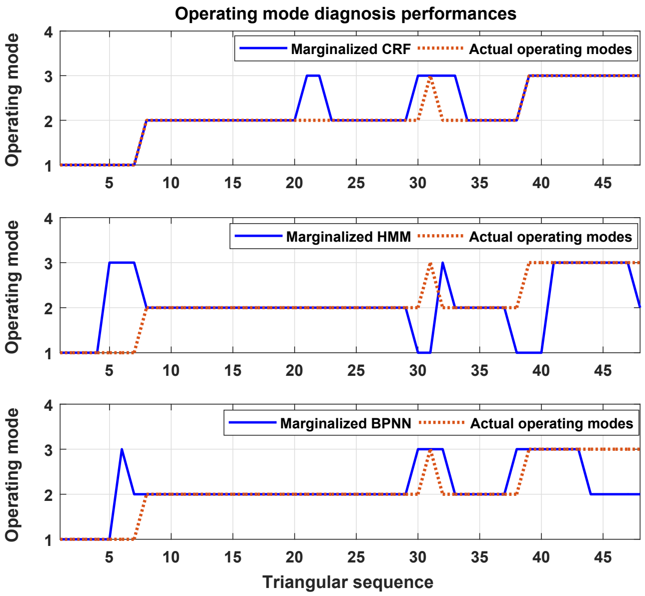


Figure 4.16: The hybrid tank system operating mode diagnosis performances compared among marginalized CRF, HMM and BPNN with 12% missing measurements. Here, the operating mode numbers 1, 2 and 3 indicate the *Normal*, *Abnormal* 1 and *Abnormal* 2 operating modes, respectively. The operating mode diagnosis accuracies of the marginalized CRF, HMM and BPNN approaches in this particular case are 89.58%, 79.17% and 79.17%, correspondingly

Table 4.5: Hybrid tank system operating mode diagnosis accuracy in the presence of different missing percentages

Data missing percentage	5 %	10 %	15 %
Marginalized CRF diagnosis accuracy (%)	93.75	85.42	83.33
Marginalized HMM diagnosis accuracy (%)	87.50	85.33	77.19
Marginalized BPNN diagnosis accuracy (%)	85.42	70.83	68.75

4.5 Conclusions

In this chapter, the CRF framework, a probabilistic discriminative model, was employed for process operating mode diagnosis problem. Due to the relaxation from the inherent assumptions of HMM, the CRF algorithm has shown to be more effective in describing complex autocorrelations among the observations, thereby more accurately estimating the current operating modes. Moreover, to deal with the missing measurement problem, a new marginalized CRF framework has been designed and corresponding inference algorithms are also developed. Finally, the proposed approach has been tested by performing a simulation study on the CSTR system and an experimental study on a hybrid tank system. The results indicate that the CRF framework can lead to a better and more robust operating mode diagnosis, and can be a potentially good tool to solve process monitoring problems with temporally correlated data.

Chapter 5

Two-stage Time-varying Hidden Conditional Random Fields with Variable Selection for Process Operating Mode Diagnosis

5.1 Introduction

As the rapid development of modern industrial technologies, process production scales are gaining increasingly large. In addition to ensuring production efficiency, process and operation safety needs to be guaranteed during a large scale process operation. Owing to the applications of distributed control system, a large volume of process data are available, resulting in a rapid development of data-based process monitoring approaches in recent decades [12]. Among the existing data-based process monitoring algorithms, the MSPM approaches have been well developed, such as PCA [115], PLS method [58], etc.. However, since industrial processes are often operated under multiple modes, the inherent assumptions of the conventional MSPM approaches may be violated, causing degradation of monitoring performance. To compensate the weakness of MSPM methods, many multimodal process monitoring strategies have been established, among which the HMMs based monitoring algorithms are gaining wide attentions owing to their capability of modeling mode transitions [37].

However, limited by certain independence assumptions of HMMs, the HMM based process monitoring techniques can become unsatisfactory when the HMM independence assumptions are violated [116]. To compensate such limitations, a probabilistic discriminative model, namely CRF, has been proposed and employed to address both multimodal and dynamic process monitoring problems, with demonstrated superior performances over the HMM based approaches [117]. Moreover, owing to the probabilistic discriminative modeling structure of CRFs, there is no requirement to explicitly model the observation with certain probabilistic distributions, and in theory arbitrary features can be selected for process operating mode diagnosis. Such advantage makes CRFs more flexible and expressive than the HMMs.

During CRF modeling, in order to reduce the computational load and increase the modeling accuracy, a combination of effective techniques is commonly seen when solving the CRF related problems. Instead of using raw observations as CRF model inputs, some feature extraction approaches are employed to pre-process the data at the first stage and then use the extracted features to proceed with the subsequent conventional CRF modeling. Moreover, certain discriminative classifiers, such as logistic regression model and SVM, are also used as the first-stage feature extractors by training a local classifier [118]. The proposed two-stage CRF structure creates discriminative features in the first stage and models the temporal-spatial correlations among the labels in the second stage, which has been demonstrated to outperform the conventional single stage CRFs and provide similar accuracy to more complicated approaches [119]. However, to the best of authors' knowledge, in the existing related literatures of two-stage CRFs, the first-stage local classifiers are simply used to generate discriminative features and no related works of feature selection have been done in the first-stage classifier training.

Since the quality of the data used to build the process monitoring models can have a great impact on the final monitoring performance, using all the available features might not achieve a satisfactory monitoring performance compared with using well selected features [120]. As CRF models allow arbitrarily large number of features to be included, feature selection is particularly important for the training of models due to potential redundancy of PVs. To address the feature selection issue of CRFs, two types of algorithms, namely, filtering and embedding, have been considered in the existing literatures [121]. For instance, in the filtering category, multiple evaluation indices have been proposed to rank and prune some features [122], and as a typical example of the embedding approach, an efficient feature inducing algorithm is implemented by assessing and adding features that can improve the training performance [123]. However, the main drawbacks of the filtering approaches are that the selected features likely have no contributions to the final performance, and due to the large number of features available, the embedding approaches tend to have high computational cost, making the solution intractable.

Therefore, in this thesis, a novel two-stage HCRF model with feature selection is proposed for real-time process operating mode diagnosis. In the first stage, on the basis of the max-margin training strategy [124], the HCRF model is obtained as a local classifier, with features ranked according to the fault relevance. In the second stage, to adapt to the dynamic characteristic of the real processes, a time-varying structure is proposed on the basis of the first-stage HCRF outputs. The innovations of the proposed approach can be summarized as follows: (i) the relevant feature selection is first considered during the first-stage training process of CRFs; (ii) with a time-varying model structure, the proposed algorithm is able to adapt to the process changes in real time, effectively avoiding model performance degradation compared with the existing approaches.

The remainder of this chapter is organized as follows: Section 5.2 summarizes preliminaries and comparisons of the conventional HCRFs and LCCRFs, and then proposes the two-stage HCRF model. In section 5.3, the first-stage HCRF model is formulated, and the second-stage HCRF model is illustrated in section 5.4. The online implementation procedures are integrated in section 5.5. To demonstrate the performance of the proposed algorithm, a numerical case study is performed and detailed in section 5.6. Finally, in section 5.7, conclusions are presented.

5.2 Preliminary and Introduction of the Proposed Two-stage HCRF Model

5.2.1 Preliminaries of HCRFs and LCCRFs

As a probabilistic discriminative model, CRFs directly model the conditional probability between the labels and the observations. Different from the conventional LCCRFs, the conventional HCRF has a conditional probabilistic formulation with latent variables, which can be used to solve structured classification problems [125]. Under the probabilistic framework, a mapping from process observations $\boldsymbol{O} = [O_1, \cdots, O_t, \cdots, O_T]$, $O_t \in \Re^M$, to class label $h \in \mathcal{H}$ is established, where $\mathcal{H}$ is a set of all possible labels. In between process observations $\boldsymbol{O}$ and label h, a set of latent variables $\boldsymbol{d} = [d_1, \cdots, d_l, \cdots, d_L]$, $d_l \in \mathcal{D}$, are introduced for more complicated dynamic modeling, where $\mathcal{D}$ is a finite set containing all the possible latent states. The topology of the latent variables varies from problems to problems. For instance, as illustrated in Fig. 5.1, the one dimensional linear-chain structure is applied on the latent variables. Mathematically, a HCRF model can be represented in the following conditional probabilistic form:

$$P(h, \boldsymbol{d}|\boldsymbol{O}; \Theta) = \frac{e^{F(h,\boldsymbol{d},\boldsymbol{O};\Theta)}}{\sum_{h',\boldsymbol{d}'} e^{F(h',\boldsymbol{d}',\boldsymbol{O};\Theta)}} \tag{5.1}$$

where Θ denotes the unknown parameters of the HCRF model. $F(h, \boldsymbol{d}, \boldsymbol{O}; \Theta) \in \Re$ represents the feature function parameterized by Θ, which is created to model correlations among $\boldsymbol{O}$, $\boldsymbol{d}$ and h.

In parallel, the conventional LCCRFs model the conditional probability between a se-

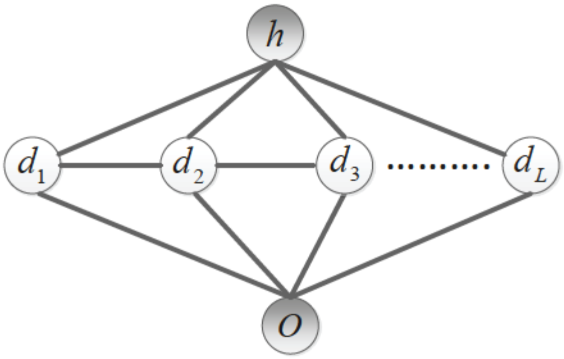


Figure 5.1: The HCRF with linear-chain structure among the latent variables [126], where the shaded nodes represent the observed variables

quence of label $h_{1:T}$ and the observations $O_{1:T}$ with a graphical structure shown in Fig. 5.2. The formulation of LCCRFs is defined as follows:

$$P(h_{1:T}|O_{1:T};\Lambda) = \frac{e^{G(h_{1:T},O_{1:T};\Lambda)}}{\sum_{h'_{1:T}} e^{G(h'_{1:T},O_{1:T};\Lambda)}} \tag{5.2}$$

where Λ represents the unknown parameters of the LCCRFs, which need to be estimated in the training process.

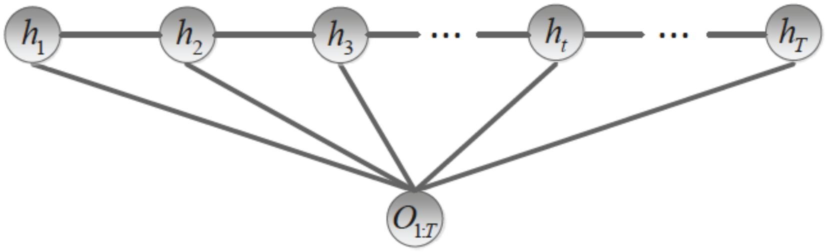


Figure 5.2: The graphical structure of the conventional LCCRF, with the shaded nodes representing the observed variables

The main differences between HCRFs and LCCRFs lie in the existence of the latent layer and the probabilistic modeling of an individual label h or a label sequence $h_{1:T}$. In fact, their difference is similar to the difference between classical state space models and input-output transfer functions. HCRF is analogous to the state space formulation where $\boldsymbol{d}$ is similar to the state, while LCCRF is analogous to the transfer function formulation. Due to the modeling characteristics, both HCRFs and LCCRFs have their own advantages. By adding the latent layer, the HCRFs have higher modeling flexibility and are able to describe the latent features of the process. By modeling the probability between the label sequence and observations, the LCCRFs take the correlations among the labels into consideration. In this work, a two-stage HCRF model is proposed by integrating the advantages of both HCRFs and LCCRFs to solve the process operating mode diagnosis problem. Detailed explanation will be provided in the next section.

5.2.2 Two-stage HCRF Model for Process Operating Mode Diagnosis

By making use of the CRF framework, the process operating mode diagnosis problem can be solved as a sequential classification problem. In this work, the process under consideration is assumed to have multiple operating modes, such as *Normal*, *Abnormal* and *Faulty*, etc.. In different operating modes, the process can exhibit different dynamics or statistical properties, meaning that from the process observations, the current process operating mode may be recognized through dynamical or statistical analysis. The process operating mode at the sampling instant t is represented by $h_t \in \mathcal{H}$, where $\mathcal{H} = \{1, 2, \cdots, P\}$, with P operating modes in total. The objective is to find the most likely operating mode h_t^* given all available process observations $O_{1:t}$.

By removing the data from the transit of switchings between different operating modes, the observations are first separated according to their operating modes and the HCRF model is trained as a local classifier. This formulates the first-stage of the proposed strategy to discriminate different operating modes. In this stage, the max-margin training strategy is employed and the most relevant variables are selected while maintaining the discriminative capacity of the local classifier. In the second-stage of the proposed strategy, the first-stage outputs act as the inputs to the second stage and the dynamic correlations between the operating modes are considered by using a LCCRF model. As illustrated in Fig. 5.3, at each sampling instant, a moving window including its previous observations are taken and evaluated by the first-stage HCRF classifier. A longer moving window is beneficial to clearly discriminate one operating mode from others in the instant where there is no switching occurs, as denoted by the window L_l in Fig. 5.3. However, toward switching instant, a shorter moving window is better in capturing the change of features, as denoted by the window L_s in Fig. 5.3. Therefore, an automatic selection of the window length is desirable and this is conducted in the second stage of the proposed method. In the second-stage modeling, the moving window length is determined which is adaptive to the mode switching.

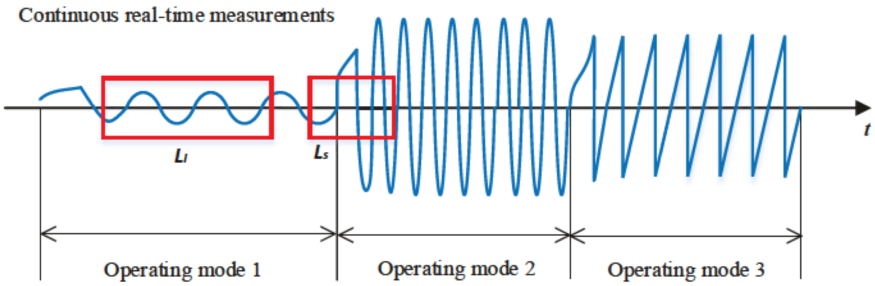


Figure 5.3: The graphical illustration of the process dynamics and moving window strategy

The detailed formulation and training strategies of the two-stage HCRF model are provided in the subsequent sections.

5.3 First-stage HCRF with Process Variable Selection

5.3.1 Problem Formulation

For the first-stage HCRF training, the observations from the same operating mode are extracted and integrated into multiple sequences. Each sequence starts with a transient period which is followed by a steady period. To avoid the impacts of transient period on the training performance of the local classifier, the first few sampling points of each sequence are removed from training dataset, which is shown in Fig. 5.4.

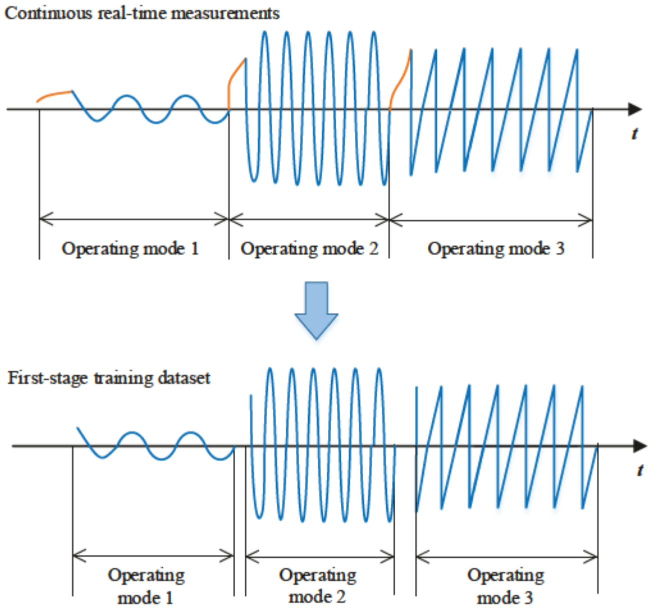


Figure 5.4: The graphical illustration of transient periods removal

In each operating mode, the first-stage HCRF classifier takes an observation sequence with a fixed length L as input, and the corresponding operating mode as output. This local classifier has an internal graphical structure shown in Fig. 5.5 in its simplest form. To increase modeling flexibility, based on Fig. 5.5, each Y_t can be extended to include several of its previous observations as shown in Fig. 5.6, and each Y_t has its corresponding length d_l. Therefore, a latent vector $\boldsymbol{d} = [d_1, \cdots, d_l, \cdots, d_L]$, $d_l \in \{1, 2, \cdots, D\}$ is introduced as shown in Fig. 5.6, which breaks the long chain into smaller windows. With this layout, the observations are re-arranged with reduced number of model parameters comparing with the conventional HCRF. The detailed examples of $D = 1$ and $D = 2$ will be represented later in this section. During the offline training period, longer L is more beneficial to differentiate one operating mode from others. Therefore, L in the first stage model training will be chosen much longer than D. The evolution of $d_{1:L}$ is modeled by a first-order Markov chain, where transition characteristics can be obtained from the process data. As indicated in Fig. 5.6, same graphical structure will be repeated over time.

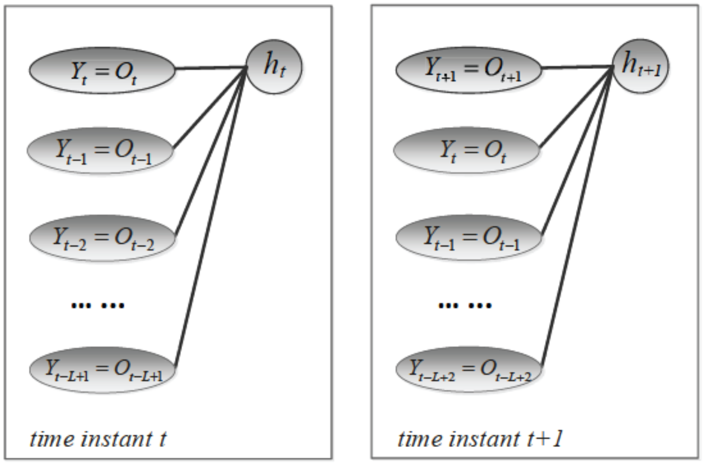


Figure 5.5: The graphical structure of the first-stage HCRF at time t and $t+1$ in one operating mode, which is the simplest form with $D = 1$. The shaded nodes represent the observed variables

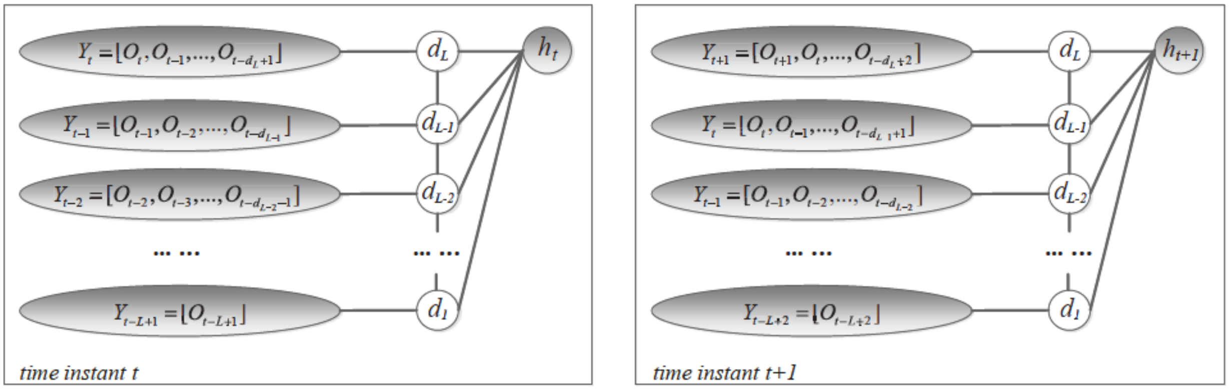

$Y_t = [O_t, O_{t-1}, ..., O_{t-d_L+1}]$
d_L
h_t
$Y_{t-1} = [O_{t-1}, O_{t-2}, ..., O_{t-d_{L-1}}]$
d_{L-1}
$Y_{t-2} = [O_{t-2}, O_{t-3}, ..., O_{t-d_{L-2}-1}]$
d_{L-2}
... ...
... ...
$Y_{t-L+1} = [O_{t-L+1}]$
d_1
time instant t
$Y_{t+1} = [O_{t+1}, O_t, ..., O_{t-d_L+2}]$
d_L
h_{t+1}
$Y_t = [O_t, O_{t-1}, ..., O_{t-d_{L-1}+1}]$
d_{L-1}
$Y_{t-1} = [O_{t-1}, O_{t-2}, ..., O_{t-d_{L-2}}]$
d_{L-2}
... ...
... ...
$Y_{t-L+2} = [O_{t-L+2}]$
d_1
time instant t+1

Figure 5.6: The graphical structure of the first-stage HCRF at time t and $t+1$ in one operating mode. The shaded nodes represent the observed variables

Based on the above structure, two types of feature functions are defined as follows: (i) the feature functions modeling the transitions between d_l and d_{l-1} in the operating mode h_t, namely, $f_{k_1}(h_t, d_l, d_{l-1})$; (ii) the feature functions modeling the connections between the operating mode h_t and the process observations, i.e., $f_{k_2}(h_t, d_l, O_{n_l:n_l-D+1})$ and

$f_{k_3}(h_t, d_l, O_{n_l:n_l-D+1})$, with $n_l = t - L + l$.

The first type of transition feature function has the following binary indicator format:

$$f_{k_1}(h_t, d_l, d_{l-1}) = \begin{cases} 1 \quad \text{if } h_t = p \;\text{ and }\; d_l = i \;\text{ and }\; d_{l-1} = j \\ 0 \qquad\qquad\qquad \text{otherwise} \end{cases} \tag{5.3}$$

where the process operating mode $p \in \{1, 2, \cdots, P\}$, and $i, j \in \{1, 2, \cdots, D\}$. The total number of these transition feature functions is PD^2, and the possibility of transition from $d_{l-1} = j$ to $d_l = i$ in the p^{th} operating mode is evaluated by a weighting factor θ_{k_1}, which is unknown and needs to be identified from the process data.

As to the second type of feature function, the detailed formulation is:

$$f_{k_2}(h_t, d_l, O_{n_l:n_l-D+1}) = \begin{bmatrix} f_{k_{21}}(h_t, O_{n_l}) \\ f_{k_{22}}(h_t, O_{n_l-1}) \\ \vdots \\ f_{k_{2d_l}}(h_t, O_{n_l-d_l+1}) \end{bmatrix} \tag{5.4}$$

where each element $f_{k_{2j}}(h_t, O_{n_l-j+1}) \in \Re^M$ can be written into the following form:

$$f_{k_{2j}}(h_t, O_{n_l-j+1}) = \begin{cases} O_{n_l-j+1} & \text{if } h_t = p \\ 0 & \text{otherwise} \end{cases} \tag{5.5}$$

where a weighting vector $\theta_{k_{2j}} \in \Re^M$ is assigned to evaluate the above feature function.

The feature functions $f_{k_3}(h_t, d_l, O_{n_l:n_l-D+1})$ are formulated similar to $f_{k_2}(h_t, d_l, O_{n_l:n_l-D+1})$. The only difference is that each element in $f_{k_3}(h_t, d_l, O_{n_l:n_l-D+1})$ has a quadratic form as below:

$$f_{k_{3j}}(h_t, O_{n_l-j+1}) = \begin{cases} O_{n_l-j+1} \odot O_{n_l-j+1} & \text{if } h_t = p \\ 0 & \text{otherwise} \end{cases} \tag{5.6}$$

where $\odot$ represents the element-wise product between two matrices.

The above feature functions also have their corresponding weighting factors θ_{k_3}, and the maximal total number of θ_{k_2} and θ_{k_3} are $2MDP$, which can be reduced according to the optimal estimations of latent variables $\boldsymbol{d}$.

Specifically, considering the cases with $D = 1$ in Fig. 5.5, for the operating mode $h_t = p$, f_{k_2} and f_{k_3} can be formulated as $f_{k_2} = f_{k_{21}} = O_{n_l}$ and $f_{k_3} = f_{k_{31}} = O_{n_l}^2$, with two unknown weighting parameters $\theta_{k_{21}}$ and $\theta_{k_{31}}$. When D increases to $D = 2$ as illustrated in Fig. 5.6, for the operating mode $h_t = p$, f_{k_2} and f_{k_3} can be formulated as $f_{k_2} = [f_{k_{21}}, f_{k_{22}}]^T = [O_{n_l}, O_{n_l-1}]^T$ and $f_{k_3} = [f_{k_{31}}, f_{k_{32}}]^T = [O_{n_l}^2, O_{n_l-1}^2]^T$, with four unknown weighting parameters $[\theta_{k_{21}}, \theta_{k_{22}}, \theta_{k_{31}}, \theta_{k_{32}}]$.

