AGH
University of Science and Technology in Krakow

Faculty of Electrical Engineering, Automatics, Computer Science and Biomedical Engineering

DEPARTMENT OF AUTOMATICS AND ROBOTICS

DOCTORAL DISSERTATION

TIAN CONG

STATISTICAL REASONING ANALYSIS OF FAULT OCCURRENCES IN INDUSTRIAL APPLICATIONS

DISCIPLINE:
Automatics and Robotics
SUPERVISOR:
Jerzy Baranowski Ph.D
ASSISTANT SUPERVISOR:
James R. Ottewill Ph.D

Krakow 2020

Akademia Górniczo-Hutnicza
im. Stanisława Staszica w Krakowie

Wydział Elektrotechniki, Automatyki, Informatyki i Inżynierii Biomedycznej

KATEDRA AUTOMATYKI I ROBOTYKI

ROZPRAWA DOKTORSKA

TIAN CONG

WYKORZYSTANIE WNIOSKOWANIA STATYSTYCZNEGO DO ANALIZY WYSTĘPOWANIA USTEREK W ZASTOSOWANIACH PRZEMYSŁOWYCH

DYSCYPLINA NAUKOWA:
Automatyka i Robotyka
PROMOTOR:
Dr. hab. inż. Jerzy Baranowski
PROMOTOR POMOCNICZY:
Dr. James R. Ottewill

Kraków 2020

Abstract

As modern industrial plants are instrumented with a large number of sensors, advanced monitoring algorithms are required to extract actionable insights from the vast quantities of process measurements. The main intention of this thesis is to provide a workflow which is able to handle the monitoring of processes with different range of complexity.

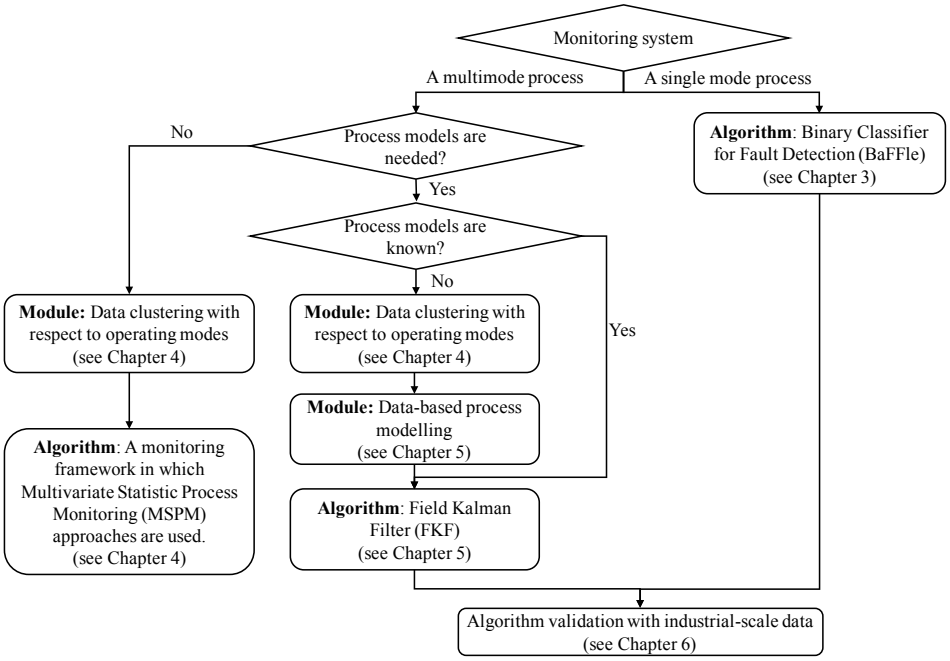


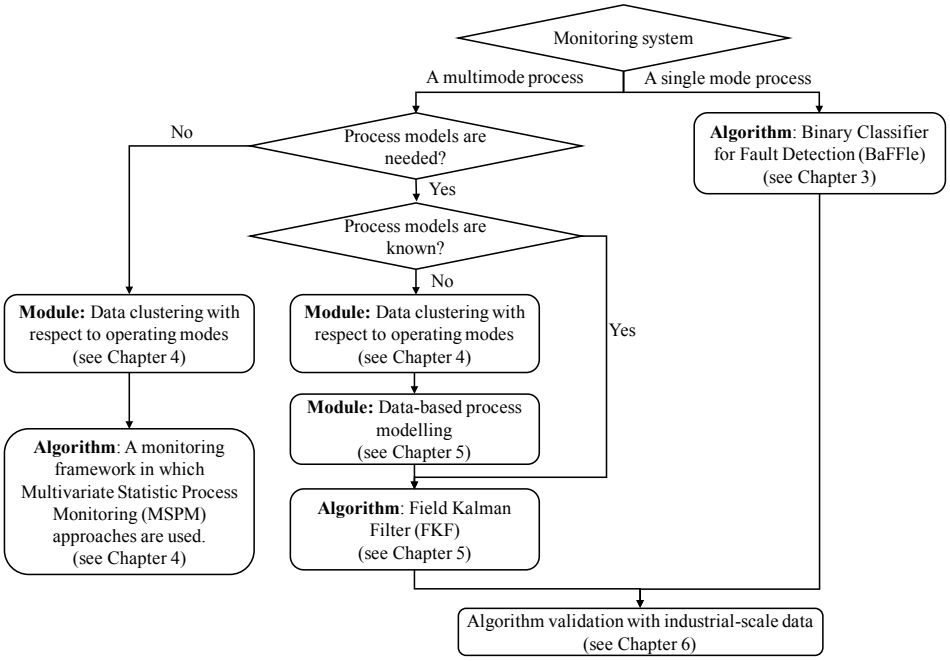
Figure 1: Main achievements in this thesis

This workflow is shown as the flowchart in Fig. 1. It is possible to choose a suitable monitoring approach according to the number of operating modes in the monitored systems. For monitoring a single mode process, a Binary Classifier for Fault Detection (BaFFle), is designed, featuring adaptability to smoothly accommodate itself to monitor single mode systems. In addition, the BaFFle can mitigate the influence of an inappropriate monitoring model by continuously incorporating the incoming data. In terms of monitoring a process with multiple operating modes, this thesis investigated various application scenarios. In the situations where process models are needed, but unknown, a data clustering method, Dirichlet Process-Gaussian Mixture Models (DP-GMMs), was introduced to automatically partition measured data with respect to operating modes without the number of clusters being known in advance. Additionally, this thesis explored the ways of exploiting recorded data for process modelling. The

Multivariate Autoregressive State-Space (MARSS) method is used for deriving the state-space models with the clustered data. Given process models, the Field Kalman Filter (FKF) algorithm can be used for monitoring processes. The modules in Fig. 1, data clustering and data-based process modelling, can also be incorporated into other monitoring algorithms. When there is no need of explicit mathematical process models, a monitoring framework is proposed, in which cluster-based Multivariate Statistic Process Monitoring (MSPM) approaches are the core technique for detecting faults from multimode processes. To validate the BaFFle and the FKF algorithms, industrial-scale multiphase flow data are used. The results show that these two algorithms can improve the detection performance, particularly shortening the detection time and reducing the false and missed alarm rates.

Streszczenie

Nowoczesne instalacje przemysłowe są wyposażone w dużą liczbę urządzeń pomiarowych, stąd też potrzeba zaawansowanych algorytmów monitorujących. Pozwolą one na ekstrakcję praktycznych informacji z ogromnej liczby pomiarów zmiennych procesowych. Głównym celem tej pracy jest przedstawienie sposobu postępowania, który pozwoli monitorować procesy o różnym stopniu złożoności.



Rys. 1. Główne osiągnięcia w rozprawie.