By selecting both linear and quadratic feature functions f_{k_2} and f_{k_3}, it has been proven that the statistics of a dataset following Gaussian distributions can be sufficiently described [127], based on which more complicated data structure has been formulated in the HCRF model. On the basis of the feature functions defined above, the function $F(h_t, \boldsymbol{d}, \boldsymbol{O}; \Theta)$ formulated in Equation (5.1) can be specified as below:

$$\begin{aligned} F(h_t, \boldsymbol{d}, \boldsymbol{O}; \Theta) = & \sum_{l=1}^{L} \{ \sum_{k_1} \theta_{k_1} f_{k_1}(h_t, d_l, d_{l-1}) + \sum_{k_2} \theta_{k_2} f_{k_2}(h_t, d_l, O_{n_l:n_l-D+1}) + \\ & \sum_{k_3} \theta_{k_3} f_{k_3}(h_t, d_l, O_{n_l:n_l-D+1}) \} \end{aligned} \tag{5.7}$$

In summary, the diagram of the first-stage HCRF classifier can be found in Fig. 5.7, with specified input and output. The unknown parameters in the first-stage HCRF model are estimated by the max-margin training strategy and during the training process, the relevant PVs are selected to achieve a better classification performance. The details are introduced in the subsequent section.

5.3.2 Training of the First-stage HCRF and Variable Selection

Instead of estimating the HCRF parameters through MLE approach, the max-margin training strategy is applied to find a parameter estimation solution by maximizing the margins

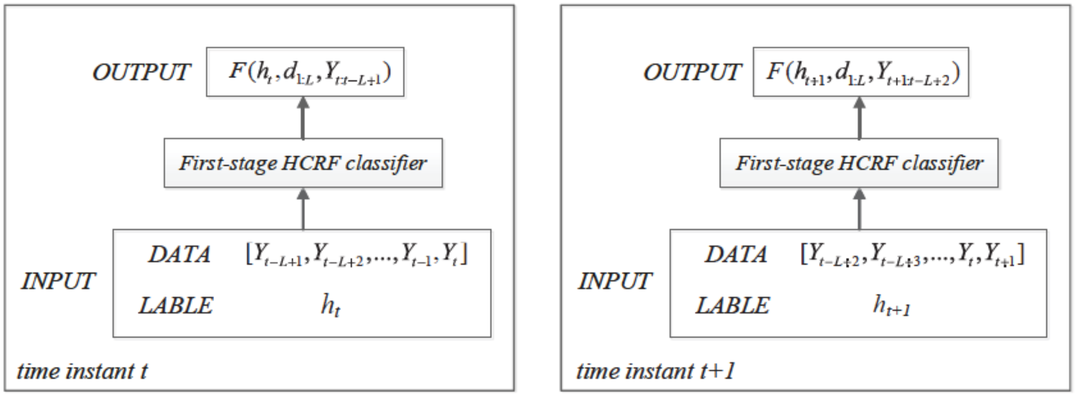

OUTPUT
$F(h_t, d_{1:L}, Y_{t:t-L+1})$
First-stage HCRF classifier
INPUT
DATA
$[Y_{t-L+1}, Y_{t-L+2}, \ldots, Y_{t-1}, Y_t]$
LABLE
h_t
time instant t
OUTPUT
$F(h_{t+1}, d_{1:L}, Y_{t+1:t-L+2})$
First-stage HCRF classifier
INPUT
DATA
$[Y_{t-L+2}, Y_{t-L+3}, \ldots, Y_t, Y_{t+1}]$
LABLE
h_{t+1}
time instant t+1

Figure 5.7: The illustrative diagram of the first-stage HCRF at time t and $t+1$

between the true label and the other labels [124], as illustrated in Fig. 5.8. By training in this way, the HCRF model is named max-margin HCRF (MMHCRF) [124]. In the way that is described in the above section, the training dataset $\{h^{(n)}, \boldsymbol{O}^{(n)}\}_{n=1}^{N}$ are collected by integrating the observations from multiple operating modes, where $h^{(n)}$ and $\boldsymbol{O}^{(n)}$ denote the operating mode identity and the corresponding observations with length L, respectively. With this fully labeled training dataset, the max-margin training process of HCRF can be performed by iterating between the following two steps [124]:

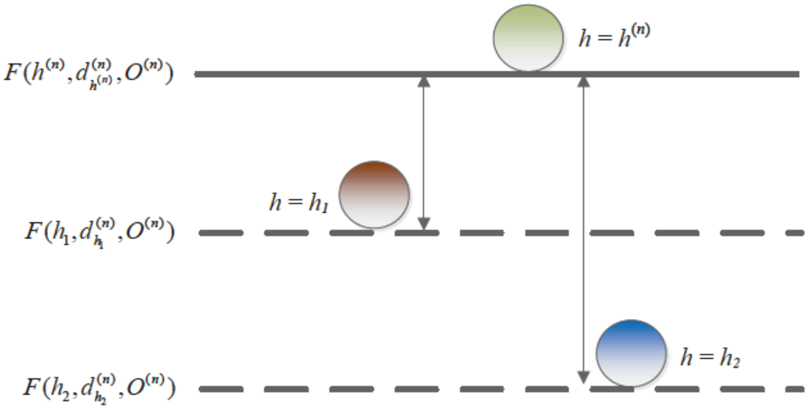

$h = h^{(n)}$
$F(h^{(n)}, d^{(n)}_{h^{(n)}}, O^{(n)})$
$h = h_1$
$F(h_1, d^{(n)}_{h_1}, O^{(n)})$
$h = h_2$
$F(h_2, d^{(n)}_{h_2}, O^{(n)})$

Figure 5.8: The illustration of the MMHCRF training, where the true label is $h^{(n)}$

(1) Fix the HCRF parameters $\Theta = [\theta_{1:K_1}\ \theta_{1:K_2}\ \theta_{1:K_3}]$, and find the optimal latent variable

$\boldsymbol{d}_h^{(n)}$ for each training sample $\{h^{(n)}, \boldsymbol{O}^{(n)}\}$ with respect to all possible labels:

$$\boldsymbol{d}_h^{(n)} = \arg\max_{\boldsymbol{d}} F(h, \boldsymbol{d}, \boldsymbol{O}^{(n)}; \Theta) \tag{5.8}$$

(2) Fix the optimal latent variable $\boldsymbol{d}_h^{(n)}$ derived from the first step for all the training samples, then optimize the HCRF parameters by solving the following optimization problem:

$$\begin{aligned} &\min_{\Theta,\xi} \ \frac{1}{2}||\Theta||^2 + C\sum_{n=1}^{N}\xi_n \\ &s.t. \quad F(h, \boldsymbol{d}_h^{(n)}, \boldsymbol{O}^{(n)}; \Theta) - F(h^{(n)}, \boldsymbol{d}_{h^{(n)}}^{(n)}, \boldsymbol{O}^{(n)}; \Theta) \leqslant \xi_n - \delta(h, h^{(n)}), \quad \forall n, \forall h \end{aligned} \tag{5.9}$$

where ξ_n is the slack variable of the n^{th} training sample, and C is a regularization factor. $\delta(h, h^{(n)})$ has the following form:

$$\delta(h, h^{(n)}) = \begin{cases} 1 & \text{if } h \neq h^{(n)} \\ 0 & \text{otherwise} \end{cases} \tag{5.10}$$

By comparing above equations with the illustration of Fig. 5.8, the first step is implemented to maximize the score $F(h, \boldsymbol{d}_h, O^{(n)})$ for all the nodes with different labels by searching for and fixing the optimal latent variable. Due to the linear-chain structure of the latent variable in this model, the Viterbi algorithm [27] can be applied to obtain the optimal latent variable. Then in the second step, with the selected latent variable fixed, HCRF parameters are estimated by maximizing the margins between the node with true label and the other nodes.

The optimization problem in Equation (5.9) can be solved by optimizing the following dual form through the quadratic programming (QP) strategy [128]:

$$
\begin{aligned}
\max_{\boldsymbol{\alpha}} \quad & \sum_n \sum_h \alpha_{n,h}\delta(h, h^{(n)}) - \frac{1}{2}||\sum_n \sum_h \alpha_{n,h}\psi(\boldsymbol{O}^{(n)}, h)||^2 \\
s.t. \quad & \sum_h \alpha_{n,h} = C, \ \forall n \\
& \alpha_{n,h} \geqslant 0, \ \forall n, \ \forall h
\end{aligned}
\tag{5.11}
$$

where $\psi(\boldsymbol{O}^{(n)}, h)$ represents the feature difference between the node with predicted label h and the node with true label $h^{(n)}$ in the form $\psi(\boldsymbol{O}^{(n)}, h) = f(h, \boldsymbol{d}_h^{(n)}, \boldsymbol{O}^{(n)}) - f(h^{(n)}, \boldsymbol{d}_{h^{(n)}}^{(n)}, \boldsymbol{O}^{(n)})$, and $f(h, \boldsymbol{d}_h^{(n)}, \boldsymbol{O}^{(n)})$ is the concatenation of the HCRF feature functions.

Then, the unknown parameter Θ can be retrieved from the optimized dual variables $\boldsymbol{\alpha}$, as follows:

$$
\Theta = -\sum_{n=1}^{N} \sum_h \alpha_{n,h}\psi(\boldsymbol{O}^{(n)}, h) \tag{5.12}
$$

The above max-margin training strategy is quite similar to the SVM training, therefore in this case, the idea of recursive feature elimination (RFE) strategy in SVM [129] is employed to perform variable selection. The goal of SVM-RFE is to search for a subset of variables among all the available variables which can maximize the classification performance, by iteratively eliminating the most irrelevant variables. For the first-stage HCRF, the evaluation criterion is formulated as

$$
W(\boldsymbol{\alpha}) = \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_h \sum_{h'} \alpha_{i,h}\alpha_{j,h'}\psi^T(\boldsymbol{O}^{(i)}, h)\psi(\boldsymbol{O}^{(j)}, h') \tag{5.13}
$$

For each variable, removal is attempted and the following evaluation criterion is calculated:

$$
W_{(-m)}(\boldsymbol{\alpha}) = \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_h \sum_{h'} \alpha_{i,h}\alpha_{j,h'}\psi^T(\boldsymbol{O}_{(-m)}^{(i)}, h)\psi(\boldsymbol{O}_{(-m)}^{(j)}, h') \tag{5.14}
$$

where m means the m^{th} variable and $\boldsymbol{O}_{(-m)}^{(i)}$ represents the observations after removing the m^{th} variable.

From $m = 1$ to M, the difference $\Delta W_{(-m)} = |W(\boldsymbol{\alpha}) - W_{(-m)}(\boldsymbol{\alpha})|$ is calculated and the variable with the smallest $\Delta W_{(-m)}$ is determined to be removed. The above procedures are equivalent to removing the variables with the weighting factors $[\theta_{k_2}, \theta_{k_3}]$ closest to zero. The pseudocode of variable selection can be found in Appendix C.1.

5.4 Second-stage Time-varying HCRF for Process Operating Mode Diagnosis

5.4.1 The Connection between the First-stage and the Second-stage HCRF

As shown in Fig. 5.9, the identified first-stage HCRF model with variable selection is used as a local classifier. Based on the outputs of the local classifier, the correlations between different operating modes are taken into consideration for sequential classification, known as the second-stage HCRF modeling. In this sense, as shown in Fig. 5.3, a shorter moving window length is desired to quickly capture the switching of process operating mode by involving fewer observations from the previous operating mode. Therefore, an adaption of L over time is motivated in the second-stage HCRF modeling.

As shown in Fig. 5.9, the input features of the second-stage HCRF are from the outputs of the first-stage classifier. A softmax function is applied on the first-stage HCRF outputs at each sampling instant to calculate the probability of each operating mode, namely $X_t(L_t) = [X_t^1(L_t), \cdots, X_t^p(L_t), \cdots, X_t^P(L_t)]^T$. Each element $X_t^p(L_t)$ is formulated as:

$$X_t^p(L_t) = \frac{e^{F(h_t=p, \boldsymbol{d}_p(L_t), \boldsymbol{O}^{(t)}; \Theta)}}{\sum_{h'_t=1}^{P} e^{F(h'_t, \boldsymbol{d}_{h'_t}(L_t), \boldsymbol{O}^{(t)}; \Theta)}} \tag{5.15}$$

where L_t represents the length of the latent variable at time t, which is equivalent to the moving window length at time t.

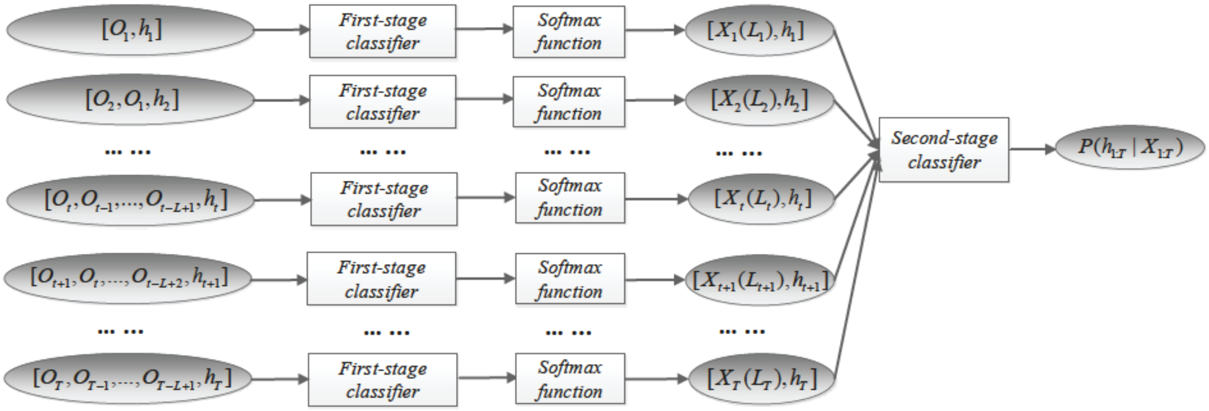


Figure 5.9: The illustration of the two stage connection during the second-stage HCRF training period

5.4.2 Formulation and Parameter Estimation of the Second-stage HCRF

In order to model the adaption of moving window length L_t, an auxiliary label sequence $y_{1:T}$ is introduced and assigned according to the operating mode. As shown in Fig. 5.10, the first L samples in each operating mode are labeled with $y = 2$ and the other samples are labeled with $y = 1$. In this sense, $y = 2$ means that the process has just left the previous operating mode and switched to a new operating mode, requiring a reduced L_t to track the most recent process status more quickly.

With this auxiliary label $y_{1:T}$, the structure of the second-stage HCRF model is illustrated in Fig. 5.11, based on which the following conditional probability is formulated and factorized:

$$\begin{aligned} P(h_{1:T}, y_{1:T}|X_{1:T}; \Lambda, R) &= P(h_{1:T}|y_{1:T}, X_{1:T}; \Lambda) \cdot P(y_{1:T}|X_{1:T}; R) \\ &= \frac{e^{\sum_{L_{1:T}} G(h_{1:T}, y_{1:T}, L_{1:T}, X_{1:T}; \Lambda)}}{\sum_{h'_{1:T}} e^{\sum_{L_{1:T}} G(h'_{1:T}, y_{1:T}, L_{1:T}, X_{1:T}; \Lambda)}} \cdot \frac{\sum_{L_{1:T}} e^{E(y_{1:T}, L_{1:T}, X_{1:T}; R)}}{\sum_{y'_{1:T}} \sum_{L_{1:T}} e^{E(y'_{1:T}, L_{1:T}, X_{1:T}; R)}} \end{aligned} \tag{5.16}$$

The objective of the second-stage HCRF is to maximize the following conditional log

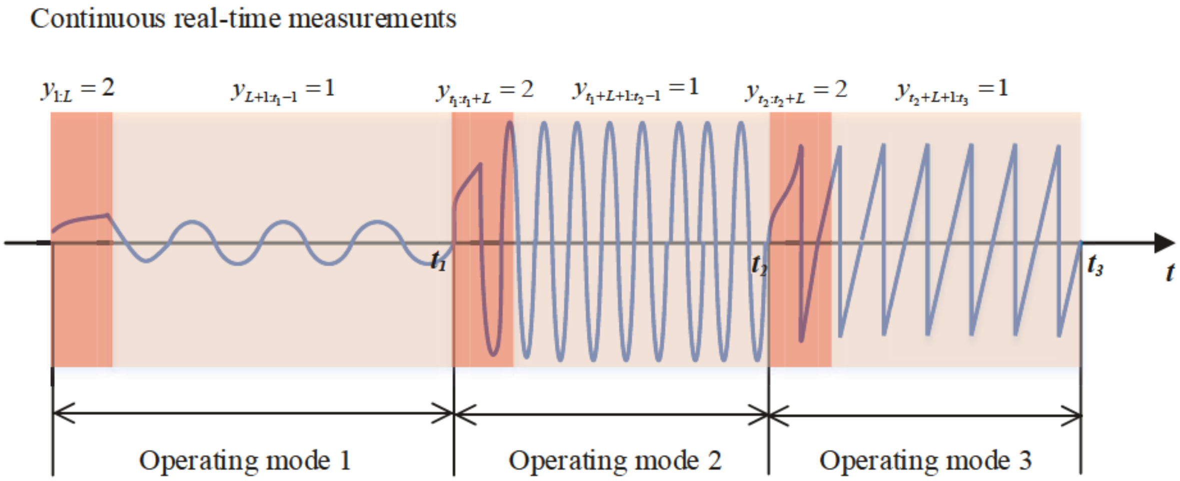


Figure 5.10: The illustration of $y_{1:T}$ labeling

likelihood function:

$$\log P(h_{1:T}, y_{1:T}|X_{1:T}; \Lambda, R) = \log P(h_{1:T}|y_{1:T}, X_{1:T}; \Lambda) + \log P(y_{1:T}|X_{1:T}; R) \tag{5.17}$$

In the above equation, the first term models the conditional probability of actual process operating modes with input features and the auxiliary label sequence $y_{1:T}$, and the second term models the correlation between the auxiliary label sequence and the input features. With such formulation, during online implementation, the auxiliary label is first inferred from the first-stage HCRF outputs, and then the actual operating mode is estimated based on the adjusted window length suggested by the inferred auxiliary label. The two log likelihoods in Equation (5.17) are trained separately: $\log P(y_{1:T}|X_{1:T})$ is first maximized by using the VB approach, and then $\log P(h_{1:T}|y_{1:T}, X_{1:T})$ is optimized by using the CMLE.

As shown in Fig. 5.11, $P(y_{1:T}|X_{1:T})$ has a simplified HCRF structure, with two types of feature functions defined as: (i) the feature functions connecting y_t and L_t, namely, $e_{u_1}(y_t, L_t)$; (ii) the feature functions connecting L_t and X_t, namely, $e_{u_2}(L_t, X_t)$. The upper bound of L_t is assumed to be L and the uncertainty of L_t is modeled by two Dirichlet

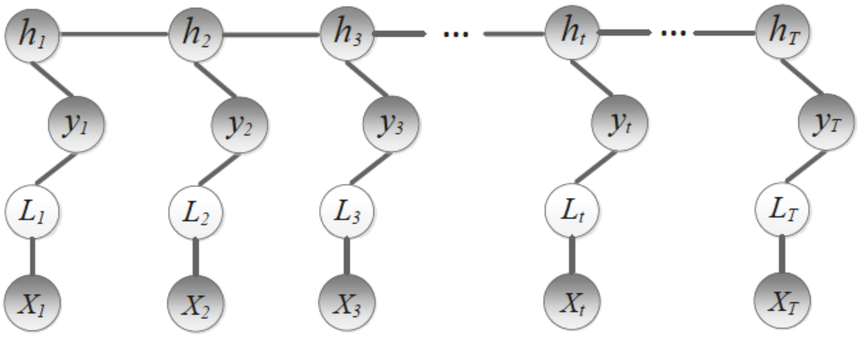


Figure 5.11: The illustration of the second-stage HCRF with shaded nodes representing the observed variables

distributions with respect to different y_t, as follows:

$$e_{u_1}(y_t, L_t; \zeta_{y_t}) = P(L_t = l_1, l_2, \cdots, L | y_t; \zeta_{y_t}) = [\zeta_{1,y_t}, \zeta_{2,y_t}, \cdots, \zeta_{L,y_t}] \sim Dir(K_L, \eta_{y_t}) \quad (5.18)$$

where η_{y_t} is the concentration parameters of the Dirichlet distribution and K_L is the total number of the possible values of L_t. For the sake of reducing the computational load, instead of taking all the values from 1 to L, the possible values of L_t are selected as a subset $\mathcal{L} = \{l_1, l_2, \cdots, L\} \subseteq \{1, 2, \cdots, L\}$.

To evaluate the discriminative capacity of the first-stage HCRF, according to the identified first-stage HCRF model, all the L_t values are tested at each sampling instant t, and within a fixed length window L, the percentages of different estimated operating modes are computed as follows:

$$w_{L_t}(p) = \frac{\#(H^*_{1:L-1}(L_t) = p)}{L-1}, \quad p = 1, 2, \cdots, P \quad (5.19)$$

where the operator $\#(\cdot)$ counts the occurrence times of the events in the bracket, and $H^*_l(L_t) = \text{argmax}_p X^p_{t-l+1}(L_t)$ represents the most likely operating mode estimated from the observations.

An illustrative example of the above calculation is shown in Fig. 5.12. In this example, every adjacent two observations within a window of length L is evaluated based on the first-stage HCRF model. The most likely operating mode sequence $H^*_{1:5}$ is inferred based on every pair of the adjacent observations. In this case, it is observed that with smaller L_t, the potential switching from one operating mode to another can be detected quicker. Using Equation (5.19), in this example, $w_{L_t=2}(1)$ and $w_{L_t=2}(2)$ are calculated as 0.6 and 0.4, respectively.

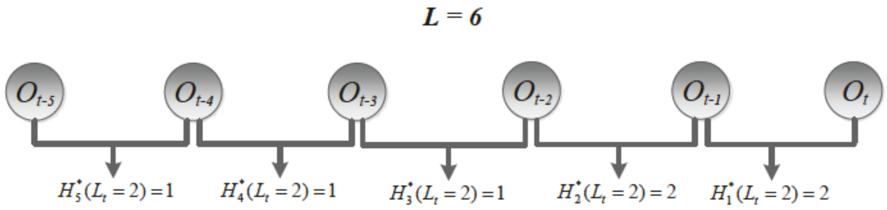


Figure 5.12: An illustrative example of the first-stage HCRF discriminative capacity evaluation

Generally, within the same operating mode, both larger and smaller L_t can give consistent $w_{L_t}(p)$, while longer L_t can provide more reliable classification result than smaller L_t. If $w_{L_t}(p)$ decreases from a relatively high value to a low value, there is a possibility of the operating mode switching. Hence, the negative changing slope of $w_{L_t}(p)$ in a fixed length short period is inspected and selected as a feature to reflect the switching at t, denoted as V_{L_t}. Then the feature function $e_{u_2}(L_t, X_t)$ can be designed as:

$$e_{u_2}(L_t, X_t) = \begin{cases} e^{-\frac{\|V_{L_t}\|^2}{\varepsilon_1^2}} & \text{if } L_t \in L_s \\ e^{-\frac{\|V_{L_t} - V_{l_m}\|^2}{\varepsilon_1^2}} & \text{if } L_t \in L_l \end{cases} \tag{5.20}$$

where L_s and L_l represent the subsets of smaller and larger L_t, respectively. l_m denotes the median value in L_s.

Then, the conditional log likelihood $\log P(y_{1:T}|X_{1:T}; R)$ can be optimized by iteratively performing the following VB-E and VB-M steps.

In the VB-E step, the KL divergence $D_{KL}(q(L_{1:T})q(\boldsymbol{\zeta})||P(L_{1:T},\boldsymbol{\zeta}|y_{1:T},X_{1:T}))$ is first minimized to obtain the variational posteriors $q(L_t)$ and the variational parameters of $q(\boldsymbol{\zeta})$. The factorization of D_{KL} can be found as follows:

$$\begin{aligned}
&D_{KL}(q(L_{1:T})q(\boldsymbol{\zeta})||P(L_{1:T},\boldsymbol{\zeta}|y_{1:T},X_{1:T})) = \int_{\boldsymbol{\zeta}}\sum_{L_{1:T}} q(L_{1:T})q(\boldsymbol{\zeta})\log\frac{q(L_{1:T})q(\boldsymbol{\zeta})}{P(L_{1:T},\boldsymbol{\zeta}|y_{1:T},X_{1:T})}d\boldsymbol{\zeta}\\
=&\int_{\boldsymbol{\zeta}} q(\boldsymbol{\zeta})\log q(\boldsymbol{\zeta})d\boldsymbol{\zeta} + \sum_{L_{1:T}} q(L_{1:T})\log q(L_{1:T}) - \int_{\boldsymbol{\zeta}}\sum_{L_{1:T}} q(L_{1:T})q(\boldsymbol{\zeta})\log P(L_{1:T},\boldsymbol{\zeta},y_{1:T},X_{1:T})d\boldsymbol{\zeta}\\
&+ C_{q(L_{1:T})q(\boldsymbol{\zeta})}\\
=&\int_{\boldsymbol{\zeta}} q(\boldsymbol{\zeta})\log q(\boldsymbol{\zeta})d\boldsymbol{\zeta} + \sum_{L_{1:T}} q(L_{1:T})\log q(L_{1:T}) - \int_{\boldsymbol{\zeta}}\sum_{L_{1:T}} q(L_{1:T})q(\boldsymbol{\zeta})\log P(L_{1:T},y_{1:T},X_{1:T}|\boldsymbol{\zeta})d\boldsymbol{\zeta}\\
&- \int_{\boldsymbol{\zeta}} q(\boldsymbol{\zeta})\log P(\boldsymbol{\zeta}|\eta)d\boldsymbol{\zeta} + C_{q(L_{1:T})q(\boldsymbol{\zeta})}
\end{aligned} \tag{5.21}$$

where $\boldsymbol{\zeta} = [\boldsymbol{\zeta}_{yt=1},\ \boldsymbol{\zeta}_{yt=2}]$, $C_{q(L_{1:T})q(\boldsymbol{\zeta})}$ incorporates the terms irrelevant to the variational posteriors $q(L_{1:T})$ and $q(\boldsymbol{\zeta})$, and can be treated as a constant.