Ten sposób przedstawiono w postaci schematu blokowego na rys. 1. Istnieje możliwość wyboru odpowiedniego podejścia do monitorowania w zależności od liczby trybów pracy w monitorowanych systemach. Do monitorowania procesu jednomodowego zaprojektowano klasyfikator binarny do wykrywania usterek (BaFFle, Binary Classifier for Fault Detection), charakteryzujący się możliwością adaptacji do płynnego dostosowania się do monitorowania systemów jednomodalnych. Ponadto BaFFle może złagodzić wpływ niewłaściwego modelu monitorowania poprzez ciągłe uwzględnianie napływających danych. Dla obszaru monitorowania procesu o wielu trybach pracy, w tej pracy badano różne scenariusze zastosowań. W sytuacjach, w których modele procesów są potrzebne, ale nieznane, wprowadzono metodę grupowania danych, Dirichlet Process-Gaussian Mixture Models (DP-GMMs), aby automatycznie dzielić zmierzone dane w odniesieniu do trybów pracy bez znanej z góry liczby klastrów. Ponadto

w pracy rozważano sposoby wykorzystania zarejestrowanych danych do modelowania procesów. Metoda wielowymiarwych autoregresywnych równań stanu (MARSS, Multivariate Autoregressive State-Space) służy do tworzenia modeli w przestrzeni stanów z danymi z klasteryzacji. Jeżeli modele procesów są dostępne, algorytm Field Kalman Filter (FKF) może być używany do ich monitoringu. Moduły z rys. 1, grupowanie danych (ang. data clustering) i modelowanie procesów w oparciu o dane, można również włączyć do innych algorytmów monitorowania. Gdy nie ma potrzeby stosowania jawnych matematycznych modeli procesów, proponuje się metodykę monitorowania opartą na klasteryzacji. Podstawową techniką wykrywania błędów w procesach wielomodalnych jest wtedy statystyczne monitorowanie procesów wielowymiarowych (MSPM, Multivariate Statistic Process Monitoring). Do walidacji algorytmów BaFFle i FKF wykorzystuje się dane z wielofazowego przepływu w skali przemysłowej. Wyniki pokazują, że te dwa algorytmy mogą poprawić wydajność wykrywania, w szczególności skrócić czas wykrywania i zmniejszyć częstość fałszywych i pominiętych alarmów.

Istnieje możliwość wyboru odpowiedniego monitorowania podejście w zależności od liczby trybów pracy w monitorowanych systemach. Do monitorowania pojedynczego pliku Proces trybu, Binary Classifier for Fault Detection (BaFFle), został zaprojektowany z możliwością adaptacji płynnie dostosowuje się do monitorowania systemów jednomodowych. Ponadto BaFFle może złagodzić wpływ niewłaściwego modelu monitorowania poprzez ciągłe uwzględnianie napływających danych. Jeśli chodzi o monitorowanie procesu w wielu trybach pracy, w tej pracy badano różne zastosowania scenariusze. W sytuacjach, w których modele procesów są potrzebne, ale nieznane, grupowanie danych metoda Dirichlet Process-Gaussian Mixture Models (DP-GMMs), została wprowadzona w celu automatycznego podziału zmierzone dane w odniesieniu do trybów pracy bez wcześniejszej znajomości liczby klastrów. Ponadto praca ta dotyczyła sposobów wykorzystania zarejestrowanych danych do modelowania procesów. Metoda wielowymiarowej autoregresywnej przestrzeni stanów (MARSS) służy do tworzenia modeli w przestrzeni stanów z danymi skupionymi. Przy danych modelach procesów można zastosować algorytm Field Kalman Filter (FKF) monitorowanie procesów. Moduły na rys. 1, grupowanie danych i modelowanie procesów w oparciu o dane, również mogą być włączone do innych algorytmów monitorowania. Kiedy nie ma potrzeby stosowania wyraźnego procesu matematycznego W modelach proponuje się ramy monitorowania, w których oparty na klastrach wielowymiarowy proces statystyczny Podejścia do monitorowania (MSPM) są podstawową techniką wykrywania błędów w procesach wielomodowych. Do walidacji algorytmów BaFFle i FKF wykorzystuje się wielofazowe dane przepływu na skalę przemysłową. Wyniki pokazują, że te dwa algorytmy mogą poprawić wydajność wykrywania, zwłaszcza skracając czas wykrywania i zmniejszenie liczby fałszywych i pominiętych alarmów.

Acknowledgement

First, I would like to thank my family for giving me the freedom and respect to forge my own path. Their unconditional love and extreme support are what allowed me to be who I am now.