While minimizing D_{KL} with respect to $q(L_{1:T})$, by substituting the complete log likelihood $\log P(L_{1:T},y_{1:T},X_{1:T}|\boldsymbol{\zeta})$ and arranging all the irrelevant terms into constant $C_{q(L_{1:T})}$, Equation (5.21) can be rewritten as

$$\begin{aligned}
D_{KL}(q(L_{1:T})) =& \sum_{L_{1:T}} q(L_{1:T}) \log q(L_{1:T}) - \int_{\boldsymbol{\zeta}} \sum_{L_{1:T}} q(L_{1:T}) q(\boldsymbol{\zeta}) \log P(L_{1:T}, y_{1:T}, X_{1:T}|\boldsymbol{\zeta}) d\boldsymbol{\zeta} \\
&+ C_{q(L_{1:T})} \\
=& \sum_{L_{1:T}} q(L_{1:T}) \log q(L_{1:T}) - \int_{\boldsymbol{\zeta}} \sum_{L_{1:T}} q(L_{1:T}) q(\boldsymbol{\zeta}) \{ \sum_{t=1}^{T} [\sum_{u_1} e_{u_1}(y_t, L_t; \boldsymbol{\zeta}) + \\
&\sum_{u_2} \gamma_{u_2} e_{u_2}(L_t, X_t)]\} + C_{q(L_{1:T})} \\
=& \sum_{t=1}^{T} \sum_{L_t} q(L_t) \log q(L_t) - \sum_{t=1}^{T} \sum_{L_t} q(L_t) \langle \sum_{u_1} e_{u_1}(y_t, L_t; \boldsymbol{\zeta}) \rangle_{q(\boldsymbol{\zeta})} - \sum_{t=1}^{T} \\
&\sum_{L_t} q(L_t) (\sum_{u_2} \gamma_{u_2} e_{u_2}(L_t, X_t)) + C_{q(L_{1:T})}
\end{aligned} \tag{5.22}$$

By taking derivative of the above equation with respect to $q(L_t)$, the variational posterior $q(L_t)$ can be computed as below:

$$q(L_t) \propto \exp[\langle \sum_{u_1} e_{u_1}(y_t, L_t; \boldsymbol{\zeta}) \rangle_{q(\boldsymbol{\zeta})} + \sum_{u_2} \gamma_{u_2} e_{u_2}(L_t, X_t)] \tag{5.23}$$

Similarly, when minimizing D_{KL} with respect to $q(\boldsymbol{\zeta})$, KL divergence is reformulated as

$$\begin{aligned}
D_{KL}(q(\boldsymbol{\zeta})) = & \int_{\boldsymbol{\zeta}} q(\boldsymbol{\zeta}) \log q(\boldsymbol{\zeta}) d\boldsymbol{\zeta} - \int_{\boldsymbol{\zeta}} \sum_{L_{1:T}} q(L_{1:T}) q(\boldsymbol{\zeta}) \log P(L_{1:T}, y_{1:T}, X_{1:T}|\boldsymbol{\zeta}) d\boldsymbol{\zeta} \\
& - \int_{\boldsymbol{\zeta}} q(\boldsymbol{\zeta}) \log P(\boldsymbol{\zeta}|\eta) d\boldsymbol{\zeta} + C_{q(\boldsymbol{\zeta})}
\end{aligned} \tag{5.24}$$

Since $\boldsymbol{\zeta}$ follows a Dirichlet distribution, assuming the variational posterior $q(\boldsymbol{\zeta}) \sim Dir(\nu)$ and substituting the complete log likelihood $\log P(L_{1:T}, y_{1:T}, X_{1:T}|\boldsymbol{\zeta})$ into the above equation, then by minimizing D_{KL}, the variational parameters are obtained in the following equation.

The detailed derivations can be found in Appendix C.2.

$$\nu_{y_t,l} = \eta_{y_t,l} + \sum_{t=1}^{T} q(L_t = l) \tag{5.25}$$

In the VB-M step, the following log likelihood function is the objective function to be optimized:

$$\log P(y_{1:T}|X_{1:T};r) = \log \sum_{L_{1:T}} e^{E(y_{1:T},L_{1:T},X_{1:T};r)} - \log \sum_{y'_{1:T}} \sum_{L_{1:T}} e^{E(y'_{1:T},L_{1:T},X_{1:T};r)} \tag{5.26}$$

From the VB-E step, the feature function $e_{u_1}(y_t, L_t)$ has been parameterized. Therefore in the above objective function, we treat only γ_{u_2} as the unknown parameter to be identified. By taking partial derivative of Equation (5.26) with respect to γ_{u_2}, the partial derivative can be obtained as

$$\frac{\partial \log P(y_{1:T}|X_{1:T};r)}{\partial \gamma_{u_2}} = \sum_{t=1}^{T} \sum_{L_t} P(L_t|y_t, X_t) \cdot e_{u_2}(L_t, X_t) - \sum_{t=1}^{T} \sum_{y'_t} \sum_{L_t} P(y'_t, L_t|X_t) \cdot e_{u_2}(L_t, X_t) \tag{5.27}$$

Then the quasi-Newton algorithms such as L-BFGS approach can be employed to solve the above optimization problem for estimation of γ_{u_2} [105].

After optimizing the second term in Equations (5.16) and (5.17), the first term can then be maximized as a conventional LCCRF. Similar to the first-stage HCRF, the feature functions of $\log P(h_{1:T}|y_{1:T}, X_{1:T}; \Lambda)$ are formulated as follows:

$$g_{w_1}(h_t, h_{t-1}) = \begin{cases} 1 \quad \text{if } h_t = p_1 \quad \text{and} \quad h_{t-1} = p_2 \\ 0 \qquad\qquad \text{otherwise} \end{cases} \tag{5.28}$$

$$g_{w_2}(h_t, y_t, X_t(L_t)) = \begin{cases} X_t(L_t) \cdot P(L_t|y_t) & \text{if } h_t = p_1 \\ 0 & \text{otherwise} \end{cases} \tag{5.29}$$

$$g_{w_3}(h_t, y_t, X_t(L_t)) = \begin{cases} (X_t(L_t))^2 \cdot P(L_t|y_t) & \text{if } h_t = p_1 \\ 0 & \text{otherwise} \end{cases} \tag{5.30}$$

where the probability $P(L_t|y_t)$ is obtained from the feature function $e_{u_1}(y_t, L_t)$ in Equation (5.18).

Corresponding to the feature functions shown in the above equations, a set of weighting factors $\Lambda = \{\{\lambda_{w_1}\}_{w_1=1}^{W_1}, \{\lambda_{w_2}\}_{w_2=1}^{W_2}, \{\lambda_{w_3}\}_{w_3=1}^{W_3}\}$ are treated as unknown parameters of the CRF model, which can be estimated by maximizing the conditional log likelihood function $\log P(h_{1:T}|y_{1:T},\ X_{1:T}; \Lambda)$, through the following partial derivatives with respect to the unknown parameters:

$$\begin{aligned} \frac{\partial \log P(h_{1:T}|y_{1:T}, X_{1:T}; \Lambda)}{\partial \lambda_{w_1}} &= \sum_{t=1}^{T} g_{w_1}(h_t, h_{t-1}) - \sum_{t=1}^{T} \sum_{h'_t, h'_{t-1}} P(h'_t, h'_{t-1}|y_{1:T}, X_{1:T}) \cdot g_{w_1}(h'_t, h'_{t-1}) \\ \frac{\partial \log P(h_{1:T}|y_{1:T}, X_{1:T}; \Lambda)}{\partial \lambda_{w_2}} &= \sum_{t=1}^{T} g_{w_2}(h_t, y_t, X_t) - \sum_{t=1}^{T} \sum_{h'_t} P(h'_t|y_{1:T}, X_{1:T}) \cdot g_{w_2}(h'_t, y_t, X_t) \\ \frac{\partial \log P(h_{1:T}|y_{1:T}, X_{1:T}; \Lambda)}{\partial \lambda_{w_3}} &= \sum_{t=1}^{T} g_{w_3}(h_t, y_t, X_t) - \sum_{t=1}^{T} \sum_{h'_t} P(h'_t|y_{1:T}, X_{1:T}) \cdot g_{w_3}(h'_t, y_t, X_t) \end{aligned} \tag{5.31}$$

where the conditional probabilities $P(h'_t, h'_{t-1}|y_{1:T}, X_{1:T})$ and $P(h'_t|y_{1:T}, X_{1:T})$ can be solved by the forward-backward algorithm [27].

5.5 Online Implementation

After completing the training, the proposed two-stage HCRF algorithm is employed for online operating mode diagnosis. From the first-stage HCRF, the variable selection is completed from the training and applied to the continuous online process observations, and then the second stage HCRF is deployed for real-time process operating mode diagnosis. The objective of the online operating mode diagnosis is to find the optimal current mode h_t^*, which maximizes the conditional probability $P(h_t|X_{1:t}; \Lambda^*, R^*)$ conditioned on the estimated model parameters and data.

As shown below, the marginal conditional probability can be derived following a two-step inference procedure:

$$\begin{aligned} P(h_t|y^*_{1:t}, X_{1:t}; \Lambda^*) &= \sum_{h_{1:t-1}} P(h_{1:t}|y^*_{1:t}, X_{1:t}; \Lambda^*) \\ &= \frac{\alpha_t(h_t)}{\sum_{h'_t} \alpha_t(h'_t)} \end{aligned} \tag{5.32}$$

where $\alpha_t(h_t) = \sum_{h_{1:t-1}} e^{\sum_{L_{1:t}} G(h_{1:t}, y^*_{1:t}, L_{1:t}, X_{1:t}; \Lambda^*)}$ is an intermediate variable which can be solved recursively through $\alpha_t(h_t) = \sum_{h_{t-1}} \alpha_{t-1}(h_{t-1}) \cdot e^{\sum_{L_t} G(h_{t:t-1}, y^*_t, L_t, X_t; \Lambda^*)}$ from sampling instant 1 to t.

The optimal $y^*_{1:t}$ is estimated as follows:

$$\begin{aligned} y^*_{1:t} &= \arg\max_{y_{1:t}} P(y_{1:t}|X_{1:t}; R^*) \\ &= \arg\max_{y_{1:t}} \prod_{t'=1}^{t} P(y_{t'}|X_{t'}; R^*) \end{aligned} \tag{5.33}$$

Then the current operating mode can be estimated by $h^*_t = \arg\max_{h_t} P(h_t|y^*_{1:t}, X_{1:t}; \Lambda^*)$.

5.6 Case Study

In this section, a simulation is conducted to validate the performance of the proposed two-stage HCRF algorithm for process operating mode diagnosis. As comparisons, the conventional LCCRF is employed to demonstrate the performances of the proposed two-stage HCRF algorithm.

5.6.1 Simulation

In this numerical study, a process with eight variables is simulated which operates in three different operating modes 1 - 3. In different process operating modes, observations are simulated to follow different statistical distributions and autocorrelations. As shown in Table 5.1, for each operating mode, a multivariate Gaussian distribution of five variables is first designed as a base distribution. By taking the data generated from the base distributions as inputs to a set of autoregressive models to generate intermediate variable $o_{1:T}$, autocorrelated data are generated. To make the monitoring problem more challenging, the final process observations $O_{1:T}$ are simulated by summing up $o_{1:T}$ in a moving window as indicated in Table 5.1. In addition to the above five PVs, another three PVs are simulated, which are not related to the actual process operating modes, as redundant variables.

Table 5.1: Basis distributions and autoregressive formulations for different operating modes

Operating mode	**Operating mode 1**	**Operating mode 2**	**Operating mode 3**
Basis distribution parameters	$\mu_1 = [6, 6, 3, 8, 10]$ $\Sigma_1 = \begin{bmatrix} 3 & 0 & 0 & 0 & 0 \\ 0 & 3 & 0.2 & 0.2 & 0.2 \\ 0 & 0.2 & 3 & 0.5 & 0.5 \\ 0 & 0.2 & 0.5 & 10 & 0.1 \\ 0 & 0.2 & 0.5 & 0.1 & 5 \end{bmatrix}$ $s_1^{(1:5)} \sim \mathcal{N}(\mu_1, \Sigma_1)$	$\mu_2 = [5, 5, 4, 8, 10]$ $\Sigma_2 = \begin{bmatrix} 2 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0.2 & 0.2 & 0.2 \\ 0 & 0.2 & 3 & 0.5 & 0.5 \\ 0 & 0.2 & 0.5 & 10 & 0.1 \\ 0 & 0.2 & 0.5 & 0.1 & 5 \end{bmatrix}$ $s_2^{(1:5)} \sim \mathcal{N}(\mu_2, \Sigma_2)$	$\mu_3 = [5, 5, 3, 8, 11]$ $\Sigma_3 = \begin{bmatrix} 2 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0.2 & 0.2 & 0.2 \\ 0 & 0.2 & 3 & 0.5 & 0.5 \\ 0 & 0.2 & 0.5 & 3 & 0.1 \\ 0 & 0.2 & 0.5 & 0.1 & 5 \end{bmatrix}$ $s_3^{(1:5)} \sim \mathcal{N}(\mu_3, \Sigma_3)$
Autoregressive formulation	$o^{(1)}(t) = s^{(1)}(t)$ $o^{(2)}(t) = 0.5o^{(2)}(t-1) - 0.8o^{(2)}(t-2) - 0.3o^{(2)}(t-3) + 0.1o^{(2)}(t-4) - 0.1o^{(2)}(t-5) + s^{(2)}(t)$ $o^{(3)}(t) = 0.5o^{(3)}(t-1) - 0.1o^{(3)}(t-2) + s^{(3)}(t)$ $o^{(4)}(t) = 0.5o^{(4)}(t-1) - 0.1o^{(4)}(t-2) - 0.8o^{(4)}(t-3) + 0.25o^{(4)}(t-4) + s^{(4)}(t)$ $o^{(5)}(t) = 0.5o^{(5)}(t-1) - 0.6o^{(5)}(t-2) + 0.25o^{(5)}(t-3) + s^{(5)}(t)$		
Moving summation formulation	$O(t) = o(t) + o(t-1) + o(t-2) + o(t-3)$		

A training dataset with 9000 samples is then selected to estimate the parameters of the proposed two-stage HCRF model and a validation dataset with 8000 samples is selected to test the performance of the proposed algorithm. In the first-stage HCRF modeling, D and L are selected as 6 and 50, respectively. From the first-stage HCRF training, importance of variables is estimated and the importance rank of all the eight variables is depicted in Fig. 5.13. As shown in Fig. 5.13, the first five PVs are ranked with higher importance than the last three variables, which is consistent to the data generation. By evaluating the training and validation performances with different selection of variables among the first five variables, the first four variables are selected from the first-stage HCRF training. To evaluate the classification performance of using the selected PVs, the confusion matrix of the test dataset is indicated in Fig. 5.14.

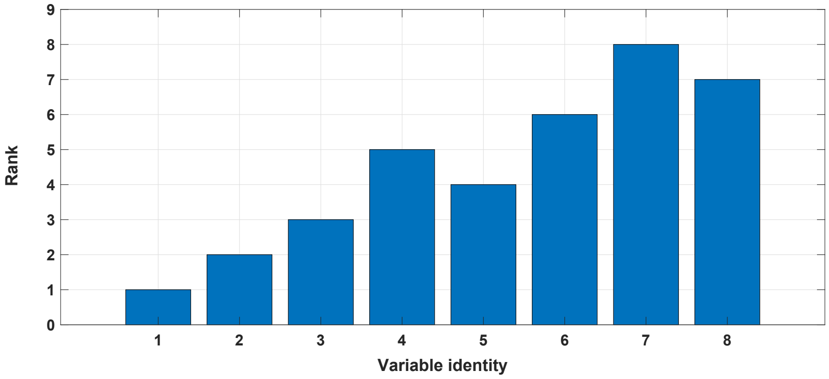


Figure 5.13: The rank of the eight PVs, with smaller rank indicating higher importance

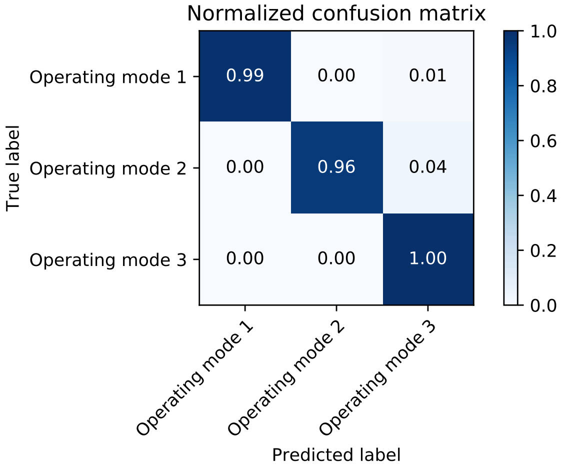


Figure 5.14: The confusion matrix with selected variables of the first-stage HCRF for the numerical case study

In order to take the operating mode switching dynamics into consideration, the second-stage HCRF is deployed for online process operating mode diagnosis with time-varying L_t.

From the training results of the Dirichlet distribution, the distributions of L_t with respect to $y_t = 2$ and $y_t = 1$ are illustrated in Fig. 5.15. It indicates that in the operating mode switching transient periods, smaller window length L_t is more reliable to identify the actual operating mode, and in each steady operating mode, larger window length L_t can provide more accuracy diagnosis result. The mode switching period detection results can be found in Fig. 5.16.

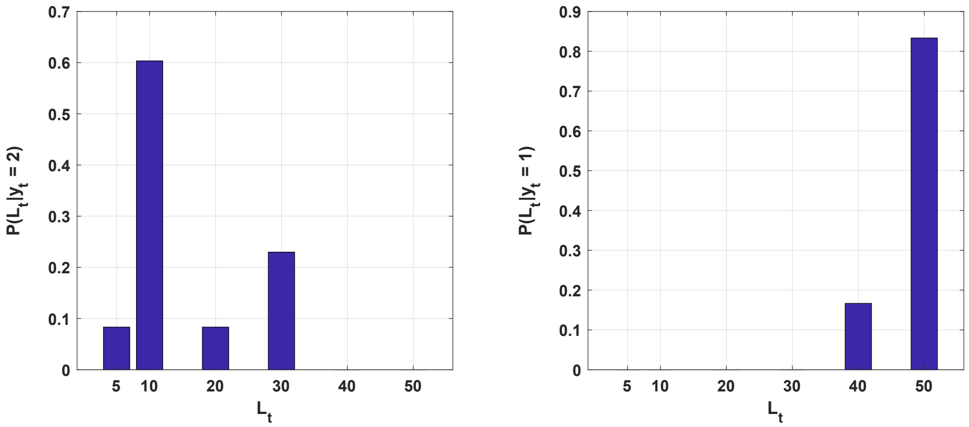


Figure 5.15: The distributions of L_t with respect to $y_t = 2$ and $y_t = 1$

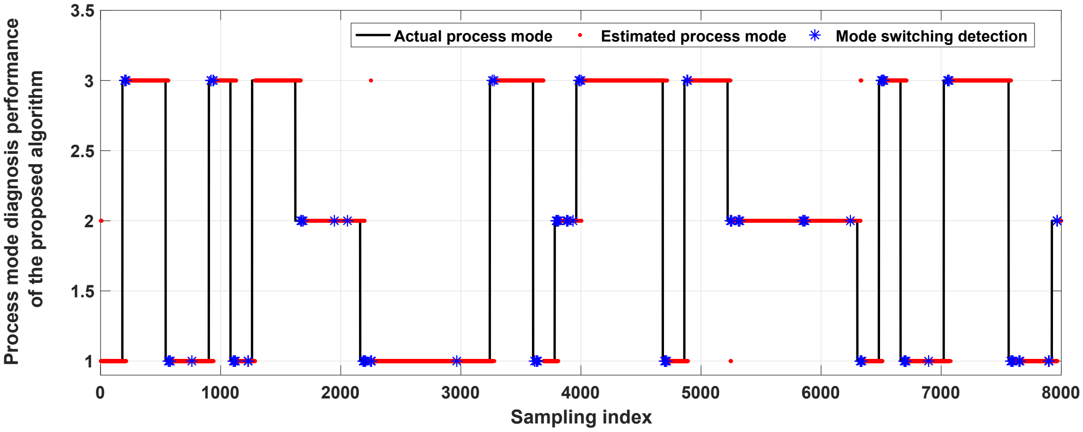


Figure 5.16: The operating mode switching period detections of the proposed algorithm

With variable selection, the proposed algorithm can effectively handle the changes on the fault-independent PVs and detection results are not being affected. For the simulation purposes, two kinds of disturbances are introduced to the last two PVs, as shown in Fig. 5.17. Starting from the 4000^{th} sampling instant, a randomly generated bias and a gradually ramping disturbance are added onto the 7^{th} and 8^{th} variables, respectively. The operating mode diagnosis performances of the conventional LCCRF and the proposed two-stage HCRF algorithms are compared and illustrated in Fig. 5.18. Starting from the 4000^{th} sampling index, due to the disturbances acting on the last two PVs, the diagnosis performance of the conventional LCCRF degrades and finally loses the diagnosis capability. In contrast, as to the proposed algorithm, because the most relevant PVs are selected in the first-stage training period, the subsequent disturbances on the fault-independent PVs do not cause the loss of diagnosis capacity. Moreover, even before introducing the disturbance, the proposed algorithm has slight better diagnosis performance than the LCCRF. For more details, the diagnosis accuracy can be found in Table 5.2.

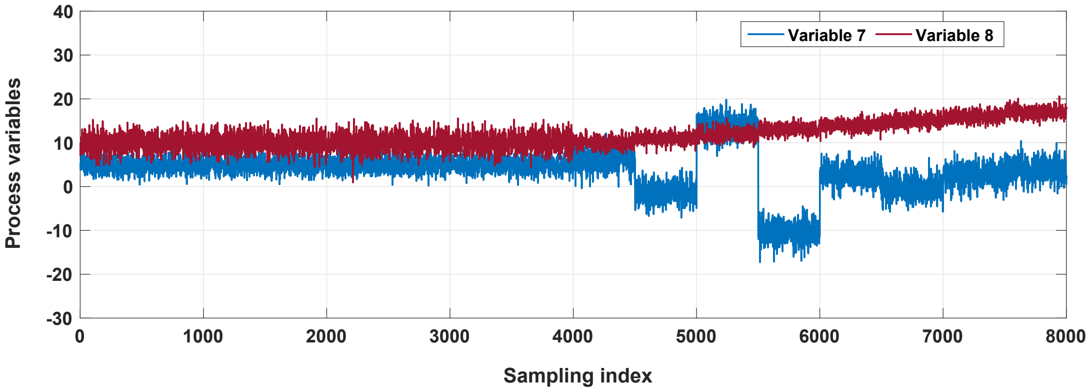


Figure 5.17: The validation dataset illustration

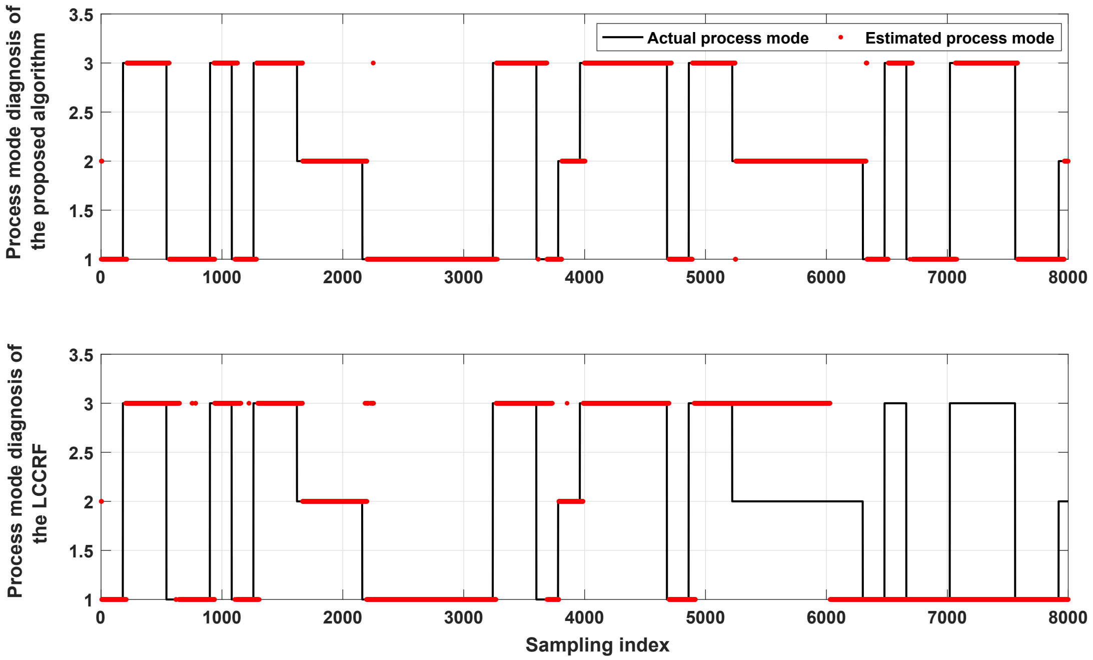


Figure 5.18: The process operating mode diagnosis performance comparison of the numerical case

Table 5.2: Process operating mode diagnosis performance comparison

	Diagnosis accuracy without disturbances	Overall diagnosis accuracy
Proposed algorithm	89.30%	90.73%
LCCRF	86.00%	68.65%

5.7 Conclusions

In this chapter, a two-stage HCRF algorithm for real-time process operating mode diagnosis is proposed and explained in details. Considering that the archived industrial PVs contain both relevant and irrelevant information to the actual abnormalities, the first-stage HCRF is designed to explore selection of PV subsets based on recursively eliminating the irrelevant variables, which reduces the number of PVs used for modeling. Meanwhile, by taking the dynamic characteristics of the actual processes into consideration, the second-stage HCRF is proposed by making use of the outputs of the first-stage HCRF. In the second-stage HCRF, an effective algorithm is designed to detect operating mode switchings and adapt the dynamics of switchings by adjusting the moving window length in real time. The VB approach is employed for parameter estimation of the second-stage HCRF. To demonstrate the performance of the proposed algorithm, a numerical simulation is studied and explained in details, and superior operating mode diagnosis performance is achieved when comparing with the conventional approaches. In conclusion, the proposed algorithm has the capability to select the abnormality relevant PVs and track the process dynamic variations, which contributes to a more reliable abnormality detection strategy.