Then, it is my great pleasure to acknowledge the European Union's Horizon 2020 research and innovation programme under grant agreement No. 675215-PRONTO-H2020-MSCA-ITN-2015 (known as PRONTO). This program provided an open platform for individuals like me who have passion and enthusiasm in academic and industrial research work.

Additionally, I would like to thank Professor Jerzy Baranowski and AGH University for accepting me to be part of the PRONTO project. Professor Jerzy Baranowski is an inspirational and creative supervisor. His guidance gave me helpful insights and brought up the fundamental idea of my research. I would also like to thank AGH Administrator, Barbara Bysiewicz-Tokarz, who made my living, study and work in Poland smooth.

Moreover, I would thank PRONTO program for providing the placement chance in ABB Corporate Research Center. It was a pleasant experience for me to work with my industrial advisor Dr. James Ottewill in ABB Corporate Research Center, Poland. His enthusiasm for research and patience in advising me are admirable, and would continue to be an inspiration for me in the future. My fellow Early Stage Researcher with whom I collaborated at ABB Corporate Research Center in Poland, Ruomu Tan, was always keen to discuss academic problems, and offered help when I required. My sincere gratitude would also go to Professor Nina Thornhill who provided me with many helpful paper revision suggestions.

The last thank would be given to everyone who has been or is being around me. Thank you all for being nice and kind, and enabling me to carry on with full courage and passion.

Contents

1. Introduction

Process industries are concerned with conversion of raw materials into useful products. It is essential to guarantee the safety and efficiency of processing operations. For this purpose, monitoring systems are required to reflect the health status of operations in real-time. Nowadays, such monitoring systems may take into account the readings from a wide range of sensors that are distributed across the process plants.

The availability of massive amounts of process measurements, produced by the sensors, leads to the prosperity of data-based Process Condition Monitoring (PCM). The use of these measurements can directly or indirectly reflect the health index of processes. Still, simultaneously supervising multiple process variables might be inefficient and cumbersome. Also, in the course of operation, the number of measurements is continuously growing. Thus, data analytics are required to manage information from different sensors, and to draw interpretable monitoring results.

Additionally, due to the varying loading conditions or production regimes, the recorded data would be a mix of various operating modes. There is a need to separate them according to the operating modes such that the characteristics of each individual mode can be further analysed. To alleviate the effort required to manually label the data according to operating modes, the data partition methods should be able to work in an automatic manner and require little prior knowledge regarding the processes, e.g. the number of operating modes. Then, the efficiency of data management could be improved as much as possible.

One of the barriers of applying model-based PCM methods is the process modelling of modern industrial processes. The modelling difficulty arises from the growing complexity of industrial plants. For example, it is challenging to describe the physics of the intricate interlinked equipment with first-principles. With data analytics, the process models might be obtained using historical data.

Due to the demand of more reliable monitoring, new methods and algorithms are required to be developed for PCM. This thesis investigates new monitoring approaches from following aspects:

- Development of data-driven monitoring algorithms which are practically relevant: when applied to monitoring industrial-scale processes, besides the accuracy of monitoring results, the desirable behaviour of monitoring algorithms should also take data management, computational efficiency, the scalability and complexity of algorithms and other implementation issues into account.

- Investigation of nonparametric methods for describing the probability distributions of process data: this aims to allow the estimation of probability distributions with generalisation.

- Development of monitoring algorithms with Bayesian statistic decisions: the decision given by monitoring algorithms should be intuitive, traceable and interpretable to assist end-users to make decisions with more confidence.

- Development of monitoring algorithms which can perform fault diagnosis, mode identification and anomaly detection: the algorithms should distinguish various faults or operating modes with low misclassifications, and identify anomalies with acceptable numbers of missed and false alarms.
- Development of monitoring algorithms accounting for the dynamics of time-series measurements.
- Validation and evaluation of monitoring algorithms in real-life case studies.