Chapter 6

Real-time Mode Diagnosis for Processes with Multiple Operating Conditions Using Switching Conditional Random Fields

6.1 Introduction

In order to ensure process safety and the product quality, effective strategies for real-time process monitoring are necessary. The primary objectives of process monitoring include detection and diagnosis of abnormal modes during the process operation. By making use of the available process information, both knowledge based and model based approaches [130] have been developed to solve the process monitoring problem. As a result, some advanced methods [131, 132] have also been developed as the foundation of more complicated process monitoring solutions. On the other hand, in the recent decade, abundance of real industrial

[3] Part of this chapter has been published as Mengqi Fang, Hariprasad Kodamana, and Biao Huang. Real-time mode diagnosis for processes with multiple operating conditions using switching conditional random fields. *IEEE Transactions on Industrial Electronics*, 67(6):5060–5070, 2020.

process data has resulted in wide spread popularity of data based approaches for monitoring, for example, the MSPM. Recently, probabilistic counterparts of MSPM approaches are also gaining wide attentions [133, 66]. On the other hand, HMM based strategies have also been employed to deal with process mode diagnosis problems for dynamic systems, owing to its capability to model temporal correlations and multi-modal dynamics [37, 134]. Even though the application of HMMs in process monitoring has shown a lot of promising results, there are some shortcomings. As a probabilistic generative model, HMM has two inherent conditional independence assumptions. If those assumptions are not satisfied in reality, the modeling capability of HMMs might not be sufficient to describe the real process [86]. Therefore, it might fail or lead to degraded performance in process monitoring [117, 135].

To circumvent such limitations of HMMs, a probabilistic discriminative model, namely, CRF, has been proposed [43]. To deal with particular problems, multiple types of CRF models have been developed. For instance, a HCRF model has been designed to incorporate latent information for better observation descriptions [125]. For more complicated scenario modeling, the hierarchical CRF and dense CRF models are established [136, 137]. Additionally, to improve the model training performance, some extended CRF model structures have been implemented, such as Bayesian CRF [138] and max-margin CRF [124], etc.. However, most of the existing CRF models are mainly designed to solve the computer science related problems, such as natural language processing, image processing, etc. [42]. Owing to its success in dealing with sophisticated classification problems, CRF has been considered as a promising approach to solve process monitoring problems. Recently, a marginalized CRF based approach for real-time process mode diagnosis, with incomplete measurements has been developed and shown to outperform HMM in solving process mode diagnosis problems [117].

Most of the process industries are operated under multiple operating conditions corresponding to different operating requirements, product qualities and load levels [87]. Meanwhile, in such cases, the process data also exhibit multi-modal behaviors, which needs to be specifically addressed. For instance, a sparse modeling and dictionary learning approach has been proposed recently to deal with this issue [139]. Furthermore, as indicated previously, when the datasets are simultaneously multi-modal and dynamic, HMMs may still be used for process diagnosis [140]. However, due to the previously indicated limitations of the HMMs, process mode diagnosis based on multi-modal datasets with strong temporal correlations will not be effective. To address this weakness, a SCRF based framework is proposed in this work. Under this framework, a scheduling variable is utilized to infer the current operating condition and subsequently to determine the status of the current process mode, such as normal, abnormal and failure. The proposed SCRF framework is similar to the mixture of CRF models [141] in that it uses a collection of CRF models to capture multi-modal scenarios. However, the key difference is that, the proposed framework also models the switching between operating conditions. For parameter estimation of SCRF, the EM algorithm is employed. Once a suitable SCRF model is developed, it is deployed for on-line process mode diagnosis. As a result, the contributions of this proposed SCRF framework are summarized as below: (i) from the theoretical perspective, the dynamic switching framework is firstly proposed to improve the conventional CRF modeling capability to describe the industrial processes with multiple operating conditions. Correspondingly, an innovative model parameter estimation approach has also been developed for the proposed SCRF model; (ii) from the practical aspect, as a probabilistic discriminative model, the proposed SCRF framework inherits the advantages from the conventional CRF models, and therefore it is able to compensate the weakness of HMMs with a simple and flexible framework. Such framework can make the informative process features easily involved into the SCRF model if they are useful.

This work is an extension of the conference paper by Fang et al. [142], and the additional contributions with respect to [142] include: (i) development of a simplified SCRF parameter estimation approach to improve computational efficiency, (ii) extended validations of the SCRF approach through a CSTR process and an experimental hybrid tank system. With respect to the simplified SCRF parameter estimation approach, the information retrieved from the scheduling variable has been fully used to help derive the SCRF model parameters, by introducing a latent PV for operating condition segment location. Such enhancement not only increases the computational efficiency during the model training process, but also effectively decreases the undesired oscillations of diagnosed process modes from the real ones.

The remainder of this chapter is organized as follows: Preliminaries of the LCCRF model are presented in section 6.2. In section 6.3, the SCRF formulation, the EM algorithm based parameter estimation of the proposed SCRF model and the corresponding on-line process mode diagnosis strategy are illustrated in detail. Section 6.4 presents the validation performances using a simulated CSTR example and an experimental hybrid tank system.

6.2 Preliminaries of LCCRFs for Process Mode Diagnosis

A LCCRF model is an undirected probabilistic graphical model that can be employed to describe the relationships between observation data sequence $\boldsymbol{O} = [O_1, \cdots, O_T]$, where $O_t \in \Re^Q$, and the process mode sequence $\boldsymbol{h} = [h_1, \cdots, h_T]$, where $h_t \in \{1, 2, \cdots, N\}$, as shown below [27]:

$$P(\boldsymbol{h}|\boldsymbol{O}) = \frac{1}{Z(\boldsymbol{O})} \exp \sum_{t=1}^{T} \{ \sum_{k=1}^{K} \lambda_k T_k(h_t, h_{t-1}) + \sum_{m=1}^{M} \mu_m E_m(h_t, \boldsymbol{Y}_t) \} \tag{6.1}$$

where N represents the total number of process modes, and Q and T represent the dimension of data and total number of samples, respectively. $\boldsymbol{Y}_t$ is composed of the data required for modeling at time instant t. The normalization term $Z(\boldsymbol{O})$ is obtained by marginalizing the process mode sequence $h_{1:T}$.

The function sets $\{T_k\}_{k=1}^{K}$ and $\{E_m\}_{m=1}^{M}$ are called feature functions, which can be discrete or continuous values. The selection of feature functions is generally based on the specific nature of the problems [102]. For the purpose of process mode diagnosis, the related feature functions are normally selected as:

$$T_k(h_t, h_{t-1}) = \begin{cases} 1 \quad \text{if} \quad h_{t-1} = l_1 \quad \text{and} \quad h_t = l_2 \\ 0 \qquad\qquad \text{otherwise} \end{cases} \tag{6.2}$$

where $l_1, l_2 = 1, 2, ..., N$ represent the process mode identities of sampling instants $t-1$ and t, respectively. When there exists a process mode transition from l_1 to l_2, the feature function T_k will be set as 1 to represent the activation of such transition in CRF model. Also, for modeling the sequential observation data dependency, the following feature functions are commonly formulated:

$$E_m(h_t, \boldsymbol{Y}_t) = E_m(h_t, O_t, O_{t-1}, ..., O_{t-d_w+1}) \tag{6.3}$$

where $\boldsymbol{Y}_t = [O_t, O_{t-1}, \cdots, O_{t-d_w+1}]$, with d_w being a suitably chosen window length reflecting strength of the observation dependency.

In summary, the unknown parameters are $\boldsymbol{\Lambda} = \{\lambda_k\}_{k=1}^K$ and $\boldsymbol{\mathcal{M}} = \{\mu_m\}_{m=1}^M$, which can be calculated by maximizing the conditional likelihood function in Equation (6.1) [93]. For detail information regarding the application of CRF for process monitoring, readers are referred to Fang et al. [117]. Once Equation (6.1) is determined, mode ($\boldsymbol{h}$) diagnosis can be performed based on the observed data ($\boldsymbol{O}$).

6.3 SCRF for Process Mode Diagnosis in Multiple Operating Conditions

6.3.1 Problem Statement

Chemical processes often operate under different process modes reflecting the process health status, and the switching between the process modes can be modeled by certain rules with a probability. For example, the process mode may be classified as normal, abnormal and failure when the process is subject to faults. When a process is in the abnormal mode, it can have a certain probability to go failure mode as well as a certain probability to return to normal mode. These probabilities may be described by a probabilistic model such as Markov chain. Meanwhile, the process can also operate in different operating conditions, such as, low throughput and high throughput. Note that two concepts used in this chapter, process mode and operating condition, are distinct as explained. Under different operating conditions, the switching rules between different process modes can be different. For example, the Markov chain model for mode switching at low throughput can be different from that at high throughput as illustrated in Fig. 6.1. Process mode diagnosis problem in this case becomes more difficult and conventional diagnosis approaches may result in ambiguous inferences, as there are multiple sources of switching rules. Therefore a more sophisticated model structure is needed under this circumstance. In this work, we employ a SCRF approach to address this issue.

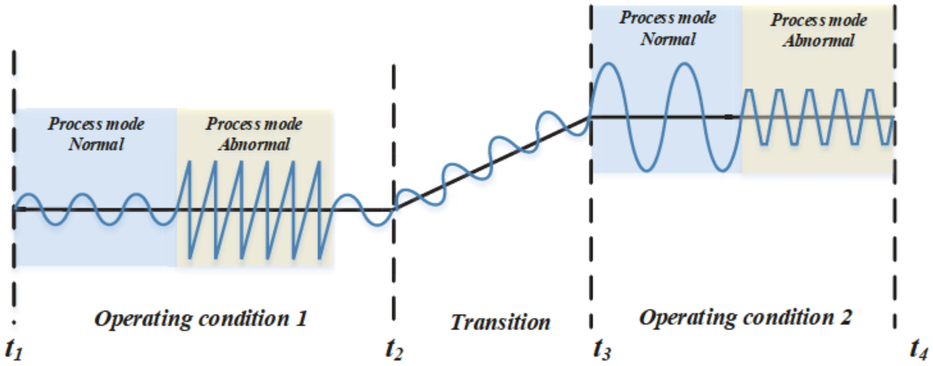


Figure 6.1: An illustration of the relation between the operating conditions and process modes

The whole process is assumed to operate in P different operating conditions, i.e., $\boldsymbol{I} = [I_1, I_2, ..., I_t, ..., I_T]$, where $I_t \in \{1, 2, \cdots P\}$. So a data point at t has two attributes: (i) its process mode h_t and (ii) its operating condition I_t. It is assumed that the process mode of the system at a particular time instant t is associated with a scheduling variable

$\boldsymbol{S} = [S_1, S_2, \cdots, S_T]$, but with uncertainties. Mathematically, we are interested in modeling $P(\boldsymbol{h}|\boldsymbol{O}, \boldsymbol{S})$, namely, the probability of the process mode given the process data to determine the process mode sequence $h_{1:T}$, regardless of varying operating conditions.

6.3.2 SCRF Model Formulation

Building on the basis of the LCCRF, a graphical illustration of the proposed SCRF is given in Fig. 6.2, where an additional operating condition layer is considered. The multiple LCCRF models are allowed to switch between each other when the change of operating conditions occurs. Further, we consider a scheduling variable S_t that reflects the operating condition with uncertainties, modeled by the following equation:

$$P(I_t = i|S_t) = \frac{\exp[-\frac{(S_t - S_i)^2}{2\sigma_i^2}]}{\sum_{p=1}^{P} \exp[-\frac{(S_t - S_p)^2}{2\sigma_p^2}]} \tag{6.4}$$

where S_i denotes the i^{th} fixed operating point and σ_i is the validity width of the scheduling variable in the i^{th} operating condition.

By referring to the system identification of linear parameter varying models [143], the above probabilistic representation in Equation (6.4) is adopted. Here, the prior probability of the current operating condition I_t is governed by the scheduling variable S_t. Industrial processes typically operate at several fixed operating conditions with occasional transitions between each other. Once the process deviates from its current operating condition, the exponential term in the numerator of Equation (6.4) becomes smaller and therefore the prior probability that indicates the process staying in the same operating condition decreases. While in the transition period between different operating conditions, by using the above priors, the system will be represented by a mixture model with the characteristics of different operating conditions. Usually in application scenarios, the operating conditions are determined beforehand to meet the desired product quality, hence, the fixed operating condition values $\{S_p\}_{p=1}^{P}$ are known in advance, whereas the validity variables $\{\sigma_p\}_{p=1}^{P}$ need to be estimated from the data.

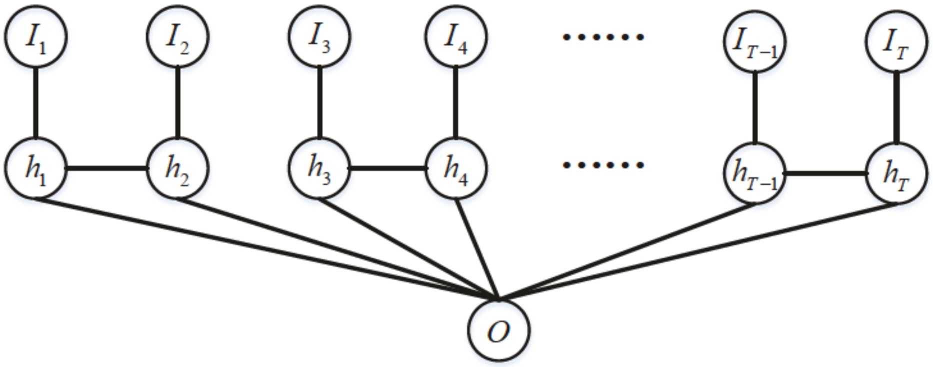


Figure 6.2: A graphical illustration of the proposed SCRF model. In this case, $I_2 \neq I_3$, i.e., it is assumed that there is no transition between h_2 and h_3, and between h_4 and h_5, etc.

On the basis of the SCRF structure, $P(\boldsymbol{h}|\boldsymbol{O},\boldsymbol{I})$ has the following form:

$$P(\boldsymbol{h}|\boldsymbol{O},\boldsymbol{I}) = \frac{1}{Z(\boldsymbol{I},\boldsymbol{O})}\exp\sum_{t=1}^{T}\{\sum_{k=1}^{K}\lambda_k T_k(h_t,h_{t-1},I_t,I_{t-1}) + \sum_{m=1}^{M}\mu_m E_m(h_t,\mathbf{Y}_t,I_t)\} \tag{6.5}$$

with the normalization term $Z(\boldsymbol{I},\boldsymbol{O})$ as:

$$Z(\boldsymbol{I},\boldsymbol{O}) = \sum_{h'_{1:T}}\exp\sum_{t=1}^{T}\{\sum_{k=1}^{K}\lambda_k T_k(h'_t,h'_{t-1},I_t,I_{t-1}) + \sum_{m=1}^{M}\mu_m E_m(h'_t,\mathbf{Y}_t,I_t)\} \tag{6.6}$$

where the process mode sequence $h'_{1:T}$ for enumeration is used to differentiate from the real process mode sequence $h_{1:T}$.

Compared with the feature functions defined in Equations (6.2) - (6.3), the feature functions in SCRF model are redefined by adding conditions $I_t = I_{t-1} = i$ and $I_t = i$ to

$T_k(h_t, h_{t-1})$ and $E_m(h_t, \boldsymbol{Y_t})$, respectively, as follows:

$$T_k(h_t, h_{t-1}, I_t, I_{t-1}) = \begin{cases} 1 \quad \text{if} \quad h_{t-1} = l_1,\ h_t = l_2,\ I_t = I_{t-1} = i \\ 0 \qquad\qquad\qquad \text{otherwise} \end{cases} \tag{6.7}$$

$$E_m(h_t, \mathbf{Y}_t, I_t = i) = [E_{m_{1,1}}, E_{m_{1,2}}, \cdots, E_{m_{d_w,1}}, E_{m_{d_w,2}}]^T$$

which means when $I_t \neq I_{t-1}$, no process mode switching is considered. The elements $E_{m_{\tau,1}}(h_t, O_{t-\tau+1}, I_t)$ and $E_{m_{\tau,2}}(h_t, O_{t-\tau+1}, I_t)$ in Equation (6.7) are taken in linear and quadratic forms, respectively, which can be employed to sufficiently describe the relevant statistics of a normally distributed dataset [42, 94]. For simplicity, in the following contents, the notations T_k and E_m will be employed to denote the SCRF feature functions.

6.3.3 Parameter Estimation Using EM Algorithm

In this section, the parameters of the proposed SCRF model are estimated by means of the EM algorithm, as the direct MLE to estimate the parameters is intractable due to existence of hidden variables. The observed dataset is denoted as $D_o = \{O_{1:T}, S_{1:T}, h_{1:T}\}$, and the latent dataset is represented as $D_m = \{I_{1:T}\}$. In the E-step, the conditional expectation of the joint log-likelihood function in presence of latent variable, known as Q-function, is formulated, and the derived Q-function is maximized in the M-step. The E-step and M-step are performed iteratively to ensure the increase of the log-likelihood function corresponding to the complete data until convergence [144, 145].

6.3.3.1 E-step

The Q-function, which is expected the log likelihood function with respect to the missing data or hidden variables, is formulated as:

$$Q(\Theta|\Theta^{(old)}) = E_{D_m|(D_o;\Theta^{(old)})}\{\log P(D_o, D_m|\Theta)\} \tag{6.8}$$

where Θ represents all the unknown parameters, i.e., $\Theta = \{\mathbf{\Lambda}, \boldsymbol{\mathcal{M}}, \mathbf{\Sigma}\}$, and $\Theta^{(old)}$ represents the estimated values of the unknown parameters from the previous iteration.

Based on the chain rule, the above Q-function can be further factorized. Moreover, given the profile of the operating conditions $I_{1:T}$, the conditional probability distribution of $h_{1:T}$ is independent of the scheduling variable profile $S_{1:T}$. Given the scheduling variable, the conditional probability distribution of the operating conditions is independent of $O_{1:T}$. Therefore, the final expression of the Q-function is formulated as follows:

$$\begin{aligned} Q(\Theta|\Theta^{(old)}) &= \sum_{I_{1:T}} \tau_{I_{1:T}}^{(old)} \log P(h_{1:T}|I_{1:T}, O_{1:T}; \Theta) + \sum_{t=1}^{T} \sum_{I_t} \tau_{I_t}^{(old)} \log P(I_t|S_t; \Theta) \\ &= Q_1(\mathbf{\Lambda}, \boldsymbol{\mathcal{M}}) + Q_2(\mathbf{\Sigma}) \end{aligned} \tag{6.9}$$

Here, the first log likelihood term has been defined in Equation (6.5), which can be treated as an overall conditional probability of the SCRF model. For simplicity, the posterior probabilities in the Q-function are defined as $P(\boldsymbol{X}|D_o; \Theta^{(old)}) = \tau_{\boldsymbol{X}}^{(old)}$, where $\boldsymbol{X}$ represents any variable or sequence, whose posterior probability is to be determined.

While maximizing the Q-function, the most challenging problem is to calculate the first log likelihood term, as it requires enumeration. As different combinations of $I_{1:T}$ could result in different factorizations of the LCCRFs, a propagation algorithm is utilized in this work for an efficient enumeration. In the proceeding contents, factorized formulation of the term Q_1 in Equation (6.9) is presented.

Based on the SCRF definition provided in Equation (6.5), the first term in Equation

(6.9) can be expanded as below:

$$\begin{aligned}Q_1(\mathbf{\Lambda}, \boldsymbol{\mathcal{M}}) =&\sum_{t=1}^{T}\sum_{I_{t-1},I_t}\tau^{(old)}_{I_{t-1},I_t}\{\sum_{k=1}^{K}\lambda_k T_k\} + \sum_{t=1}^{T}\sum_{I_t}\tau^{(old)}_{I_t}\{\sum_{m=1}^{M}\mu_m E_m\}- \\ &\sum_{I_{1:T}}\tau^{(old)}_{I_{1:T}}\log Z(\boldsymbol{I}, \boldsymbol{O}; \Theta)\end{aligned} \tag{6.10}$$

To calculate $\sum_{I_{1:T}}\tau^{(old)}_{I_{1:T}}\log Z(\boldsymbol{I}, \boldsymbol{O}; \Theta)$, a forward propagation strategy is configured. First, a series of intermediate functions are defined at the sampling instant t for simplicity, as:

$$\varphi_t(h_t, h_{t-1}, I_t, I_{t-1}, \boldsymbol{Y}_t) \overset{def}{=} \exp\{\sum_{k=1}^{K}\lambda_k T_k + \sum_{m=1}^{M}\mu_m E_m\} \tag{6.11}$$

When $I_t \neq I_{t-1}$, the feature function $T_k(h_t, h_{t-1}, I_t, I_{t-1})$ will be evaluated to zero, making the process mode sequence after time instant t independent of the previous time instants, thereby facilitating the factorization of $\sum_{I_{1:T}}\tau^{(old)}_{I_{1:T}}\log Z(\boldsymbol{I}, \boldsymbol{O}; \Theta)$. Hence, at each sampling instant t, a sequence of forward intermediate variables $\{\alpha_{t,n}\}_{n=2}^{t}$ are defined as:

$$\{\alpha_{t,n}(h_t, I_t)\}_{n=2}^{t} \overset{def}{=} \sum_{h_{t-n+1:t-1}}\varphi_1(h_{t-n+1}, I_{t-n+1}, \boldsymbol{Y}_{t-n+1})\prod_{t'=t-n+2}^{t}\varphi_{t'}(h_{t'}, h_{t'-1}, I_{t'}, I_{t'-1}, \boldsymbol{Y}_{t'}) \tag{6.12}$$

where $\alpha_{t,1}(h_t, I_t) = \varphi_1(h_t, I_t, \boldsymbol{Y}_t)$ stands for the feature functions corresponding to the initial process modes without any switching.

In the next sampling instant, if $I_t = I_{t+1}$, the forward intermediate variable $\{\alpha_{t+1,n+1}\}_{n=1}^{t}$ can be calculated via the following forward propagation rule:

$$\alpha_{t+1,n+1}(h_{t+1}, I_{t+1}) = \sum_{h_t}\varphi_{t+1}(h_{t+1}, h_t, I_{t+1}, I_t, \boldsymbol{Y}_{t+1})\cdot\alpha_{t,n}(h_t, I_t) \tag{6.13}$$

Based on the forward propagation results, the following function set is defined:

$$q_{t,n}(I_t) \ = \gamma^{(old)}_{I_{t-n+1:t}} \log \sum_{h_t} \alpha_{t,n}(h_t, I_t) \tag{6.14}$$

where $\gamma^{(old)}_{I_{t-n+1:t}}$ can be derived as:

$$\begin{aligned}
\gamma^{(old)}_{I_{t-n+1:t}} =& \frac{\sum_{j_1=1,j_1\neq i}^{P}\sum_{j_2=1,j_2\neq i}^{P} P(h_{1:T}, I_{t-n+1} =}{\sum_{i=1}^{P}\sum_{j_1=1,j_1\neq i}^{P}\sum_{j_2=1,j_2\neq i}^{P} P(h_{1:T}, I_{t-n+1} =} \\
& \frac{\cdots = I_t = i, I_{t-n} = j_1, I_{t+1} = j_2|O_{1:T}, S_{1:T};\Theta^{(old)})}{\cdots = I_t = i, I_{t-n} = j_1, I_{t+1} = j_2|O_{1:T}, S_{1:T};\Theta^{(old)})} \\
=& \frac{\sum_{j_1=1,j_1\neq i}^{P}\sum_{j_2=1,j_2\neq i}^{P} P(h_{t-n:t+1}|I_{t-n+1:t} = i,}{\sum_{i=1}^{P}\sum_{j_1=1,j_1\neq i}^{P}\sum_{j_2=1,j_2\neq i}^{P} P(h_{t-n:t+1}|I_{t-n+1:t} = i,} \\
& \frac{I_{t-n} = j_1, I_{t+1} = j_2, \boldsymbol{Y}_{t-n:t+1};\Theta^{(old)}) \cdot P(I_{t-n+1:t} = i,}{I_{t-n} = j_1, I_{t+1} = j_2, \boldsymbol{Y}_{t-n:t+1};\Theta^{(old)}) \cdot P(I_{t-n+1:t} = i,} \\
& \frac{I_{t-n} = j_1, I_{t+1} = j_2|S_{t-n:t+1};\Theta^{(old)})}{I_{t-n} = j_1, I_{t+1} = j_2|S_{t-n:t+1};\Theta^{(old)})}
\end{aligned} \tag{6.15}$$

Finally, $\sum_{I_{1:T}} \tau^{(old)}_{I_{1:T}} \log Z(\boldsymbol{I}, \boldsymbol{O}; \Theta)$ is derived like:

$$\sum_{I_{1:T}} \tau^{(old)}_{I_{1:T}} \log Z(\boldsymbol{I}, \boldsymbol{O}; \Theta) = \sum_{t=1}^{T} \sum_{I_t} R_t(I_t) \tag{6.16}$$

with $R_t(I_t) = \sum_{n=1}^{t} q_{t,n}(I_t)$. Consequently, the Q-function in Equation (6.9) becomes:

$$\begin{aligned}
& Q(\Theta|\Theta^{(old)}) = \sum_{t=1}^{T}\{ \sum_{I_{t-1},I_t} \tau^{(old)}_{I_{t-1},I_t} \sum_{k=1}^{K} \lambda_k T_k + \sum_{I_t} \tau^{(old)}_{I_t} \\
& \sum_{m=1}^{M} \mu_m E_m - \sum_{I_t} R_t(I_t) + \sum_{I_t} \tau^{(old)}_{I_t} \log P(I_t|S_t;\Theta)\}
\end{aligned} \tag{6.17}$$

The posterior probability $\tau^{(old)}_{I_{t-1},I_t}$ in Equation (6.17) can be derived as:

$$\begin{aligned}\tau^{(old)}_{I_{t-1},I_t} =& \frac{P(h_{t:t-1}|I_t = i, I_{t-1} = j, \boldsymbol{Y}_{t:t-1}; \Theta^{(old)})}{\sum_{i=1}^{P}\sum_{j=1}^{P} P(h_{t:t-1}|I_t = i, I_{t-1} = j, \boldsymbol{Y}_{t:t-1}; \Theta^{(old)})} \\ & \frac{\cdot P(I_t = i|S_t; \Theta^{(old)}) \cdot P(I_{t-1} = j|S_{t-1}; \Theta^{(old)})}{\cdot P(I_t = i|S_t; \Theta^{(old)}) \cdot P(I_{t-1} = j|S_{t-1}; \Theta^{(old)})}\end{aligned} \tag{6.18}$$

where the conditional probability $P(h_{t:t-1}|I_t = i, I_{t-1} = j, \boldsymbol{Y}_{t:t-1}; \Theta^{(old)})$ is equivalent to a local LCCRF model. The posterior probability $\tau^{(old)}_{I_t}$ can be calculated in a similar way.