The organisation of the thesis is as follows. Chapter 2 gives an overview of the fundamental concepts and techniques of PCM, also identifies the opportunities and challenges in PCM. Chapter 3 proposes a novel heuristic algorithm, named the Binary Classifier for Fault Detection (BaFFle), for monitoring a single mode process. In Chapter 4, a clustering algorithm, Dirichlet Process-Gaussian Mixture Models (DP-GMMs), is reviewed, which is a prerequisite for cluster-based monitoring algorithms. Furthermore, to improve the clustering results, a discussion regarding how to properly initialise the parameters of DP-GMMs is presented. In addition, the DP-GMMs is incorporated into a monitoring framework for clustering historical data and identifying new healthy operating modes. Chapter 5 introduces the theory behind the Field Kalman Filter (FKF) and its use in PCM with simulation examples. Chapter 6 shows the applications of the BaFFle to fault detection, and of the FKF to mode identification and anomaly detection, using industrial-scale PRONTO benchmark data. Chapter 7 concludes the achievements and contributions of this thesis, as well as the potential future work.

2. Process Condition Monitoring (PCM)

This chapter introduces Process Condition Monitoring (PCM). The chapter begins with an overview of PCM with respect to its motivations, tasks and characteristics. After the overview, this chapter reviews the methodologies of process modelling and analyzes their strengths and weaknesses. Next, the definition of multimode processes is given and the main critical characteristics of data from such processes are summarised. Subsequently, as a pre-processing step in PCM, methods of labelling data are introduced. The methods for building monitoring models are revisited and examples of available monitoring indices are provided. In order to draw monitoring results, decision-making methods in PCM are reviewed from multiple-monitoring-model scheme and single-monitoring-model scheme. Also, this chapter discusses the opportunities and challenges of PCM in industrial practice and ends with a summary. Frequently used terminologies and concepts in the context of PCM in industry can be found in Appendix A.1.

2.1. Overview of PCM

2.1.1. Motivations of PCM

Venkatasubramanian et al. (2003) noted that the advent of computer-aided process control has made enormous advances in the discipline of process industries, however governing process plants still remains largely manual activities to avoid the occurrence of abnormal events. To provide suitable control decisions and actions to maintain a process in a normal and safe operating state, two traditional supervision schemes used in industry are corrective maintenance and preventive maintenance. Corrective maintenance is "only fault repair" approach (Wang et al., 2016), intervening a system only when a failure has occurred. For a severe fault, the maintenance cost, for example, the time cost to fix the degraded components and to recover the production line, will significantly increase, compared with a moderate fault. Nevertheless, corrective maintenance is still an economic option for systems in which the interconnections are less complicated and the replacement cost of components is cheap. The preventive strategy is to schedule maintenance actions, such as system condition check and replacement of system components, at a periodic time interval (Wang et al., 2016). In this way, the failure-caused breakdowns become fewer, relative to the corrective strategy. However, the unnecessary scheduled preventive maintenance operation may happen even if the system state is healthy, which will result a high maintenance cost (Jardine et al., 2006), particularly for sophisticated apparatus and equipments.

In an ideal situation, maintenance is planned when a fault is detected. To this end, sensors installed across process plants facilitate in measuring process conditions. Given process measurements, Condition-Based Maintenance (CBM) has gained increasing attention both in theory and applications. In CBM,

the physical parameters of a process plant, such as pressure and flow rate, are detected, measured and recorded (Rao, 1996).

Due to the large scale of industrial plants and complexity of process control and optimisation, industrial processes are prone to hard failure and soft operational faults which will result in economic losses (Qin, 2012). Available incident prevention measures include PCM, reconfiguring system, installing safety instrumented systems, establishing maintenance and repair routines (Niu et al., 2010). PCM plays an important role in preventing plants from failure and shutdown whilst at the same time maintaining plant functionality (Liu and Bazzi, 2017), sustaining efficiency and safety (Dasani et al., 2015) and yielding high product quality (Zhou et al., 2014) and profitability.

2.1.2. Tasks of PCM

Diagnostics

Diagnostics in PCM involves fault detection, fault type diagnosis, fault severity diagnosis, root cause analysis of a fault and many other analysis operations. In high-cost and safety-critical processes, fault detection has enjoyed considerable attention. Increasing demands on reliability and safety of process plants requires that faults are detected as early as possible. This is because early detection can help with the interruption of abnormal events, then timely maintenance can be applied, so as to reduce the possibility of severe damages to industrial facilities and avoid large productivity loss.

Various types of faults may happen in a process plant. Data from one sensor source may contain limited diagnosis information constrained to the type of sensor, physical location, range and other miscellaneous factors. Process plants instrumented with a spectrum of sensors are able to provide heterogeneous data. This helps with a more comprehensive understanding of the behaviours happening in the monitored systems, and allows the determination of the fault type. Fault severity analysis enables to grade the level of fault evolution within an identified fault, which is complementary information for maintenance scheduling. For instance, a Self-Organizing Map (SOM) (Moshou et al., 2010) can be used for visualising the fault severity. Another task in fault diagnosis is root cause analysis which can be aided by identifying influential variables to the detected fault, based on which, practitioners can directly execute maintenance work targeting the degraded components. For example, one of the methods that can be employed to pinpoint the critical variables is contribution plots (Chiang et al., 2000).

Prognostics

Rather than indicate the current health state of monitored systems, PCM also may be used to predict the health state of a system and assess the remaining life of the system. When PCM is used in this way, the function is called prognostics. The motivation of prognostics includes alerting the user of impending faults (e.g. system fatigue, crack propagation, and spall growth), minimising repair and maintenance costs and associated operational disruptions and mitigating the risk of unscheduled downtime (Kothamasu et al., 2006). Prognostics supports the decision-making in terms of an early warning for future degradation and allows appropriate prevention actions prior to material damages to systems. It efficiently extends operation term and life cycle of systems.

2.1.3. Characteristics of PCM

Considering the desirable characteristics of a fault diagnosis system (Venkatasubramanian et al., 2003; Akhlaghi et al., 2017), the following requirements are put forward to guarantee the reliability and efficiency of PCM.

– **Sensitivity**: the PCM system should quickly respond to an abnormal operation in a process. To evaluate the response, detection time is used for measuring the lag between the fault starting time and fault detected time. Usually, shorter detection time is desired. However, PCM with high sensitivity will incur high false alarm rates because they are also sensitive to noise. Frequent false alarms will result in frequent process disruptions. Thus, a trade-off between sensitivity and noise-tolerance is required.

– **Distinguishability**: it is the ability of distinguishing various healthy and/or faulty operations. Under ideal conditions free of noise and modelling uncertainties, PCM should have a 100% certainty associated with the class label generated according to the current operation condition. Venkatasubramanian et al. (2003) also pointed out that it is usually challenging for PCM with high distinguishability performance to tackle modelling uncertainties. The uncertainties will cause misidentification.

– **Robustness**: to reduce the false alarm rate and mis-identification, PCM is required to be robust to noise and uncertainties. To avoid false alarms caused by noise, monitoring limits should be set conservatively. Moreover, PCM should preclude deterministic classification in the presence of uncertainties.

– **Adaptability**: adaptability here include two aspects. First, due to varying operation demands and loading conditions on equipments, process changes often occur in real industrial plants. To adapt to the potential changes, PCM should timely update the monitoring scope of systems as new process behaviours emerge. Second, the designed PCM should be user friendly and applicable to a wide range of practical scenarios, requiring minimal re-training and adjustments. This is also called modularity and scalability in Chapter 2 of the thesis of Stief (2019).

– **Interpretability**: Apart from monitoring the plant, the duties of the operators also include making decisions based on real-time data interpretation, knowledge and past experience (Adhitya et al., 2014). Therefore, PCM should provide interpretable monitoring results so as to assist operators in learning about the health state of processes, identifying fault types, locating fault causes and other condition-based tasks. Consequentially, operators can make appropriate decisions and actions.

– **Storage and computational requirements**: Some algorithms need to utilise historical process data and historical monitoring results, thus PCM should consider its storage ability. Also, on-line PCM usually requires algorithms featuring less computational complexity so as to output results within desirable time. In particular, the occupied time caused by fetching and processing huge amount of data is one important concern.