After the Q-function formulation in the E-step, the unknown parameters Θ will be estimated in the M-step by maximizing the formulated Q-function.

6.3.3.2 M-step

This step is to determine the unknown parameters by maximizing the Q-function derived early. In this work, the unknown weighting parameters of the SCRF model are computed by the L-BFGS optimization algorithm [105]. In order to increase the computational efficiency of the gradients, a backward propagation strategy is proposed. Similar to the forward propagation strategy, at each time instant t, a series of backward intermediate variables $\{\beta_{t,n}\}_{n=2}^{T-t+1}$ are introduced:

$$\{\beta_{t,n}(h_t, I_t)\}_{n=2}^{T-t+1} \overset{def}{=} \sum_{h_{t+1:t+n-1}} \prod_{t'=t}^{t+n-1} \varphi_{t'}(h_{t'}, h_{t'-1}, I_{t'}, I_{t'-1}, \boldsymbol{Y}_{t'}) \tag{6.19}$$

where $\beta_{t,1}(h_t, I_t)$ is defined as $\varphi_1(h_t, I_t, \boldsymbol{Y}_t)$.

Similar to the forward propagation strategy, an intermediate variable $\{\beta_{t-1,n+1}\}_{n=1}^{T-t+1}$ with $I_{t-1} = I_t$ is derived as shown below:

$$\beta_{t-1,n+1}(h_{t-1}, I_{t-1}) = \sum_{h_t} \varphi_t(h_t, h_{t-1}, I_t, I_{t-1}, \boldsymbol{Y}_t) \cdot \beta_{t,n}(h_t, I_t) \tag{6.20}$$

Gradient calculation of Q-function with respect to unknown parameters yields the following terms:

$$
\begin{aligned}
\frac{\partial Q(\Theta|\Theta^{(old)})}{\partial \lambda_k} &= \sum_{t=1}^{T}\sum_{I_t, I_{t-1}} \tau^{(old)}_{I_t, I_{t-1}} T_k - \sum_{t=1}^{T}\sum_{I_t} \frac{\partial R_t(I_t)}{\partial \lambda_k} \\
\frac{\partial Q(\Theta|\Theta^{(old)})}{\partial \mu_m} &= \sum_{t=1}^{T}\sum_{I_t} \tau^{(old)}_{I_t} E_m - \sum_{t=1}^{T}\sum_{I_t} \frac{\partial R_t(I_t)}{\partial \mu_m}
\end{aligned}
\tag{6.21}
$$

where the partial derivative terms in right hand side of Equation (6.21) are given as follows:

$$
\frac{\partial R_t(I_t)}{\partial \lambda_k} = \sum_{n=1}^{t} \gamma^{(old)}_{I_{t-n+1:t}} \sum_{t'=t-n+1}^{t} \sum_{h'_{t'}, h'_{t'-1}} T_k \cdot \frac{\alpha_{t'-1, n-t+t'-1}(h'_{t'-1}, I_{t'-1}) \cdot \beta_{t', n-t+t'}(h'_{t'}, I_{t'})}{\sum_{h'_t} \alpha_{t,n}(h'_t, I_t)}
\tag{6.22}
$$

$$
\frac{\partial R_t(I_t)}{\partial \mu_m} = \sum_{n=1}^{t} \gamma^{(old)}_{I_{t-n+1:t}} \sum_{t'=t-n+1}^{t} \sum_{h'_{t'}} E_m \cdot \frac{\alpha_{t', n-t+t'}(h'_{t'}, I_{t'}) \cdot \beta_{t', n-t+t'}(h'_{t'}, I_{t'})}{\sum_{h'_t} \alpha_{t,n}(h'_t, I_t)}
\tag{6.23}
$$

Then the unknown parameters λ_k and μ_m can be updated by the L-BFGS approach. The validity variables σ_i are estimated by performing the following optimization:

$$
\begin{aligned}
\sigma_i^{(new)} &= \underset{\sigma_i}{\text{argmax}}\ \ Q_2(\sigma_i) \\
s.t. \quad & \sigma_{i,min} \le \sigma_i \le \sigma_{i,max}
\end{aligned}
\tag{6.24}
$$

where the two parameters $\sigma_{i,min}$ and $\sigma_{i,max}$ represent the lower and upper bounds of the unknown parameter σ_i, respectively. Many existing nonlinear optimization algorithms can be selected to solve this problem, for example, the sequential QP algorithm [128] and the nonlinear interior point local optimization algorithm [37], etc..

6.3.4 Simplified SCRF Parameter Estimation

Throughout the entire steps of EM algorithm, enumeration of $\sum_{I_{1:T}} \tau^{(old)}_{I_{1:T}} \log Z(\boldsymbol{I}, \boldsymbol{O}; \Theta)$ demands high computational complexity. Hence, a simplified algorithm is presented for parameter estimation considering the following facts: (i) the operating conditions can be retrieved from the scheduling variable with uncertainties, and (ii) the operating conditions are smoothly transferred between each other. Hence, for the sake of computational efficiency, we can consider the stationary and transition periods of the operating conditions separately. During a specific operating period, the system is naturally considered to operate in the same operating condition, and while during transition periods between two adjacent operating conditions, the system will transit from one operating condition to another. As illustrated in Fig. 6.3, the entire time segment can be decomposed to two stationary operation periods, namely, stationary periods 1 and 2, and one transition period between them. The transition period can be further decomposed to two segments. It is assumed that the first segment has similar properties as the stationary period 1, and the second segmentation has similar properties as the stationary period 2. Let d_{tr} be the duration of the first portion of the tr^{th} transition period, which is not known and hence, can be considered as a latent variable. In this case, the noises during the stationary periods and two half transition periods are assumed to be independent of each other.

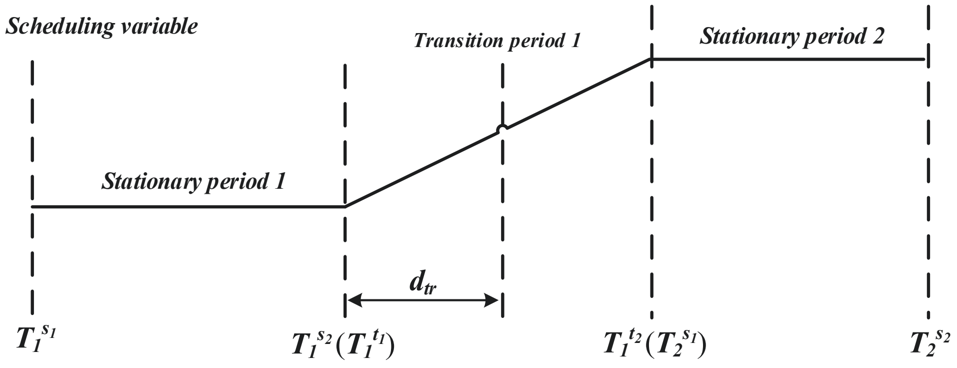


Figure 6.3: An illustration of the stationary and transition periods indicated by the scheduling variable

As a result, the entire operating condition sequence can be decomposed into small segments based on the two stationary periods and two transition periods. Therefore, the proposed SCRF framework can be simplified and the unknown parameters are estimated by the EM algorithm, as presented in the following.

For the transition segment decomposition, the latent variable d_{tr} is considered to follow a normal distribution, i.e., $d_{tr} \sim N(\mu_d, \sigma_d^2)$, with the unknown parameters μ_d and σ_d.

As a result, the unknown parameter set becomes $\Theta = \{\mathbf{\Lambda}, \boldsymbol{\mathcal{M}}, \mathbf{\Sigma}, \mu_d, \sigma_d\}$, and the objective Q-function in the EM algorithm can be rearranged into the following form:

$$\begin{aligned} Q(\Theta|\Theta^{(old)}) &= \sum_{d_{1:T_r}} \sum_{I_{1:T}} \tau_{I_{1:T}}^{(old)} P(d_{1:Tr}|I_{1:T}, h_{1:T}, O_{1:T}, S_{1:T}; \Theta^{(old)}) \log P(h_{1:T}|d_{1:T_r}, I_{1:T}, O_{1:T}; \Theta) \\ &\quad + \sum_{t=1}^{T} \sum_{I_t} \tau_{I_t}^{(old)} \cdot \log P(I_t|S_t; \Theta) + \sum_{t_r=1}^{T_r} \sum_{d_{t_r}} \tau_{d_{t_r}}^{(old)} \log P(d_{t_r}; \Theta) \\ &= Q_1(\mathbf{\Lambda}, \boldsymbol{\mathcal{M}}) + Q_2(\mathbf{\Sigma}) + Q_3(\mu_d, \sigma_d) \end{aligned} \tag{6.25}$$

where T_r represents the total number of transition periods.

Based on the scheduling variable, the prior probability of current operating conditions can be estimated according to Equation (6.4). By calculating the prior probability $P(I_t|S_t)$, several probabilistic thresholds can be set to determine the segmentations of stationary and transition periods. Assume there are S_d stationary periods and T_r transition periods in total, and the time segments of stationary and transition periods are represented by $[T_{s_d}^{s_1},\ T_{s_d}^{s_2}], s_d = 1, \cdots, S_d$ and $[T_{t_r}^{t_1},\ T_{t_r}^{t_2}], t_r = 1, \cdots, T_r$, respectively. Therefore, $Q_1(\mathbf{\Lambda}, \boldsymbol{\mathcal{M}})$

can be derived as shown in the following:

$$\begin{aligned}Q_1(\mathbf{\Lambda}, \boldsymbol{\mathcal{M}}) = &\sum_{t=1}^{T}\sum_{I_t}\tau_{I_t}^{(old)}\sum_{m=1}^{M}\mu_m E_m + \sum_{s_d=1}^{S_d}\sum_{t=T_{s_d}^{s1}}^{T_{s_d}^{s2}}\sum_{i=1}^{P}\tau^{(old)}_{I_{T_{s_d}^{s1}:T_{s_d}^{s2}}=i}\\ &\cdot\sum_{k=1}^{K}\lambda_k T_k + \sum_{t_r=1}^{T_r}\sum_{d_{t_r}}\{\sum_{t=T_{t_r}^{t1}}^{T_{t_r}^{t1}+d_{t_r}}\sum_{i=1}^{P}\tau^{(old)}_{I_{T_{t_r}^{t1}:T_{t_r}^{t1}+d_{t_r}=i}}P(d_{t_r}|I_{1:T}, C_{obs};\\ &\Theta^{(old)})\sum_{k=1}^{K}\lambda_k T_k + \sum_{t=T_{t_r}^{t1}+d_{t_r}}^{T_{t_r}^{t2}}\sum_{j=1,j\neq i}^{P}\tau^{(old)}_{I_{T_{t_r}^{t1}+d_{t_r}:T_{t_r}^{t2}}=j}P(d_{t_r}|I_{1:T},\\ &C_{obs};\Theta^{(old)})\sum_{k=1}^{K}\lambda_k T_k\} - \{\sum_{s_d=1}^{S_d}\sum_{I_{T_{s_d}^{s2}}}q_{T_{s_d}^{s2},n_{s_d}}(I_{T_{s_d}^{s2}}) + \sum_{t_r=1}^{T_r}\sum_{d_{t_r}}\\ &P(d_{t_r}|I_{1:T}, C_{obs};\Theta^{(old)})(\sum_{I_{T_{t_r}^{t_1}+d_{t_r}}}q_{T_{t_r}^{t_1}+d_{t_r},d_{t_r}}(I_{T_{t_r}^{t_1}+d_{t_r}})\\ &+\sum_{I_{T_{t_r}^{t_2}}}q_{T_{t_r}^{t_2},n_{t_r}}(I_{T_{t_r}^{t_2}}))\}\end{aligned} \tag{6.26}$$

where n_{s_d} and n_{t_r} represent the lengths of the s_d^{th} stationary period and the second half of the t_r^{th} transition period, respectively, which can be calculated as:

$$n_{s_d} = T_{s_d}^{s_2} - T_{s_d}^{s_1} + 1 \qquad n_{t_r} = T_{t_r}^{t_2} - T_{t_r}^{t_1} - d_{t_r} \tag{6.27}$$

In the E-step, the posterior probability $P(d_{t_r}|I_{1:T}, C_{obs};\Theta^{(old)})$ needs to be calculated as shown below:

$$\begin{aligned}P(d_{t_r}|I_{1:T}, C_{obs};\Theta^{(old)}) &= P(d_{t_r}|I_{a:b}, h_{a:b}, O_{a:b}, S_{a:b};\Theta^{(old)})\\ &= \frac{P(h_{a:b}|d_{t_r}, I_{a:b}, O_{a:b};\Theta^{(old)})P(I_{a:b}|d_{t_r}, S_{a:b};\Theta^{(old)})}{\sum_{d_{t_r}}P(h_{a:b}|d_{t_r}, I_{a:b}, O_{a:b};\Theta^{(old)})P(I_{a:b}|d_{t_r}, S_{a:b};\Theta^{(old)})}\\ &\quad\frac{\cdot P(d_{t_r};\Theta^{(old)})}{\cdot P(d_{t_r};\Theta^{(old)})}\end{aligned} \tag{6.28}$$

where a and b represent the starting and ending points of the t_r^{th} transition period, namely

$T_{t_r}^{t_1}$ and $T_{t_r}^{t_2}$, respectively.

In the M-step, similar numerical optimization needs to be performed to determine the parameters, as illustrated in section 6.3.3.2.

6.3.5 Simplified Online Process Mode Diagnosis Based on the SCRF Model

In the proposed work, after training the SCRF model, the identified model will be employed for the online application. Essentially, the objective is to find the mode that maximizes the probability given all the past information, i.e., $h_t^* = \text{argmax}_{h_t} P(h_t|O_{1:t}, S_{1:t}; \hat{\Theta})$. As the current process mode mainly depends on the most recent operating stationary or transition periods rather than the entire operating sequence, $P(h_t|O_{1:t}, S_{1:t}; \hat{\Theta})$ can be simplified as:

$$\begin{aligned}
P(h_t|O_{1:t}, S_{1:t}; \hat{\Theta}) &= P(h_t|O_{T_s:t}, S_{T_s:t}; \hat{\Theta}) \\
&= \sum_{i=1}^{P} P(h_t, I_{T_s:t} = i|O_{T_s:t}, S_{T_s:t}; \hat{\Theta}) \\
&= \sum_{i=1}^{P} P(h_t|I_{T_s:t} = i, O_{T_s:t}, S_{T_s:t}; \hat{\Theta}) \cdot P(I_{T_s:t} = i|S_{T_s:t}; \hat{\Theta}) \\
&= \sum_{i=1}^{P} \frac{\alpha_{t,t-T_s+1}(h_t, i)}{\sum_{h_t'=1}^{N} \alpha_{t,t-T_s+1}(h_t', i)} \cdot \prod_{t'=T_s}^{t} P(I_{t'} = i|S_{t'}; \hat{\Theta})
\end{aligned} \tag{6.29}$$

where T_s denotes the starting point of the most recent operating condition to which the current time instant t belongs, which can be $T_{s_d}^{s_1}$, $T_{t_r}^{t_1}$ or $T_{t_r}^{t_1} + [\mu_d] + 1$, where $[\cdot]$ represents the round off operator. Finally, the optimal estimation of the current process mode is the one with the highest posterior probability $P(h_t|O_{1:t}, S_{1:t}; \hat{\Theta})$.

6.4 Validations

In this section, two application scenarios are considered to validate the performance of the proposed SCRF algorithm. For comparison purposes, the conventional LCCRF and multiple

HMMs [146] are employed.

6.4.1 Simulation: Two CSTRs in Series

In this section, a simulated system containing two CSTRs in series proposed by Henson et al. [108] is employed for performance evaluation of the proposed algorithm, whose schematic is illustrated in Fig. 6.4. The notations q, C_A and T represent the flowrate, concentration and temperature, respectively. The coolant flows through both reactors to maintain an appropriate reaction temperature. The whole system is operated in open loop condition and the final product concentration in the second reactor, C_{A2}, is the critical PV related to the product quality. Since the feed flow has different concentrations C_{Af}, the final product is generated with concentrations attributing to different qualities. Therefore, it is desirable to detect the concentration quality levels of the feed flow to meet the desired final product requirements. The first principles model of the CSTR system and related parameter settings can be found in Henson et al. [108].

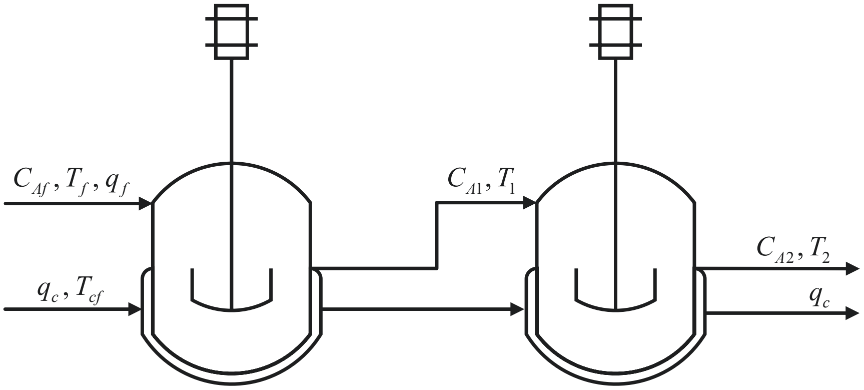


Figure 6.4: The schematic of CSTR in series [108]

As the coolant flowrate q_c has a significant influence on the entire operation, it is selected as the scheduling variable [147]. Three operating conditions, i.e., $q_c = 97L/min$, $102L/min$ and $107L/min$, are considered. The feed flowrate q_f is fixed as $100L/min$, while the feed concentration C_{Af} is manipulated to simulate various process modes, as its fluctuation results in different final product concentrations, thus different products. It is assumed that the feed concentration C_{Af} has three levels, namely high (Mode 1), medium (Mode 2), and low (Mode 3), and the switching between the process modes is simulated by following semi-Markov properties [148]. The feed concentration C_{Af} is simulated with Gaussian white noise contamination in Modes 1 and 3. In Mode 2, C_{Af} is corrupted with auto-correlated Gaussian noise. All the related simulation parameters are summarized in Table 6.1, wherein the variables TR_i and dur_i represent the Markov switching matrix and the duration of each simulated state under the i^{th} operating condition, respectively.

Table 6.1: Parameters of the simulated CSTR system

Operating condition	**Operating condition 1**	**Operating condition 2**	**Operating condition 3**
Semi-Markov switching rule	$TR_1 = \begin{bmatrix} 0.7 & 0.25 & 0.05 \\ 0.25 & 0.5 & 0.25 \\ 0.25 & 0.25 & 0.5 \end{bmatrix}$	$TR_2 = \begin{bmatrix} 0.99 & 0.005 & 0.005 \\ 0.005 & 0.99 & 0.005 \\ 0.005 & 0.005 & 0.99 \end{bmatrix}$	$TR_3 = \begin{bmatrix} 0.6 & 0.1 & 0.3 \\ 0.3 & 0.6 & 0.1 \\ 0.3 & 0.2 & 0.5 \end{bmatrix}$
	$dur_1 = 90$	$dur_2 = 120$	$dur_3 = 60$
Scheduling variable	$S_1 = 97$ $\sigma_1 = 0.6$	$S_2 = 102$ $\sigma_2 = 0.3$	$S_3 = 107$ $\sigma_3 = 0.5$
Process mode	**Feed concentration C_{Af}**	**Feed concentration C_{Af}**	**Feed concentration C_{Af}**
Process mode 1	$mean(C_{Af}) = 1.18$ $std(C_{Af}) = 0.01$	$mean(C_{Af}) = 1.18$ $std(C_{Af}) = 0.01$	$mean(C_{Af}) = 1.15$ $std(C_{Af}) = 0.01$
Process mode 2	$mean(C_{Af}) = 1.12$ $std(C_{Af}) = 0.008$	$mean(C_{Af}) = 1.12$ $std(C_{Af}) = 0.008$	$mean(C_{Af}) = 1.05$ $std(C_{Af}) = 0.008$
Process mode 3	$mean(C_{Af}) = 1.08$ $std(C_{Af}) = 0.005$	$mean(C_{Af}) = 1.02$ $std(C_{Af}) = 0.005$	$mean(C_{Af}) = 1.01$ $std(C_{Af}) = 0.005$

A training dataset with 16600 samples is employed for model development, and a validation dataset with the same length is used for performance evaluation. The product concentrations C_{A1} and C_{A2} are selected as the monitored PVs. The profile of the scheduling variable q_c and the dataset for validation can be found in Fig. 6.5.

To evaluate the performance of the proposed SCRF strategy, LCCRF and the multiple HMMs [146] approaches are compared. For the multiple HMMs strategy, under each operating condition, an HMM model is employed to model the process mode transitions. Comparison of the diagnosis results among the three algorithms is illustrated in Fig. 6.6.

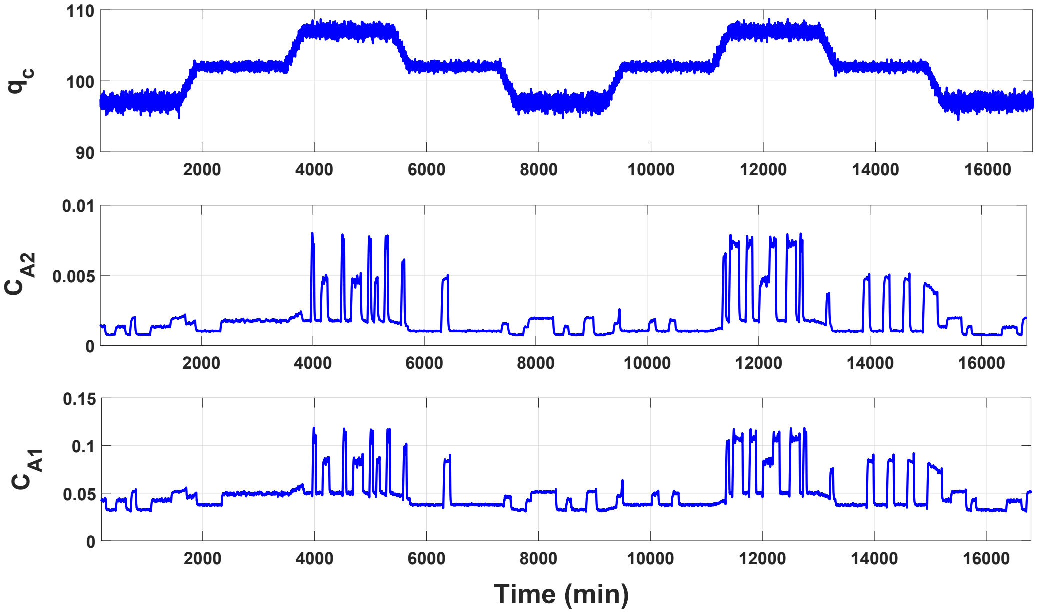


Figure 6.5: The changing profiles of the scheduling variable q_c (L/min) and the process data C_{A1} (mol/L) and C_{A2} (mol/L) in the validation dataset

Correspondingly, in order to quantify the diagnosis accuracy, we also compute the percentage of correctly identified process modes over the complete data sequence. In this case study, the diagnosis accuracies of the SCRF, the LCCRF and the multiple HMMs algorithms are 91.87%, 87.23% and 78.00%, respectively. From this comparison, the multiple HMMs cannot provide better diagnosis results compared with two CRF based algorithms, because the process observations have longer range dependency than that the HMMs can describe. From Fig. 6.6, it can also be found that the diagnosis performance of HMMs gets severely degraded under the operating conditions 1 and 2, where the process observations have relatively small magnitudes and process mode transitions are harder to be detected. In contrast, by modeling long range observation dependency, the CRF based algorithms exhibit better performances than HMMs. Furthermore, from both qualitative and quantitative comparisons, the proposed SCRF algorithm has achieved the best diagnosis performance. The conventional LCCRF can detect most process modes correctly, but it always provides delayed process mode detections. Especially during the operating condition transition periods, the LCCRF model tends to provide wrong diagnosis results compared with the proposed SCRF, because only a single CRF is not sufficient to model the process during the operating condition transition periods. On the contrary, the proposed SCRF algorithm, that employs multiple LCCRF models to differentiate the process modes under different operating conditions, provides the most accurate process mode diagnosis results.