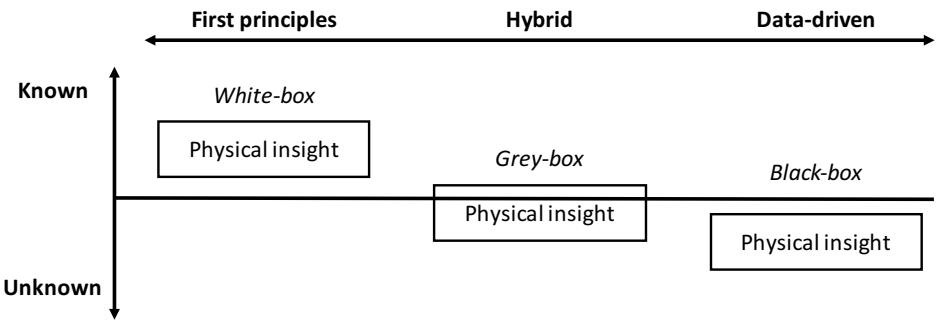


Figure 2.1: Illustration of the differences between process modelling methods (based on Czop et al. 2011)

2.2. Methodologies of process modelling

In the process industries, the use of modelling as a decision-making tool has a long history (Cameron and Ingram, 2008). Process models are the foundation that many applications build upon, including the planning and design of process plants, process monitoring, model predictive control and others related to process operations. In this thesis, we tightly define process model as a description of the process behaviour. The behaviour under certain conditions may have strong auto-correlations (static relationship) and cross-correlations (dynamic relationship). Therefore, one critical problem should be considered in modelling and monitoring is how to effectively model static and dynamic characteristics.

Common modelling methodologies can be categorised into first-principles methods, data-driven methods, and hybrids of both. According to the extent to the availability of prior process physical knowledge, the resulting models are white-, black-, and grey-box models (Bhutani et al., 2006). First-principles modelling is a white-box method as the physics of the process are well-known. Conversely, data-driven modelling is extensively used in black-box situations. Hybrid methods are viewed as grey-box. The differences between these modelling methodologies are illustrated in Fig. 2.1.

2.2.1. First-principles methods

Process models derived using first principles take the form of explicit mathematical equations. The first-principles model has the advantage of encapsulating a large amount of process knowledge. The purpose of using these equations is to provide a detailed description of the effects of the process inputs on the process outputs.

Large-scale industrial plants are usually multi-layer structured: a plant is composed of several subsystems (e.g. mechanical, electrical and process subsystems) and each subsystem is composed of several units. The multi-layer structure results in the high complexity of interactions. On one hand, the complexity, at a minimum, will increase the difficulty in the physics-based description of plant-wide operations which is typically time consuming. On the other hand, it may also indicate the existence of nonlinearity. The solvability and tractability of such nonlinear models within a desirable timeframe is of central importance in terms of their real-time applications (Pantelides and Renfro, 2013).

Usually, more expert knowledge at advanced level contributes to higher accuracy of first-principles models. Moreover, sophisticated models, which capture and deploy process knowledge across the process

lifecycle, reduce the cost of future model development and maintenance. However, due to a limited understanding of process mechanisms and possible unknown process parameters, first-principles derived models in terms of describing process dynamics are not reliable (Wu et al., 2019), for example, in certain classes of applications such as those that involve chemical reactions.

To summarise, inevitably, the issues related to model sustainability, which include complexity, solvability, tractability and maintainability, still remain more or less, and have to be considered to practical and satisfactory performance. This requires the tradeoff between model sustainability and applicability needs to take account of where the models are to be used.

2.2.2. Data-driven methods

A large number of sensors are usually installed in industrial processing plants. The main goal of installing such sensors is to deliver critical process measurements for process monitoring and control. Industrial practitioners believe that most process industries, particularly chemical industry, are in a "data rich and information poor" state, and benefits can be gained from analysing the available process data (Piovoso and Owens, 1991; Kosanovich and Piovoso, 1991). The availability of these data, the increasing computational power and improved modelling algorithms alter the traditional modelling ways to cost-effective data-based approaches. Most data-based modelling methods can be divided into two categories: one is the use of Statistical Data Analysis (SDA) and another is the use of Artificial Neural Networks (ANNs) (Qin and McAvoy, 1993).