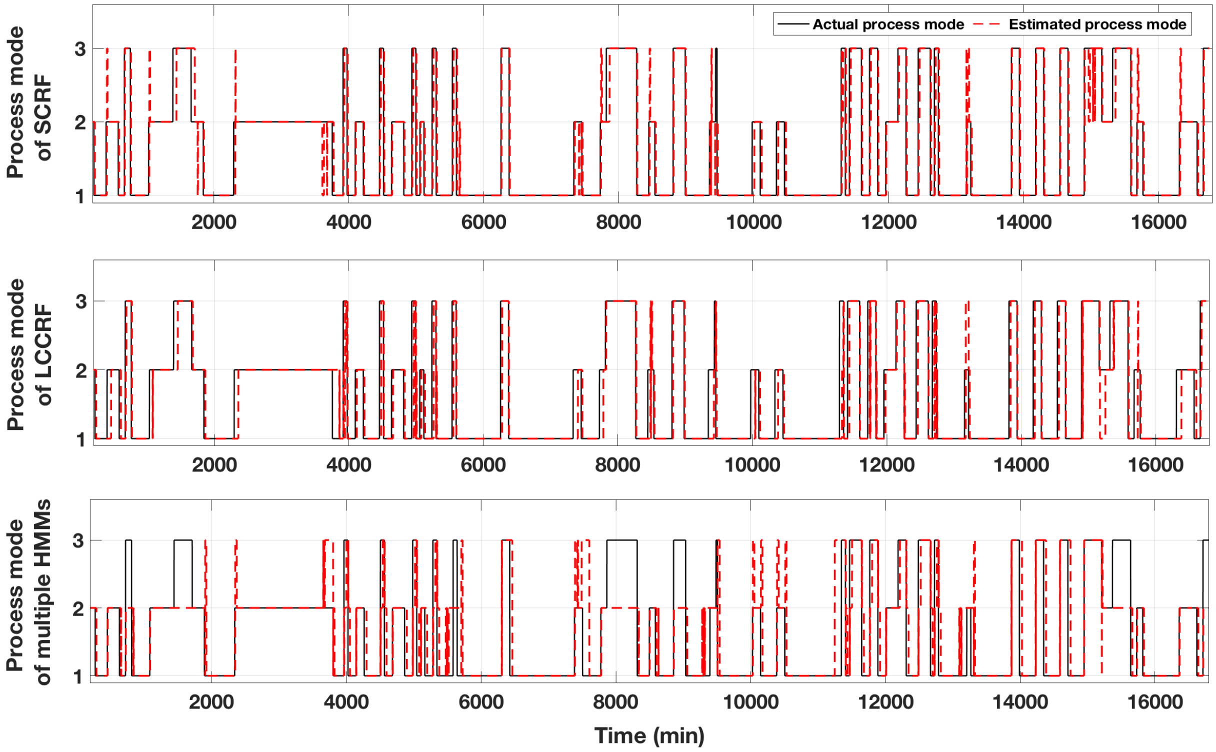


Figure 6.6: The process mode diagnosis performance comparison among the SCRF, the LCCRF and multiple HMMs algorithms in the simulated CSTR process.

6.4.2 Experimental Study through Hybrid Tank System

For further performance evaluations, a pilot-scale experimental study is conducted on a hybrid tank system, whose schematic is illustrated in Fig. 6.7. The whole system is composed of three cylindrical tanks connected in series through six valves, i.e., V_1 - V_4, V_6 and V_8. Outlet valves are provided at the bottom of each tank, i.e., V_5, V_7 and V_9 for tanks 1, 2 and 3, respectively. Water can be fed into the two side tanks via the two identical pumps driven by DC motors, and the feed flowrates can be changed by the users. The three tank levels are measured from the installed level sensors LT_1 - LT_3, individually.

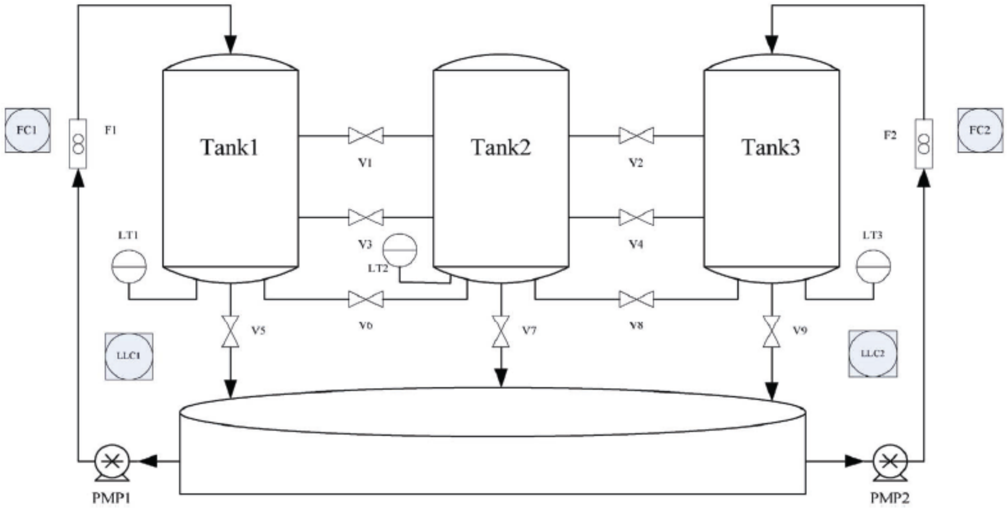


Figure 6.7: The schematic of the experimental hybrid tank system

Manipulation of the feed flowrates of tanks 1 and 3 results in two different operating conditions. It is considered that low-level operating condition 1 occurs when the feed flowrates of tanks 1 and 3 are set around 4.75 and 5.15, respectively. Increasing the two lateral feed flowrates to the values around 6.15 and 6.00 results in high-level operating condition 2. When the hybrid tank system works in the high-level operating condition 2, the tanks 1 and 3 are maintained at a level which is higher than the locations of the junction valves V_1 and V_2. Since V_1 and V_2 are kept open throughout the whole experiment process, once the water levels exceed the levels of V_1 and V_2, we consider that operating condition has changed. For the purpose of simulating different process modes, status of the two lower junction valves V_3 and V_4 are changed from open to closed, simultaneously. When V_3 and V_4 become closed, there is a chance that water might overflow in both lateral tanks, especially around the high-level operating condition 2. Therefore, the abnormal process mode is defined when V_3 and V_4 are closed, and the process mode is assumed to be normal when V_3 and V_4 are open. The left tank feed flowrate is selected as the scheduling variable in this case, and the water levels of all the three tanks form the process outputs. The parameter settings are shown

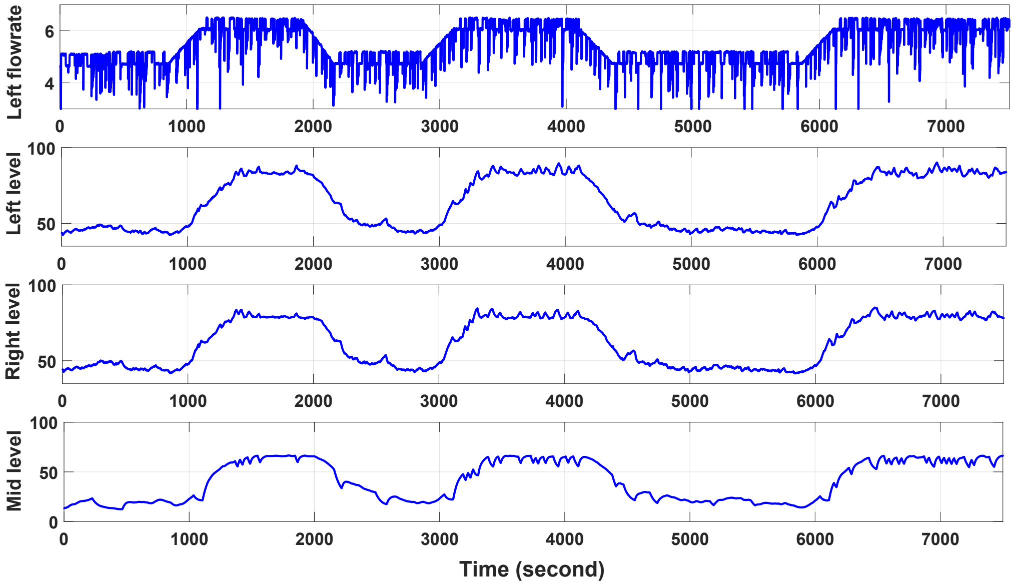


Figure 6.8: The changing profiles of the selected scheduling variable and the tank levels (%) for validation in the hybrid tank experiment

in Table 6.2. Here, the process modes 1 and 2 simulate the normal and abnormal process modes, respectively.

Table 6.2: Parameters of the experimental hybrid tank system

Operating condition	**Operating condition 1**	**Operating condition 2**
Semi-Markov switching rule	$TR_1 = \begin{bmatrix} 0.8 & 0.2 \\ 0.2 & 0.8 \end{bmatrix}$ $dur_1 = 40$	$TR_2 = \begin{bmatrix} 0.75 & 0.25 \\ 0.2 & 0.8 \end{bmatrix}$ $dur_2 = 20$
Scheduling variable left feed flowrate	$S_1 = 4.75$ $\sigma_1 = 0.2$	$S_2 = 6.15$ $\sigma_2 = 0.16$
Process mode	**Process mode 1**	**Process mode 2**
Junction valve condition	V_3, V_4 open	V_3, V_4 closed

Experiment is conducted with sampling interval 1 second and the collected training and validation datasets have a length of 7500 samples. Validation data are presented in Fig. 6.8, wherein the feed flowrate of tank 1, i.e, the scheduling variable, is shown in the first subfigure, and the water levels of tanks 1, 3 and 2 are illustrated in the second to the fourth subfigures, respectively.

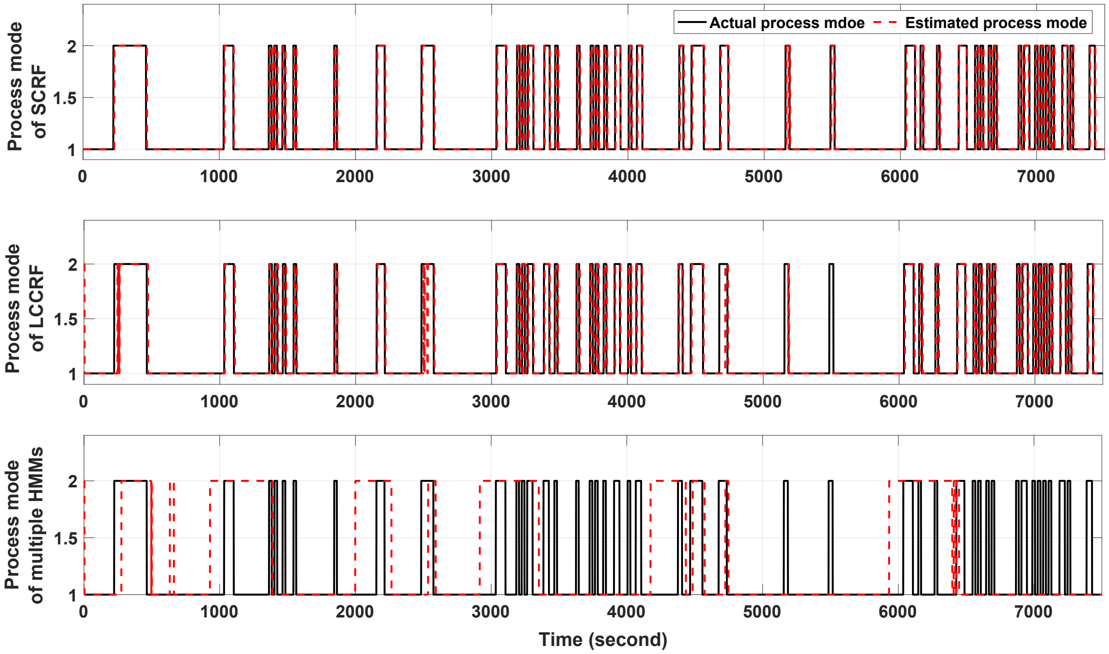


Figure 6.9: The process mode diagnosis performances of the proposed SCRF, the LCCRF and the multiple HMMs approaches.

The proposed SCRF, the LCCRF and the multiple HMMs strategies have been implemented on the validation data, and the real-time diagnosis results are displayed in Fig. 6.9. In this case, the process mode diagnosis accuracies of the SCRF, the LCCRF and the multiple HMMs approaches are 94.56%, 92.33% and 67.19%, respectively, among which the multiple HMMs get the lowest diagnosis accuracy and SCRF achieves the best performance. Due to multiple process disturbances, the level measurements of the three tank system are contaminated with noise, which increases the difficulty to differentiate the abnormal process modes from the noise contaminated observations. By including more information from the observations, the two CRF based algorithms achieve a better diagnosis performance than the multiple HMMs algorithm. Considering the diagnosis performances of the SCRF and LCCRF algorithms, both algorithms can detect the process modes accurately in the high-level operating condition, since the abnormal process mode has more obvious effects on the observations than in the low-level operating condition. However, in the low-level operating condition with smaller level measurements, the abnormal process mode is mixed with process noise and it gets more difficult to be detected. Compared with the SCRF, the LCCRF has delayed or even missing detections in this situation. By involving multiple CRF models, the SCRF can sufficiently capture the process changing properties under multiple operating conditions, therefore more accurate process mode diagnosis performance can be achieved. The more complicated the process is, the more advantages of SCRF can be exhibited compared with the existing approaches.

6.5 Conclusions

In this chapter, a novel SCRF model has been proposed to diagnose the process mode, i.e., normal and abnormal, in real time, regardless of the varying operating conditions. Under the framework of SCRF approach, multiple LCCRF models are identified and switched between each other for process monitoring. In order to increase the computational efficiency, a simplified parameter estimation strategy is proposed for SCRF model identification. The monitoring performance of the SCRF approach has been demonstrated by a CSTR simulation and a hybrid tank system experiment.

Chapter 7

Concluding Remarks and Future Works

In this chapter, conclusions of the above chapters are summarized. The main idea of this thesis is explained and connected to the previous chapters. Finally, the potential future directions are introduced.

7.1 Concluding Remarks

The focus of this thesis is to solve fault detection and diagnosis problems based on both unsupervised hierarchical MSPM approach and supervised CRF algorithm. A large number of fault detection and diagnosis algorithms both unsupervised and supervised have been developed to deal with different practical scenarios during process operation, with the aim of fully excavating the features obtained from the process data.

In Chapter 2, the mathematical backgrounds are explained in details. The modeling, training and inference of CRFs are presented. As alternative solutions to the MLE with latent variables, EM and VB algorithms are introduced and compared to demonstrate their advantages while solving problems with hidden variables.

Chapter 3 proposes an effective hierarchically distributed process monitoring scheme and applies it to solve the early flare event prediction problem for a refinery process, with limited access to process knowledge but large amounts of process data. As a practical large-scale process, there exist several challenges, such as high-dimensionality, nonstationarity, time-varying characteristics, various process changing patterns, high correlations and small number of faulty events, etc.. A hierarchically distributed monitoring framework is developed with a two-layer structure. The bottom layer is composed by monitoring individual units and the top layer is created to integrate the information from the bottom layer. The two layers are embedded with both time-domain MSPM and frequency-domain approaches. Meanwhile, the proposed algorithm is also efficient to solve fault isolation problems by tracing the fault across different units. Both of the time-domain MSPM and the frequency-domain algorithms are tested and compared, and finally the time-domain MSPM algorithm is proven to provide a better solution. The unsupervised approach is appropriate for the problem of the considered flare event prediction as there is essentially no sufficient faulty events available in the data set. The limitation of the proposed approach, similar to all other unsupervised approaches, is that no reference is used.

In the subsequent chapters, as supervised learning approaches, three main theoretical contributions are made based on the LCCRF structure:

- *The marginalized CRF.* The first theoretical contribution based on the CRF model is made in Chapter 4. As a probabilistic discriminative model, the LCCRF is first introduced as a conditional probabilistic counterpart of the HMMs, with higher modeling flexibility and improved process operating mode diagnosis performance. The equivalent conditions of the LCCRFs and HMMs are derived to demonstrate the advantages of LCCRFs. Furthermore, a marginalized CRF model to deal with the missing observation problems is proposed. Because the CRFs involve more complicated observations such as the missing observations, it makes training and inference more complicated. A new forward-backward algorithm is proposed to efficiently solve the training and inference problems of the marginalized CRF model. The performance of the proposed

CRF algorithm has been tested on both simulated and experimental studies, and the superior performance of CRFs over HMMs is demonstrated.

- *The two-stage HCRF.* The second theoretical contribution based on CRF model is made in Chapter 5. In this work, aiming at the problems of feature selection and the online adaption of process changes, a two-stage HCRF structure is proposed and implemented by making full use of the available process measurements. The process observations from different operating modes are first separated and analyzed by a MMHCRF model to select the most relevant variables to detect operating mode changes, known as the first-stage HCRF model. Then based on the outputs of the first-stage HCRF model, the second-stage HCRF model is proposed by including the transitions among different operating modes with a time-varying structure. With the prior knowledge of the second-stage HCRF model, the VB algorithm is employed to solve the unknown model parameters. Briefly, the first-stage HCRF model contributes to determining a set of local classifiers to select most relevant variables, and the second-stage HCRF model conducts an online operating mode diagnosis on the basis of the local classifiers in the first-stage HCRF. The superior performance of the two-stage HCRFs over the conventional algorithm is demonstrated on a numerical case study.

- *The switching CRF.* The third theoretical contribution based on the CRF model is given in Chapter 6. In this work, the process operating mode diagnosis problem for processes with multiple operating conditions is considered. Instead of using only one CRF model for process operating mode diagnosis, a SCRF structure is created to extend unitary LCCRF into multiple LCCRFs, which can be switched between each other according to the changes of the process operating conditions. The process operating conditions are considered to be latent and a scheduling variable is included to infer the potential changes of operating conditions. In this sense, the EM algorithm is used to solve the training problem of the proposed SCRF model. The performances are validated through several case studies.

7.2 Future Works

In this section, potential future directions are summarized.

7.2.1 Feature Dimension Reduction in Probabilistic Discriminative Models

Owing to the direct modeling of the conditional probability, one of the outstanding advantages of the probabilistic discriminative models is the capability to include any features into modeling without need to formulate the explicit distributions of these features. In this way, high dimensional features can be addressed in such a framework with the cost of increased model parameters. However, when some of the features are likely to be correlated with each other as in the MSPM algorithms, it is meaningful to combine the probabilistic counterparts of the MSPM algorithms with the discriminative probabilistic modeling. In this way, the correlations among the raw features can be more precisely addressed and the latent features with lower dimensions can be extracted by the MSPM algorithms. Then in the probabilistic discriminative framework, the latent features are used for further classification.

There are some challenges while dealing with this problem. First, the conditional probability modeling framework increases the modeling complexity and causes the integral of both latent features and the unknown labels harder to address than the probabilistic generative models combined with MSPM algorithms. Second, because of introducing the latent features, the alternatives of standard MLE algorithms, such as EM and VB, need to be further extended for model training. The increased model complexity can make the posterior probabilities of latent variables very difficult to derive. In this sense, more effective inference strategies should be developed based on the designed probabilistic discriminative model structure.

7.2.2 Transfer Learning of the CRFs

In Chapter 6, a SCRF framework has been developed to solve the process operating mode diagnosis problem for the processes with multiple operating conditions. This work is developed by extending a unitary LCCRF model suitable to a single operating condition to multiple LCCRFs to adapt to different operating conditions. To track the change of the operating conditions, a scheduling variable is selected that needs to be available in the SCRF structure. By taking considerations of the facts that some processes might not have a suitable scheduling variable, and the same operating modes in different operating conditions might have high similarity among each other, transfer learning technique is a good choice to make the developed CRF model more general and effective for different operating conditions. The knowledge that CRF learned from one operating condition may be transferred to the other operating conditions, without creating a number of CRF models.

7.2.3 Probabilistic Graphical Model Based Fault-tolerant Control Strategy

In general, fault-tolerant control can be treated as a system that integrates online fault detection and diagnosis, automatic operating condition assessment and the remedial action calculation to compensate the detected faults. In this thesis, the fault detection and diagnosis problems solved by probabilistic graphical model have been discussed, but the actions after fault diagnosis have not been considered. Moreover, not limited to calculating the posterior probabilities, the probabilistic graphical models can also be employed for decision making, such as influence diagrams and Markov decision processes [149], where the decisions are obtained by certain strategies. The rewards of making a specific decision vary according to the states and can be uncertain and model-free. This contributes to a unified fault-tolerant probabilistic structure with advantages of both fault detection and diagnosis and decision making.

7.2.4 Survival Analysis for Remaining Useful Life Prediction

When solving fault detection and diagnosis problems, the conventional MSPM algorithms are used as unsupervised learning algorithms. By selecting a portion of normal operating data, the most suitable MSPM algorithm is employed to build a model, which can be used to detect process abnormalities. However, when targeting on a permanent failure prediction problem, such as motor pump failure, the abnormalities detected by the designed MSPM model can be anything that is different from the normal operations, and might not be the failures that are interested. Unless a unique signature to the final failure is given, otherwise the MSPM algorithms will provide high false positives that are not directly related to the final failure.

To solve this problem, the concept of the time-to-event distribution is introduced and combined with the conventional MSPM algorithms. Survival analysis covers a series of approaches to model the time-to-failure distribution and therefore creates a survival curve for specific process or equipment under monitoring. The survival curve depicts the distribution of survival life which is modeled by integrating the process features into a probabilistic model. By making use of the latent features extracted by MSPM algorithms, the connection between the process abnormalities and the process survival time is established, by which the process remaining useful life can be predicted with the potential likelihood. Moreover, since the individual survival analysis problem can be treated as a probabilistic multi-task classification problem, the probabilistic discriminative models, for example, CRFs, can also be involved for structural modeling for more complicated and correlated features. By this means, more reliable permanent failure prediction algorithms can be developed.

Bibliography

[1] L. H. Chiang, E. L. Russell, and R. D. Braatz. *Fault detection and diagnosis in industrial systems*. Springer Science & Business Media, 2000.

[2] R. Isermann and P. Ballé. Trends in the application of model based fault detection and diagnosis of technical processes. *IFAC Proceedings Volumes*, 29(1):6325–6336, 1996.

[3] R. Isermann. Model based fault detection and diagnosis methods. In *Proceedings of 1995 American Control Conference-ACC'95*, volume 3, pages 1605–1609. IEEE, 1995.

[4] V. Venkatasubramanian, R. Rengaswamy, K. Yin, and S. N. Kavuri. A review of process fault detection and diagnosis: Part I: Quantitative model-based methods. *Computers & Chemical Engineering*, 27(3):293–311, 2003.

[5] N. Sammaknejad. Fault detection and isolation based on hidden Markov models. PhD thesis, University of Alberta, 2015.

[6] V. Venkatasubramanian, R. Rengaswamy, S. N. Kavuri, and K. Yin. A review of process fault detection and diagnosis: Part III: Process history based methods. *Computers & Chemical Engineering*, 27(3):327–346, 2003.

[7] W. A. Shewhart. *Economic control of quality of manufactured product*. Macmillan And Co Ltd, London, 1931.

[8] E. S. Page. Continuous inspection schemes. *Biometrika*, 41(1/2):100–115, 1954.

[9] K. Pearson. LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901.

[10] S. Wold, A. Ruhe, H. Wold, and W. D., III. The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses. *SIAM Journal on Scientific and Statistical Computing*, 5(3):735–743, 1984.

[11] H. Hotelling. Multivariate quality control, illustrated by the air testing of sample bombsights. *Selected Techniques of Statistical Analysis (Ed MWHC Eisenhart, and WA Wallis). New York, NY, USA: Mc-Graw-Hill*, 1947.

[12] S. Yin, S. X. Ding, X. Xie, and H. Luo. A review on basic data-driven approaches for industrial process monitoring. *IEEE Transactions on Industrial Electronics*, 61(11):6418–6428, 2014.

[13] P. Nomikos and J. F. MacGregor. Monitoring batch processes using multiway principal component analysis. *AIChE Journal*, 40(8):1361–1375, 1994.

[14] S. J. Qin and T. J. McAvoy. Nonlinear PLS modeling using neural networks. *Computers & Chemical Engineering*, 16(4):379–391, 1992.

[15] D. Dong and T. J. McAvoy. Batch tracking via nonlinear principal component analysis. *AIChE Journal*, 42(8):2199–2208, 1996.

[16] J. Yu. A nonlinear kernel Gaussian mixture model based inferential monitoring approach for fault detection and diagnosis of chemical processes. *Chemical Engineering Science*, 68(1):506–519, 2012.

[17] N. Sammaknejad, B. Huang, and Y. Lu. Robust diagnosis of operating mode based on time-varying hidden Markov models. *IEEE Transactions on Industrial Electronics*, 63(2):1142–1152, 2015.

[18] T. Chen, E. Martin, and G. Montague. Robust probabilistic PCA with missing data and contribution analysis for outlier detection. *Computational Statistics & Data Analysis*, 53(10):3706–3716, 2009.

[19] W. Ku, R. H. Storer, and C. Georgakis. Disturbance detection and isolation by dynamic principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 30(1):179–196, 1995.

[20] J. H. Cho, J. M. Lee, S. W. Choi, D. Lee, and I. B. Lee. Fault identification for process monitoring using kernel principal component analysis. *Chemical Engineering Science*, 60(1):279–288, 2005.

[21] M. Kano, S. Tanaka, S. Hasebe, I. Hashimoto, and H. Ohno. Monitoring independent components for fault detection. *AIChE Journal*, 49(4):969–976, 2003.

[22] J. M. Lee, C. Yoo, and I. B. Lee. Statistical process monitoring with independent component analysis. *Journal of Process Control*, 14(5):467–485, 2004.

[23] C. Shang, F. Yang, X. Gao, X. Huang, J. A. Suykens, and D. Huang. Concurrent monitoring of operating condition deviations and process dynamics anomalies with slow feature analysis. *AIChE Journal*, 61(11):3666–3682, 2015.

[24] L. Eciolaza, M. Alkarouri, N. D. Lawrence, V. Kadirkamanathan, and P. J. Fleming. Gaussian process latent variable models for fault detection. In *2007 IEEE Symposium on Computational Intelligence and Data Mining*, pages 287–292. IEEE, 2007.

[25] A. Abdollahi, K. R. Pattipati, A. Kodali, S. Singh, S. Zhang, and P. B. Luh. Probabilistic graphical models for fault diagnosis in complex systems. In *Principles of Performance and Reliability Modeling and Evaluation*, pages 109–139. Springer, 2016.

[26] L. E. Sucar. Probabilistic graphical models. *Advances in Computer Vision and Pattern Recognition. London: Springer London. doi* 10:978–1, 2015.

[27] C. Sutton, A. McCallum. An introduction to conditional random fields. *Foundations and Trends in Machine Learning*, 4(4):267–373, 2012.

[28] B. Cai, L. Huang, and M. Xie. Bayesian networks in fault diagnosis. *IEEE Transactions on Industrial Informatics*, 13(5):2227–2240, 2017.

[29] J. Zhu, Z. Ge, and Z. Song. Distributed Gaussian mixture model for monitoring plant-wide processes with multiple operating modes. *IFAC Journal of Systems and Control*, 6:1–15, 2018.

[30] J. Yu and S. J. Qin. Multimode process monitoring with Bayesian inference-based finite Gaussian mixture models. *AIChE Journal*, 54(7):1811–1829, 2008.

[31] J. Zhu, Z. Ge, and Z. Song. Robust modeling of mixture probabilistic principal component analysis and process monitoring application. *AIChE Journal*, 60(6):2143–2157, 2014.

[32] Z. Ge and Z. Song. Mixture Bayesian regularization method of PPCA for multimode process monitoring. *AIChE Journal*, 56(11):2838–2849, 2010.

[33] Q. Wen, Z. Ge, and Z. Song. Data-based linear Gaussian state-space model for dynamic process monitoring. *AIChE Journal*, 58(12):3763–3776, 2012.

[34] L. Zhou, G. Li, Z. Song, and S. J. Qin. Autoregressive dynamic latent variable models for process monitoring. *IEEE Transactions on Control Systems Technology*, 25(1):366–373, 2016.

[35] L. Zhou, J. Zheng, Z. Ge, Z. Song, and S. Shan. Multimode process monitoring based on switching autoregressive dynamic latent variable model. *IEEE Transactions on Industrial Electronics*, 65(10):8184–8194, 2018.

[36] J. Zhu, Z. Ge, and Z. Song. Dynamic mixture probabilistic PCA classifier modeling and application for fault classification. *Journal of Chemometrics*, 29(6):361–370, 2015.

[37] N. Sammaknejad, B. Huang, W. Xiong, A. Fatehi, F. Xu, and A. Espejo. Operating condition diagnosis based on HMM with adaptive transition probabilities in presence of missing observations. *AIChE Journal*, 61(2):477–493, 2015.

[38] I. Stanculescu, C. K. Williams, and Y. Freer. Autoregressive hidden Markov models for the early detection of neonatal sepsis. *IEEE Journal of Biomedical and Health Informatics*, 18(5):1560–1570, 2013.

[39] W. K. Ching, E. S. Fung, and M. K. Ng. Higher-order hidden Markov models with applications to DNA sequences. In *International Conference on Intelligent Data Engineering and Automated Learning*, pages 535–539. Springer, 2003.

[40] D. Pandya, S. Upadhyay, and S. P. Harsha. Fault diagnosis of rolling element bearing by using multinomial logistic regression and wavelet packet transform. *Soft Computing*, 18(2):255–266, 2014.

[41] Y. Cai, M. Y. Chow, W. Lu, and L. Li. Statistical feature selection from massive data in distribution fault diagnosis. *IEEE Transactions on Power Systems*, 25(2):642–648, 2010.

[42] E. F. Lussier, Y. He, P. Jyothi, and R. Prabhavalkar. Conditional random fields in speech, audio, and language processing. *Proceedings of the IEEE*, 101(5):1054–1075, 2013.

[43] J. Lafferty, A. McCallum, and F. C. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. June 2001.

[44] Y. Hifny and S. Renals. Speech recognition using augmented conditional random fields. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(2):354–365, 2009.

[45] G. Wang, X. Feng, and C. Liu. Bearing fault classification based on conditional random field. *Shock and Vibration*, 20(4):591–600, 2013.

[46] P. Tang and T. W. Chow. Wireless sensor-networks conditions monitoring and fault diagnosis using neighborhood hidden conditional random field. *IEEE Transactions on Industrial Informatics*, 12(3):933–940, 2016.

[47] S. Borman. The expectation maximization algorithm - a short tutorial. *Submitted for Publication*, 41, 2004.

[48] G. J. McLachlan and T. Krishnan. *The EM algorithm and extensions*, volume 382. John Wiley & Sons, 2007.

[49] D. G. Tzikas, A. C. Likas, and N. P. Galatsanos. The variational approximation for Bayesian inference. *IEEE Signal Processing Magazine*, 25(6):131–146, 2008.

[50] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

[51] M. Soltanieh, A. Zohrabian, M. J. Gholipour, and E. Kalnay. A review of global gas flaring and venting and impact on the environment: Case study of Iran. *International Journal of Greenhouse Gas Control*, 49:488–509, 2016.

[52] S. Casadio, O. Arino, and D. Serpe. Gas flaring monitoring from space using the ATSR instrument series. *Remote Sensing of Environment*, 116:239–249, 2012.

[53] O. G. Fawole, X. M. Cai, and A. MacKenzie. Gas flaring and resultant air pollution: A review focusing on black carbon. *Environmental Pollution*, 216:182–197, 2016.

[54] E. A. Emam. Environmental pollution and measurement of gas flaring. *International Journal of Scientific Research in Science & Technology*, 2:252–262, 2016.

[55] C. E. B. Jr. *The John Zink Hamworthy combustion handbook: Volume 1-Fundamentals.* CRC press, 2012.

[56] N. Bell and M. Nixon. Early alerts of flare events. In *Proceedings of the American Fuels and Petrochemical Manufacturers Annual Meeting*, pages AM–17–74, 2017.

[57] H. Abdi and L. J. Williams. Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(4):433–459, 2010.

[58] P. Geladi and B. R. Kowalski. Partial least-squares regression: A tutorial. *Analytica Chimica Acta*, 185:1–17, 1986.

[59] E. L. Russell, L. H. Chiang, and R. D. Braatz. Fault detection in industrial processes using canonical variate analysis and dynamic principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 51(1):81–93, 2000.

[60] X. Yin, K. Arulmaran, J. Liu, and J. Zeng. Subsystem decomposition and configuration for distributed state estimation. *AIChE Journal*, 62(6):1995–2003, 2016.

[61] X. Yin, J. Zeng, and J. Liu. Forming distributed state estimation network from decentralized estimators. *IEEE Transactions on Control Systems Technology*, 27(6):2430–2443, 2018.

[62] Z. Ge and Z. Song. Distributed PCA model for plant-wide process monitoring. *Industrial & Engineering Chemistry Research*, 52(5):1947–1957, 2013.

[63] S. J. Qin, S. Valle, and M. J. Piovoso. On unifying multiblock analysis with application to decentralized process monitoring. *Journal of Chemometrics*, 15(9):715–742, 2001.

[64] X. Yin and J. Liu. Distributed output-feedback fault detection and isolation of cascade process networks. *AIChE Journal*, 63(10):4329–4342, 2017.

[65] X. Yin and J. Liu. Distributed moving horizon state estimation of two-time-scale nonlinear systems. *Automatica*, 79:152–161, 2017.

[66] R. Raveendran, H. Kodamana, and B. Huang. Process monitoring using a generalized probabilistic linear latent variable model. *Automatica*, 96:73–83, 2018.

[67] L. Wiskott and T. J. Sejnowski. Slow feature analysis: Unsupervised learning of invariances. *Neural Computation*, 14(4):715–770, 2002.

[68] P. Norouzi, M. R. Ganjali, and L. Hajiaghababaei. Fast monitoring of nano-molar level of gentamycin by fast Fourier transform continuous cyclic voltammetry in flowing solution. *Analytical Letters*, 39(9):1941–1953, 2006.

[69] W. T. Peter, Y. Peng, and R. Yam. Wavelet analysis and envelope detection for rolling element bearing fault diagnosis - their effectiveness and flexibilities. *Journal of Vibration and Acoustics*, 123(3):303–310, 2001.

[70] K. A. Kosanovich and M. J. Piovoso. PCA of wavelet transformed process data for monitoring. *Intelligent Data Analysis*, 1(2):85–99, 1997.

[71] J. A. Westerhuis, T. Kourti, and J. F. MacGregor. Analysis of multiblock and hierarchical PCA and PLS models. *Journal of Chemometrics*, 12(5):301–321, 1998.

[72] S. Wold, K. Esbensen, and P. Geladi. Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 2(1-3):37–52, 1987.

[73] C. Shang, B. Huang, F. Yang, and D. Huang. Slow feature analysis for monitoring and diagnosis of control performance. *Journal of Process Control*, 39:21–34, 2016.

[74] R. Ganesan, T. K. Das, and V. Venkataraman. Wavelet-based multiscale statistical process monitoring: A literature review. *IIE Transactions*, 36(9):787–806, 2004.

[75] B. K. Alsberg, A. M. Woodward, and D. B. Kell. An introduction to wavelet transforms for chemometricians: A time-frequency approach. *Chemometrics and Intelligent Laboratory Systems*, 37(2):215–239, 1997.

[76] H. Zhang, A. K. Tangirala, and S. Shah. Dynamic process monitoring using multiscale PCA. In *1999 IEEE Canadian Conference on Electrical and Computer Engineering*, volume 3, pages 1579–1584. IEEE, 1999.

[77] C. F. Alcala and S. J. Qin. Analysis and generalization of fault diagnosis methods for process monitoring. *Journal of Process Control*, 21(3):322–330, 2011.

[78] C. F. Alcala and S. J. Qin. Reconstruction-based contribution for process monitoring. *Automatica*, 45(7):1593–1600, 2009.

[79] S. Zhao, B. Huang, and F. Liu. Fault detection and diagnosis of multiple-model systems with mismodeled transition probabilities. *IEEE Transactions on Industrial Electronics*, 62(8):5063–5071, 2015.

[80] X. Yin and J. Liu. Distributed moving horizon state estimation of two-time-scale nonlinear systems. *Automatica*, 79:152–161, 2017.

[81] Z. Ge, Z. Song, and F. Gao. Review of recent research on data-based process monitoring. *Industrial & Engineering Chemistry Research*, 52(10):3543–3562, 2013.

[82] Y. Tao, D. Shen, M. Fang, and Y. Wang. Reliable H_∞ control of discrete-time systems against random intermittent faults. *International Journal of Systems Science*, 47(10):2290–2301, 2016.

[83] D. Zhao, D. Shen, and Y. Wang. Fault diagnosis and compensation for two-dimensional discrete time systems with sensor faults and time-varying delays. *International Journal of Robust and Nonlinear Control*, 2017.

[84] W. Li, H. H. Yue, S. V. Cervantes, and S. J. Qin. Recursive PCA for adaptive process monitoring. *Journal of Process Control*, 10(5):471–486, 2000.

[85] S. Sedghi and B. Huang. Real-time assessment and diagnosis of process operating performance. *Engineering*, 3(2):214–219, 2017.

[86] F. Qi and B. Huang. Bayesian methods for control loop diagnosis in the presence of temporal dependent evidences. *Automatica*, 47(7):1349–1356, 2011.

[87] S. J. Qin. Survey on data-driven industrial process monitoring and diagnosis. *Annual Reviews in Control*, 36(2):220–234, 2012.

[88] N. Sammaknejad, B. Huang, A. Fatehi, Y. Miao, F. Xu, and 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. *Computers & Chemical Engineering*, 71:281–297, 2014.

[89] S. Zhao, J. Zhang, and Y. Xu. Monitoring of processes with multiple operating modes through multiple principle component analysis models. *Industrial & Engineering Chemistry Research*, 43(22):7025–7035, 2004.

[90] M. M. Rashid and J. Yu. Hidden Markov model based adaptive independent component analysis approach for complex chemical process monitoring and fault detection. *Industrial & Engineering Chemistry Research*, 51(15):5506–5514, 2012.

[91] C. Ning, M. Chen, and D. Zhou. Hidden Markov model-based statistics pattern analysis for multimode process monitoring: An index-switching scheme. *Industrial & Engineering Chemistry Research*, 53(27):11084–11095, 2014.

[92] X. Yin, Z. Li, L. Zhang, and M. Han. Distributed state estimation of sensor-network systems subject to Markovian channel switching with application to a chemical process. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 48(6):864-874, 2017.

[93] C. Sutton and A. McCallum. An introduction to conditional random fields for relational learning. *Introduction to Statistical Relational Learning*, pages 93–128, 2006.

[94] G. Heigold, H. Ney, P. Lehnen, T. Gass, and R. Schluter. Equivalence of generative and log-linear models. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(5):1138–1148, 2011.

[95] G. Zweig and P. Nguyen. A segmental CRF approach to large vocabulary continuous speech recognition. In *Automatic Speech Recognition & Understanding, 2009. ASRU 2009. IEEE Workshop on*, pages 152–157. IEEE, 2009.

[96] D. Zhang, K. Xu, Y. Lu, C. Pan, and H. Peng. Abnormal crowd motion detection with hidden conditional random fields model. *International Journal of Multimedia and Ubiquitous Engineering*, 10(10):91–98, 2015.

[97] J. Zhang and S. Gong. Action categorization with modified hidden conditional random field. *Pattern Recognition*, 43(1):197–203, 2010.

[98] K. Zhang, R. Gonzalez, B. Huang, and G. Ji. Expectation-maximization approach to fault diagnosis with missing data. *IEEE Transactions on Industrial Electronics*, 62(2):1231–1240, 2015.

[99] F. Koushanfar and M. Potkonjak. Markov chain-based models for missing and faulty data in mica2 sensor motes. In *Sensors, 2005 IEEE*, pages 4–pp. IEEE, 2005.

[100] A. J. Quattoni. *Object recognition with latent conditional random fields.* PhD thesis, Massachusetts Institute of Technology, 2005.

[101] T. G. Dietterich, G. Hao, and A. Ashenfelter. Gradient tree boosting for training conditional random fields. *Journal of Machine Learning Research*, 9(Oct):2113–2139, 2008.

[102] L. Liao, D. Fox, and H. Kautz. Extracting places and activities from GPS traces using hierarchical conditional random fields. *The International Journal of Robotics Research*, 26(1):119–134, 2007.

[103] H. M. Wallach. Conditional random fields: An introduction. *Technical Reports (CIS)*, page 22, 2004.

[104] Z. Ghahramani. An introduction to hidden Markov models and Bayesian networks. *International Journal of Pattern Recognition and Artificial Intelligence*, 15(01):9–42, 2001.

[105] D. C. Liu and J. Nocedal. On the limited memory BFGS method for large scale optimization. *Mathematical Programming*, 45(1-3):503–528, 1989.

[106] N. Mittag. Imputations: Benefits, risks and a method for missing data. *Unpublished Manuscript*, 2013.

[107] M. Collins. The forward-backward algorithm. Technical report, Department of Computer Science, Columbia University, Columbia, 2013.

[108] M. A. Henson and D. E. Seborg. Input-output linearization of general nonlinear processes. *AIChE Journal*, 36(11):1753–1757, 1990.

[109] M. Henson and D. Seborg. A unified differential geometric approach to nonlinear process control. In *AIChE Annual Meeting, San Francisco, CA*, 1989.

[110] C. Bishop. Pattern recognition and machine learning (information science and statistics), 1st edn. 2006. corr. 2nd printing edn. *Springer, New York*, 2007.

[111] C. Torrence and G. P. Compo. A practical guide to wavelet analysis. *Bulletin of the American Meteorological society*, 79(1):61–78, 1998.

[112] J. Y. Cheung and G. Stephanopoulos. Representation of process trends - Part I. A formal representation framework. *Computers & Chemical Engineering*, 14(4-5):495–510, 1990.

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

[114] A. A. Popov, T. A. Gultyaeva, and V. E. Uvarov. Training hidden Markov models on incomplete sequences. In *Actual Problems of Electronics Instrument Engineering (APEIE), 2016 13th International Scientific-Technical Conference on*, volume 2, pages 317–320. IEEE, 2016.

[115] I. Jolliffe. *Principal component analysis*. Springer, 2011.

[116] M. Fang, H. Kodamana, and B. Huang. Real-time mode diagnosis for processes with multiple operating conditions using switching conditional random fields. *IEEE Transactions on Industrial Electronics*, 67(6):5060–5070, 2020.

[117] M. Fang, H. Kodamana, B. Huang, and N. Sammaknejad. A novel approach to process operating mode diagnosis using conditional random fields in the presence of missing data. *Computers & Chemical Engineering*, 111:149–163, 2018.

[118] J. Behmann, K. Hendriksen, Ute Müller, Wolfgang Büscher, and L. Plümer. Support vector machine and duration-aware conditional random field for identification of spatio-temporal activity patterns by combined indoor positioning and heart rate sensors. *Geoinformatica*, 20(4):693–714, 2016.

[119] G. Hoefel and C. Elkan. Learning a two-stage SVM/CRF sequence classifier. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, pages 271–278, 2008.

[120] K. Ghosh, M. Ramteke, and R. Srinivasan. Optimal variable selection for effective statistical process monitoring. *Computers & Chemical Engineering*, 60:260–276, 2014.

[121] T. T. Tran. *On conditional random fields: Applications, feature selection, parameter estimation and hierarchical modelling*. PhD thesis, Curtin University of Technology, 2008.

[122] R. Klinger and C. M. Friedrich. Feature subset selection in conditional random fields for named entity recognition. In *Proceedings of the International Conference RANLP-2009*, pages 185–191, 2009.

[123] A. McCallum. Efficiently inducing features of conditional random fields. In *Proceedings*

of the Nineteenth conference on Uncertainty in Artificial Intelligence, pages 403–410. Morgan Kaufmann Publishers Inc., 2002.

[124] Y. Wang and G. Mori. Max-margin hidden conditional random fields for human action recognition. In *CVPR*, volume 9, pages 872–879, 2009.

[125] A. Quattoni, S. Wang, L. P. Morency, M. Collins, and T. Darrell. Hidden conditional random fields. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (10):1848–1852, 2007.

[126] S. Wang, A. Quattoni, L. Morency, D. Demirdjian, and T. Darrell. Hidden conditional random fields for gesture recognition. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1521–1527. IEEE, 2006.

[127] A. K. Menon and C. Elkan. A log-linear model with latent features for dyadic prediction. In *2010 IEEE International Conference on Data Mining*, pages 364–373. IEEE, 2010.

[128] P. T. Boggs and J. W. Tolle. Sequential quadratic programming. *Acta Numerica*, 4:1–51, 1995.

[129] I. Guyon, S. Gunn, M. Nikravesh, and L. A. Zadeh. *Feature extraction: foundations and applications*, volume 207. Springer, 2008.

[130] D. Zhao, S. X. Ding, Y. Wang, and Y. Li. Krein-space based robust H-infinity fault estimation for two-dimensional uncertain linear discrete time-varying systems. *Systems & Control Letters*, 115:41–47, 2018.

[131] D. Zhao, S. X. Ding, H. R. Karimi, Y. Li, and Y. Wang. On robust Kalman filter for two-dimensional uncertain linear discrete time-varying systems: A least squares method. *Automatica*, 99:203–212, 2019.

[132] D. Zhao, S. X. Ding, H. Karimi, and Y. Li. Robust H-infinity filtering for two-dimensional uncertain linear discrete time-varying systems: A Krein space-based method. *IEEE Transactions on Automatic Control*, April 2019.

[133] Z. Zhao, Q. Li, B. Huang, F. Liu, and Z. Ge. Process monitoring based on factor analysis: Probabilistic analysis of monitoring statistics in presence of both complete and incomplete measurements. *Chemometrics and Intelligent Laboratory Systems*, 142:18–27, 2015.

[134] J. Xu, D. W. Ho, F. Li, W. Yang, and Y. Tang. Event-triggered risk-sensitive state estimation for hidden Markov models. *IEEE Transactions on Automatic Control*, January 2019.

[135] Z. Lou and Y. Wang. Multimode continuous processes monitoring based on hidden semi-Markov model and principal component analysis. *Industrial & Engineering Chemistry Research*, 56(46):13800–13811, 2017.

[136] Q. Huang, M. Han, B. Wu, and S. Ioffe. A hierarchical conditional random field model for labeling and segmenting images of street scenes. In *CVPR 2011*, pages 1953–1960. IEEE, 2011.

[137] S. Zheng, S. Jayasumana, B. R. Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. Torr. Conditional random fields as recurrent neural networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1529–1537, 2015.

[138] Y. Qi, M. Szummer, and T. P. Minka. Bayesian conditional random fields. In *Proceedings of the 10th International Workshop on Artificial Intelligence and Statistics*, pages 269–276, 2005.

[139] X. Peng, Y. Tang, W. Du, and F. Qian. Multimode process monitoring and fault detection: A sparse modeling and dictionary learning method. *IEEE Transactions on Industrial Electronics*, 64(6):4866–4875, February 2017.

[140] N. Sammaknejad, B. Huang, and Y. Lu. Robust diagnosis of operating mode based on time-varying hidden Markov models. *IEEE Transactions on Industrial Electronics*, 63(2):1142–1152, February 2016.

[141] M. Kim. Mixtures of conditional random fields for improved structured output prediction. *IEEE Transactions on Neural Networks and Learning Systems*, 28(5):1233–1240, May 2017.

[142] M. Fang, H. Kodamana, and B. Huang. Switching conditional random field approach to process operating mode diagnosis for multi-modal processes. In *2018 IEEE Conference on Decision and Control (CDC)*, pages 5146–5151. IEEE, 2018.

[143] Y. Lu and B. Huang. Robust multiple-model LPV approach to nonlinear process identification using mixture t-distributions. *Journal of Process Control*, 24(9):1472–1488, 2014.

[144] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–38, 1977.

[145] J. A. Bilmes. A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models. *International Computer Science Institute*, 4(510):126, 1998.

[146] C. Couvreur. Hidden Markov models and their mixtures. PhD thesis, Catholic University of Louvain, 1996.

[147] X. Jin, B. Huang, and D. S. Shook. Multiple model LPV approach to nonlinear process identification with EM algorithm. *Journal of Process Control*, 21(1):182–193, 2011.

[148] B. Jiang, H. R. Karimi, Y. Kao, and C. Gao. A novel robust fuzzy integral sliding mode

control for nonlinear semi-Markovian jump T-S fuzzy systems. *IEEE Transactions on Fuzzy Systems*, 26(6):3594–3604, May 2018.

[149] M. Puterman. *Markov decision processes discrete stochastic dynamic programming.* John Wiley & Sons, New York, 1994.

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

Appendix A

Proof of the Equivalence of HMM and LCCRF Model [93]

Since it is a generative model, the HMM models the joint probability $P(\boldsymbol{h}, \boldsymbol{O})$, but the CRF models the conditional probability $P(\boldsymbol{h}|\boldsymbol{O})$ and does not consider the joint distribution. Given a set of discrete states $\{h_1, h_2, ..., h_T\}$, where $h_t \in \{1, 2, ..., N\}$ and the corresponding observations $\{O_1, O_2, ..., O_T\}$, the joint probability $P(\boldsymbol{h}, \boldsymbol{O})$ can be derived under the HMM framework as follows:

$$\begin{aligned} P(\boldsymbol{h}, \boldsymbol{O}) =& P(h_1, h_2, ..., h_T, O_1, O_2, ..., O_T) \\ =& P(O_T|O_{T-1}, ..., O_1, h_T, ..., h_1) \cdot P(h_T|O_{T-1}, ..., O_1, h_{T-1}, ..., h_1) \\ & \cdots P(O_1|h_1) \cdot P(h_1) \end{aligned} \tag{A.1}$$

Following the two conditional independence assumptions of HMMs [150], i.e. $P(h_t|h_{t-1}, ..., h_1) = P(h_t|h_{t-1})$ and $P(O_t|h_t, h_{t-1}, ..., h_1) = P(O_t|h_t)$, the above joint probability $P(\boldsymbol{h}, \boldsymbol{O})$ can be simplified as follows:

$$P(\boldsymbol{h}, \boldsymbol{O}) = \prod_{t=1}^{T} P(O_t|h_t) \cdot P(h_t|h_{t-1}) \tag{A.2}$$

And the conditional probability $P(\boldsymbol{h}|\boldsymbol{O})$ can be formulated thereafter as below:

$$
\begin{aligned}
P(\boldsymbol{h}|\boldsymbol{O}) = \frac{P(\boldsymbol{h},\boldsymbol{O})}{P(\boldsymbol{O})} = \frac{P(\boldsymbol{h},\boldsymbol{O})}{\sum_{\boldsymbol{h}'} P(\boldsymbol{h}',\boldsymbol{O})} &= \frac{\prod_{t=1}^{T} P(O_t|h_t)\cdot P(h_t|h_{t-1})}{\sum_{\boldsymbol{h}'}\prod_{t=1}^{T} P(O_t|h'_t)\cdot P(h'_t|h'_{t-1})} \\
&= \frac{\exp\sum_{t=1}^{T}\{\log P(h_t|h_{t-1}) + \log P(O_t|h_t)\}}{\sum_{\boldsymbol{h'}}\exp\sum_{t=1}^{T}\{\log P(h'_t|h'_{t-1}) + \log P(O_t|h'_t)\}}
\end{aligned}
\tag{A.3}
$$

where $P(h_t|h_{t-1})$ and $P(O_t|h_t)$ are the transition and emission probabilities of HMM, respectively, and the notation $\boldsymbol{h'}$ represents all possible combinations of the states.

Now let us consider the conditional distribution $P(\boldsymbol{h}|\boldsymbol{O})$ under the CRF framework in Equations (4.1) and (4.2) as below:

$$
\begin{aligned}
P(\boldsymbol{h}|\boldsymbol{O}) &= \frac{\exp\sum_{t=1}^{T}\{\sum_{k=1}^{K}\log P(h_t|h_{t-1})T_k(h_t,h_{t-1}) + \sum_{m=1}^{M}\log P(O_t|h_t)E_m(h_t,O_t)\}}{\sum_{\boldsymbol{h'}}\exp\sum_{t=1}^{T}\{\sum_{k=1}^{K}\log P(h'_t|h'_{t-1})T_k(h'_t,h'_{t-1}) + \sum_{m=1}^{M}\log P(O_t|h'_t)E_m(h'_t,O_t)\}} \\
&= \frac{\exp\sum_{t=1}^{T}\{\sum_{k=1}^{K}\lambda_k T_k(h_t,h_{t-1}) + \sum_{m=1}^{M}\mu_m E_m(h_t,O_t)\}}{\sum_{\boldsymbol{h'}}\exp\sum_{t=1}^{T}\{\sum_{k=1}^{K}\lambda_k T_k(h'_t,h'_{t-1}) + \sum_{m=1}^{M}\mu_m E_m(h'_t,O_t)\}}
\end{aligned}
\tag{A.4}
$$

Assuming the summations over weights λ_k and μ_m to be unity and by choosing feature functions shown below, we can demonstrate that Equation (A.4) is equivalent to Equation (A.3).

$$
T_k(h_t,h_{t-1}) = \begin{cases} 1 & \text{if } h_{t-1} = i \text{ and } h_t = j \\ 0 & \text{otherwise} \end{cases}
\tag{A.5}
$$

$$
E_m(h_t,O_t) = \begin{cases} 1 & \text{if } h_t = i \text{ and } O_t \in \boldsymbol{\mathcal{B}} \\ 0 & \text{otherwise} \end{cases}
\tag{A.6}
$$

where $i, j \in \{1, 2, ..., N\}$ are the state values and $\boldsymbol{\mathcal{B}}$ is the set of all possible observation values.

Appendix B

Detailed Derivations and Pseudocodes of Chapter 4

B.1 Detailed Steps of Forward and Backward Propagation

For the forward propagation, a set of intermediate variables $\boldsymbol{\alpha}$ is proposed to increase the computational efficiency for the normalization term $Z(\boldsymbol{O}_{obs})$. Based on the definition of $\alpha_t(h_t, \boldsymbol{h}_{t,mis}^{(f)})$, at the time point $t+1$, the intermediate variable α_{t+1} can be formulated as follows:

$$\alpha_{t+1}(h_{t+1}, \boldsymbol{h}_{t+1,mis}^{(f)}) \overset{def}{=} \sum_{h_{1:t}} \sum_{O_{1:t+1}^{(mis)}} \prod_{t'=1}^{t+1} \varphi_{t'}(h_{t'}, h_{t'-1}, \boldsymbol{Y}_{t'}^{(obs)}) \cdot \gamma_{t'}(h_{t'}, h_{t'+1}, ..., h_{t'+d-1}, O_{t'}^{(mis)}) \tag{B.1}$$

which can be calculated by the following recursion based on the result of α_t.

$$\begin{aligned}\alpha_{t+1}(h_{t+1},\boldsymbol{h}_{t+1,mis}^{(f)}) =&\sum_{h_t}\sum_{O_{t+1}^{(mis)}}\varphi_{t+1}(h_{t+1},h_t,\boldsymbol{Y}_{t+1}^{(obs)})\cdot\gamma_{t+1}(h_{t+1},...,h_{t+d},O_{t+1}^{(mis)})\\&\sum_{h_{1:t-1}}\sum_{O_{1:t}^{(mis)}}\prod_{t'=1}^{t}\varphi_{t'}(h_{t'},h_{t'-1},\boldsymbol{Y}_{t'}^{(obs)})\cdot\gamma_{t'}(h_{t'},...,h_{t'+d-1},O_{t'}^{(mis)})\\=&\sum_{h_t}\varphi_{t+1}(h_{t+1},h_t,\boldsymbol{Y}_{t+1}^{(obs)})\cdot\eta_{t+1}(h_{t+1},...,h_{t+d})\cdot\alpha_t(h_t,\boldsymbol{h}_{t,mis}^{(f)})\end{aligned}\tag{B.2}$$

where the lengths of operating mode sequences $\boldsymbol{h}_{t,mis}^{(f)}$ and $\boldsymbol{h}_{t+1,mis}^{(f)}$ depends on the missing measurements within the time range $t-d+2$ to $t+1$.

After the forward propagation procedures, the normalization term $Z(\boldsymbol{O}_{obs})$ can be calculated based on $\alpha_T(h_T)$, which can be proved as below:

$$\begin{aligned}Z(\boldsymbol{O}_{obs}) =&\sum_{h_{1:T}}\sum_{O_{1:T}^{(mis)}}\prod_{t=1}^{T}\varphi_t(h_t,h_{t-1},\boldsymbol{Y}_t^{(obs)})\cdot\gamma_t(h_t,...,h_{t+d-1},O_t^{(mis)})\\=&\sum_{h_T}\sum_{h_{1:T-1}}\sum_{O_{1:T}^{(mis)}}\prod_{t=1}^{T}\varphi_t(h_t,h_{t-1},\boldsymbol{Y}_t^{(obs)})\cdot\gamma_t(h_t,...,h_{t+d-1},O_t^{(mis)})\\=&\sum_{h_T}\alpha_T(h_T)\end{aligned}\tag{B.3}$$

Similarly, for backward propagation, a set of backward variables $\boldsymbol{\beta}$ is proposed with the definition as below:

$$\beta_t(h_{t+d-2},\boldsymbol{h}_{t,mis}^{(b)})\overset{def}{=}\sum_{O_{t:T}^{(mis)}}\sum_{h_{t+d-1:T}}\prod_{t'=t}^{T}\gamma_{t'}(h_{t'},...,h_{t'+d-1},O_{t'}^{(mis)})\prod_{t'=t+d-1}^{T}\varphi_{t'}(h_{t'},h_{t'-1},\boldsymbol{Y}_{t'}^{(obs)})\tag{B.4}$$

According to this definition, at the time point $t-1$, the corresponding intermediate variable β_{t-1} can be computed as follows:

$$\begin{aligned}
\beta_{t-1}(h_{t+d-3}, \boldsymbol{h}_{t-1,mis}^{(b)}) &= \sum_{O_{t-1:T}^{(mis)}} \sum_{h_{t+d-2:T}} \prod_{t'=t-1}^{T} \gamma_{t'}(h_{t'}, ..., h_{t'+d-1}, O_{t'}^{(mis)}) \prod_{t'=t+d-2}^{T} \varphi_{t'}(h_{t'}, h_{t'-1}, \boldsymbol{Y}_{t'}^{(obs)}) \\
&= \sum_{O_{t-1}^{(mis)}} \sum_{h_{t+d-2}} \gamma_{t-1}(h_{t-1}, ..., h_{t+d-2}, O_{t-1}^{(mis)}) \cdot \varphi_{t+d-2}(h_{t+d-2}, h_{t+d-3}, \boldsymbol{Y}_{t+d-2}^{(obs)}) \\
&\quad \sum_{O_{t:T}^{(mis)}} \sum_{h_{t+d-1:T}} \prod_{t'=t}^{T} \gamma_{t'}(h_{t'}, ..., h_{t'+d-1}, O_{t'}^{(mis)}) \prod_{t'=t+d-1}^{T} \varphi_{t'}(h_{t'}, h_{t'-1}, \boldsymbol{Y}_{t'}^{(obs)}) \\
&= \sum_{O_{t-1}^{(mis)}} \sum_{h_{t+d-2}} \gamma_{t-1}(h_{t-1}, ..., h_{t+d-2}, O_{t-1}^{(mis)}) \cdot \varphi_{t+d-2}(h_{t+d-2}, h_{t+d-3}, \boldsymbol{Y}_{t+d-2}^{(obs)}) \\
&\quad \beta_t(h_{t+d-2}, \boldsymbol{h}_{t,mis}^{(b)})
\end{aligned} \tag{B.5}$$

where the lengths of operating mode sequences $\boldsymbol{h}_{t,mis}^{(b)}$ and $\boldsymbol{h}_{t-1,mis}^{(b)}$ depend on the missing measurements within the time range $t + d - 3$ to $t - 1$.

B.2 The Pseudocodes of the Marginalized CRFs

Algorithm 2 Parameter Estimation

```
Require: The training dataset $\{h_1, h_2, ..., h_T\}$, $\{O_1, O_2, ..., O_T\}$ and the tolerance $\varepsilon$ as termination criteria;
Ensure: The estimated weighting parameters $\boldsymbol{\Theta} = \{\lambda_k, \mu_m\}$;
Initialization: assign the initial guess for $\boldsymbol{\Theta}$ randomly and initial values for gradients;
while $gradient > \varepsilon$ do
    $gradient(\lambda_k) \leftarrow 0$, $gradient(\mu_{ml}) \leftarrow 0$
    $\log Z(\boldsymbol{h}, \boldsymbol{O}_{obs}) \leftarrow 0$
    $\log Z(\boldsymbol{O}_{obs}), P(h_t, h_{t-1}|\boldsymbol{O}_{obs}), P(h_t, O_{t-l+1}^{(mis)}|\boldsymbol{O}_{obs}) \leftarrow MCRF\_Inference(O_{1:T}, d, \boldsymbol{\Theta})$
    for $t = 1 \rightarrow T$ do
        $\log Z(\boldsymbol{h}, \boldsymbol{O}_{obs}) \leftarrow \log Z(\boldsymbol{h}, \boldsymbol{O}_{obs}) + \sum_{k=1}^{K} \lambda_k T_k(h_t, h_{t-1}) + \sum_{m=1}^{M} \mu_m E_m(h_t, \boldsymbol{Y}_t^{(obs)})$
        $gradient(\lambda_k) \leftarrow gradient(\lambda_k) + T_k(h_t, h_{t-1}) - \sum_{h'_t, h'_{t-1}} P(h'_t, h'_{t-1}|\boldsymbol{O}_{obs}) T_k(h'_t, h'_{t-1})$
        if $O_t$ is missing then
            $\log Z(\boldsymbol{h}, \boldsymbol{O}_{obs}) \leftarrow \log Z(\boldsymbol{h}, \boldsymbol{O}_{obs}) + \log\{\sum_{O_t^{(mis)}} \exp \sum_{t'=t}^{T} \sum_{m=1}^{M} \mu_m E_m(h_{t'}, O_t^{(mis)})\}$
        end if
        if $O_{t-l+1}$ is missing then
            $gradient(\mu_{ml}) \leftarrow gradient(\mu_{ml}) + \sum_{O_{t-l+1}^{(mis)}} w(O_{t-l+1}^{(mis)}) E_{ml}(h_t, O_{t-l+1}^{(mis)}) -$
            $\sum_{h'_t} \sum_{O_{t-l+1}^{(mis)}} P(h'_t, O_{t-l+1}^{(mis)}|\boldsymbol{O}_{obs}) E_{m_l}(h'_t, O_{t-l+1}^{(mis)})$
        else
            $gradient(\mu_{ml}) \leftarrow gradient(\mu_{ml}) + E_{ml}(h_t, O_{t-l+1})$
            $- \sum_{h'_t} P(h'_t|\boldsymbol{O}_{obs}) E_{m_l}(h'_t, O_{t-l+1})$
        end if
    end for
    $l(\boldsymbol{\Theta}) \leftarrow \log Z(\boldsymbol{h}, \boldsymbol{O}_{obs}) + \log Z(\boldsymbol{O}_{obs})$
    $\boldsymbol{\Theta} \leftarrow L\_BFGS(l(\boldsymbol{\Theta}), gradient(\boldsymbol{\Theta}))$
end while
```

Algorithm 3 Inference of Marginalized CRFs

```
Require: The observations in training dataset $\{O_1, O_2, ..., O_T\}$, impact factor $d$ and the estimated parameters $\mathbf{\Theta}$;
Ensure: The marginal probabilities $P(h_t, h_{t-1}|\boldsymbol{O}_{obs})$, $P(h_t, O^{(mis)}_{t-l+1}|\boldsymbol{O}_{obs})$ and the normalization term $\log Z(\boldsymbol{O}_{obs})$
function $\text{MCRF\_Inference}(O_{1:T}, d, \mathbf{\Theta})$
    $\alpha_{1:T}, \beta_{1:T} \leftarrow Forward\_Backward(O_{1:T}, d, \mathbf{\Theta})$
    // Normalization term calculation
    $\log Z(\boldsymbol{O}_{obs}) \leftarrow \log \sum_{h_T} \alpha_T(h_T)$
    // Marginal probability calculation
    for $t = 2 \rightarrow T$ do
        $\zeta_{t+d-2}(h_{t-1}, ..., h_{t+d-2}) \leftarrow \alpha_{t-1}(h_{t-1}, \boldsymbol{h}^{(f)}_{t-1,mis}) \cdot \beta_t(h_{t+d-2}, \boldsymbol{h}^{(b)}_{t,mis})$
        $\nu_{t+d-2}(h_t, ..., h_{t+d-2}) \leftarrow \sum_{h_{t+d-1}} \alpha_t(h_t, \boldsymbol{h}^{(f)}_{t,mis}, O^{(mis)}_{t-l+1}) \cdot \beta_{t+1}(h_{t+d-1}, \boldsymbol{h}^{(b)}_{t+1,mis})$
        for $i = d-2 \rightarrow 1$ do
            $\zeta_{t+i-1}(h_{t-1}, ..., h_{t+i-1}) \leftarrow \sum_{h_{t+i}} \zeta_i(h_{t-1}, ..., h_{t+i}) \cdot \varphi_{t+i}(h_{t+i}, h_{t+i-1}, \boldsymbol{Y}^{(obs)}_{t+i})$
            $\nu_{t+i-1}(h_t, ..., h_{t+i-1}) \leftarrow \sum_{h_{t+i}} \nu_i(h_t, ..., h_{t+i}) \cdot \varphi_{t+i}(h_{t+i}, h_{t+i-1}, \boldsymbol{Y}^{(obs)}_{t+i})$
        end for
        $P(h_t, h_{t-1}|\boldsymbol{O}_{obs}) \leftarrow \frac{1}{Z(\boldsymbol{O}_{obs})} \cdot \varphi_t(h_t, h_{t-1}, \boldsymbol{Y}^{(obs)}_t) \cdot \zeta_t(h_{t-1}, h_t)$
        $P(h_t, O^{(mis)}_{t-l+1}|\boldsymbol{O}_{obs}) \leftarrow \frac{1}{Z(\boldsymbol{O}_{obs})} \cdot \nu_t(h_t)$
    end for
    return $P(h_t, h_{t-1}|\boldsymbol{O}_{obs})$, $P(h_t, O^{(mis)}_{t-l+1}|\boldsymbol{O}_{obs})$ and $\log Z(\boldsymbol{O}_{obs})$
end function
```

Algorithm 4 Forward-backward Propagation for Marginalized CRFs

```
Require: The observations in training dataset $\{O_1, O_2, ..., O_T\}$, impact factor $d$ and the estimated parameters $\mathbf{\Theta}$;
Ensure: The propagation intermediate variables $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$;
function FORWARD_BACKWARD($O_{1:T}, d, \mathbf{\Theta}$)
    Initialization: $\alpha_1(h_1) \leftarrow \varphi_1(h_1, h_0, O_1)$, $\beta_T(h_T) \leftarrow 1$
    // The forward propagation
    for $t = 1 \rightarrow T$ do
        $\alpha_{t+1}(h_{t+1}, h_t, \boldsymbol{h}_{t,mis}^{(f)}) \leftarrow \alpha_t(h_t, \boldsymbol{h}_{t,mis}^{(f)}) \cdot \varphi_{t+1}(h_{t+1}, h_t, \boldsymbol{Y}_{t+1}^{(obs)})$
        $\alpha_{t+1}(h_{t+1}, \boldsymbol{h}_{t,mis}^{(f)}) \leftarrow \sum_{h_t} \alpha_{t+1}(h_{t+1}, h_t, \boldsymbol{h}_{t,mis}^{(f)})$
        if $O_{t+1}$ is missing then
            $\alpha_{t+1}(h_{t+1}, \boldsymbol{h}_{t+1,mis}^{(f)}) \leftarrow \alpha_{t+1}(h_{t+1}, \boldsymbol{h}_{t,mis}^{(f)}) \cdot \eta_{t+1}(h_{t+1}, ..., h_{t+d})$
        else
            $\alpha_{t+1}(h_{t+1}, \boldsymbol{h}_{t+1,mis}^{(f)}) \leftarrow \alpha_{t+1}(h_{t+1}, \boldsymbol{h}_{t,mis}^{(f)})$
        end if
    end for
    // The backward propagation
    for $t = T \rightarrow 1$ do
        $\beta_{t-1}(h_{t+d-2}, h_{t+d-3}, \boldsymbol{h}_{t,mis}^{(b)}) \leftarrow \beta_t(h_{t+d-2}, \boldsymbol{h}_{t,mis}^{(b)}) \cdot \varphi_{t+d-2}(h_{t+d-2}, h_{t+d-3}, \boldsymbol{Y}_{t+d-2}^{(obs)})$
        if $O_{t-1}$ is missing then
            $\beta_{t-1}(h_{t+d-3}, \boldsymbol{h}_{t-1,mis}^{(b)}) \leftarrow \sum_{h_{t+d-2}} \beta_{t-1}(h_{t+d-2}, h_{t+d-3}, \boldsymbol{h}_{t,mis}^{(b)}) \eta_{t-1}(h_{t-1}, ..., h_{t+d-2})$
        else
            $\beta_{t-1}(h_{t+d-3}, \boldsymbol{h}_{t-1,mis}^{(b)}) \leftarrow \sum_{h_{t+d-2}} \beta_{t-1}(h_{t+d-2}, h_{t+d-3}, \boldsymbol{h}_{t,mis}^{(b)})$
        end if
    end for
    return $\alpha_{1:T}$ and $\beta_{1:T}$
end function
```

Appendix C

Detailed Derivations and Pseudocode of Chapter 5

C.1 The Pseudocode of Variable Selection in the First-stage HCRF

Algorithm 5 Variable selection for the first-stage HCRF

```
Require: The training dataset $\{h^{(n)}, \boldsymbol{O}^{(n)}\}_{n=1}^{N}$, full variable set $\mathcal{S}$ and empty variable rank set $R$;
Ensure: The ranked variables from the most irrelevant to the most relevant;
Initialization: use all the available variables for MMHCRF training and set $var_{count} = M$;
while $var_{count} > 1$ do
    train the first-stage HCRF model with variables in set $\mathcal{S}$ and obtain $\alpha_{var_{count}}$
    calculate $W(\alpha_{var_{count}})$
    for $m$ in $\mathcal{S}$ do
        calculate $W_{(-m)}(\alpha_{var_{count}})$
        calculate $\Delta W_{(-m)}$
    end for
    rank $\Delta W_{(-m)}$
    $m^* = \arg\min_m \Delta W_{(-m)}$
    $\mathcal{S} \leftarrow \mathcal{S} - m^*$
    $R \leftarrow R \cup m^*$
    $var_{count} \leftarrow var_{count} - 1$
end while
```

C.2 The Variational Parameter Estimation of the Dirichlet Distribution

By substituting $q(\boldsymbol{\zeta}) \sim Dir(\nu)$ into $\log P(L_{1:T}, y_{1:T}, X_{1:T}|\boldsymbol{\zeta})$ of Equation (5.24), D_{KL} is derived as

$$\begin{aligned} D_{KL}(\nu) &= \langle \log q(\boldsymbol{\zeta}) \rangle_{q(\boldsymbol{\zeta})} - \langle \log P(\boldsymbol{\zeta}|\eta) \rangle_{q(\boldsymbol{\zeta})} - \sum_{t=1}^{T} \int_{\boldsymbol{\zeta}} \sum_{L_t} q(L_t) q(\boldsymbol{\zeta}) \sum_{u_1} e_{u_1}(y_t, L_t; \boldsymbol{\zeta}) d\boldsymbol{\zeta} + C_{q(\boldsymbol{\zeta})} \\ &= \langle \log q(\boldsymbol{\zeta}) \rangle_{q(\boldsymbol{\zeta})} - \langle \log P(\boldsymbol{\zeta}|\eta) \rangle_{q(\boldsymbol{\zeta})} - \sum_{t=1}^{T} \langle \sum_{u_1} e_{u_1}(y_t, L_t; \boldsymbol{\zeta}) \rangle_{q(L_t)q(\boldsymbol{\zeta})} + C_{q(\boldsymbol{\zeta})} \end{aligned} \tag{C.1}$$

With the characteristics of Dirichlet distribution, the first two terms can be easily obtained as

$$\begin{aligned} \langle \log q(\boldsymbol{\zeta}_{y_t}|\nu_{y_t}) \rangle_{q(\boldsymbol{\zeta}_{y_t})} &= \log \Gamma(\sum_{l} \nu_{y_t,l}) - \sum_{l} \log \Gamma(\nu_{y_t,l}) + \sum_{l} (\nu_{y_t,l} - 1)(\Psi(\nu_{y_t,l}) - \Psi(\sum_{l'} \nu_{y_t,l'})) \\ \langle \log P(\boldsymbol{\zeta}_{y_t}|\eta_{y_t}) \rangle_{q(\boldsymbol{\zeta}_{y_t})} &= \log \Gamma(\sum_{l} \eta_{y_t,l}) - \sum_{l} \log \Gamma(\eta_{y_t,l}) + \sum_{l} (\eta_{y_t,l} - 1)(\Psi(\nu_{y_t,l}) - \Psi(\sum_{l'} \nu_{y_t,l'})) \end{aligned} \tag{C.2}$$

where $\Gamma(\cdot)$ and $\Psi(\cdot)$ are the gamma function and digamma function, respectively, given as

$$\begin{aligned} \Gamma(\nu_{y_t,l}) &= \int_{0}^{\infty} z^{\nu_{y_t,l}-1} e^{-z} dz \\ \Psi(\nu_{y_t,l}) &= \frac{\frac{\partial \Gamma(\nu_{y_t,l})}{\partial \nu_{y_t,l}}}{\Gamma(\nu_{y_t,l})} \end{aligned} \tag{C.3}$$

The third term of Equation (C.1) can be further derived as

$$\begin{aligned}\langle\sum_{u_1} e_{u_1}(y_t, L_t; \boldsymbol{\zeta})\rangle_{q(L_t)q(\boldsymbol{\zeta})} &= \langle\log P(L_t|y_t, \boldsymbol{\zeta})\rangle_{q(L_t)q(\boldsymbol{\zeta})} \\ &= \sum_l q(L_t = l)\langle\log \boldsymbol{\zeta}_{y_t,l}\rangle_{q(\boldsymbol{\zeta})} \\ &= \sum_l q(L_t = l)(\Psi(\nu_{y_t,l}) - \Psi(\sum_{l'} \nu_{y_t,l'}))\end{aligned} \tag{C.4}$$

By substituting the intermediate results in Equations (C.2) - (C.4), the KL divergence in Equation (C.1) can be simplified as

$$\begin{aligned}D_{KL}(\nu_{y_t}) =& \sum_l(\nu_{y_t,l} - \eta_{y_t,l} - \sum_{t=1}^{T} q(L_t = l))(\Psi(\nu_{y_t,l}) - \Psi(\sum_{l'} \nu_{y_t,l'})) + \log\Gamma(\sum_l \nu_{y_t,l}) \\ &- \sum_l \log\Gamma(\nu_{y_t,l}) + C_{\nu_{y_t}}\end{aligned} \tag{C.5}$$

Taking derivative with respect to $\nu_{y_t,l}$, one can get

$$\frac{\partial D_{KL}(\nu_{y_t,l})}{\partial \nu_{y_t,l}} = \Psi^{'}(\nu_{y_t,l})(\nu_{y_t,l} - \eta_{y_t,l} - \sum_{t=1}^{T} q(L_t = l)) - \Psi^{'}(\sum_{l'} \nu_{y_t,l'})\sum_{l'}(\nu_{y_t,l'} - \eta_{y_t,l'} - \sum_{t=1}^{T} q(L_t = l')) \tag{C.6}$$

The final result can be obtained by setting Equation (C.6) to zero as

$$\nu_{y_t,l} = \eta_{y_t,l} + \sum_{t=1}^{T} q(L_t = l) \tag{C.7}$$