Statistical Data Analysis (SDA)

SDA can process a huge amount of process data efficiently. Probability distribution models (e.g. Gaussian distribution) are easy to build and suitable to describe stationary statistics of processes. However, strong assumptions, for example, Gaussian distribution or non-Gaussian distribution, will result in a decrease in accuracy of process models. Besides, dynamic characteristics cannot be represented using probability distribution models. It has been proven that other SDA methods, such as Principal Component Analysis (PCA) and Partial Least Squares (PLS), can extract cross-correlations between process variables effectively (Zhou et al., 2018). Furthermore, dynamic PCA/PLS (Li and Qin, 2001; Russell et al., 2000, 2012), recursive implementation of PCA(Li et al., 2000)/PLS (Qin, 1998), fast moving window PCA (Wang et al., 2005) and multiple-mode PCA (Garcia-Alvarez et al., 2012) have been developed in order to deal with the dynamics in processes. In terms of the nonlinearity problems, Jiang and Yan (2015) developed kernel PCA and Zhang et al. (2009) introduced the kernel PLS model.

Artificial Neural Networks (ANNs)

ANNs are a kind of biology inspired computer tool which excels at analysing complicated nonlinear associations among data (Agatonovic-Kustrin and Beresford, 2000). The modelling sources for ANNs are input and output measurements of system and network topology (Bhutani et al., 2006). Owing to the ready-to-be-used data and its attractive ability of capturing nonlinearity, ANNs-based modelling has been widely studied and employed in various fields. For example, Bialic et al. (2009) utilised the ANNs to describe the nonlinear relationships between the input and output mass flow in a fuel feeding system. Patan (2008) pointed out that in industrial process field, ANNs can aid in fault diagnosis without specific mathematical process models.

2.2.3. Hybrid methods

As discussed in Section 2.2.1, complete and fine-tuned first-principles models are difficult to obtain. In reality, for the sake of simplicity, a number of assumptions are applied to derive first-principle methods. In addition, the data in the process industry can be incomplete and uninformative, while additional tests on demand may be expensive (de Prada et al., 2019). A hybrid way is viable, allowing the integration of process knowledge and valuable information contained in the data (Bhutani et al., 2006).

2.3. Multimode processes

2.3.1. General definition of multimode processes

A single mode process operates in quasi steady-state while its variable values fluctuate slightly; the mathematical description of such process is as follows (Srinivasan et al., 2004):

$$\left| \frac{x(t) - x(t_0)}{t - t_0} \right| < T_x, \ \forall t \in [t_0 - \Delta t, t_0 + \Delta t] \tag{2.1}$$

where T_x is a threshold defined by users and $x(t)$ is a random process variable. Eq. (2.1) shows that for a steady variable, its changing rate should be small, restricted to T_x.

Srinivasan et al. (2004, 2005) defined that a process is multimode if at least one variable violates the steady mode condition. Multimode processes are often found in process industries. There are numerous factors resulting in multimode behaviour, such as the alternations of feedstock and compositions, the changing of manufacturing strategies, the adjustment of set points, the ageing of equipment and the disturbance in the external environment (Yu and Qin, 2008; Lou and Wang, 2017; Shang et al., 2017). Multimode processes are characterised with time-varying, dynamic and nonlinear properties (Quiñones-Grueiro et al., 2019).

2.3.2. Data characteristics in the multimode processes

Computer-aided data acquisition and storage systems allows companies to organise large databases which record measurements relating to the machine, process and plant operation. Measurements from multimode processes have following main critical characteristics:

– **High dimensionality**: modern industry plants are composed of a suite of sub-systems, featuring high connectivity and functionality. Each subsystem embedded with a series of sensors may have a number of measured variables. As a result, industrial processes may generate a massive amount of data samples in high dimensions (Amini and Chang, 2018).

– **Multimodality**: typically, healthy process data recorded from a single mode follows a unimodal Gaussian distribution. However, due to the varying of process conditions, multimodality can appear in normal operating data (Tan et al., 2019).

– **Nonlinearity**: it is common that the relationships between process variables are nonlinear (Ge et al., 2013), for instance, the relationships between mass and energy balance. Moreover, various modes may have different relationships among process variables (Tan et al., 2019), which further complicates the analysis of data from multimode processes.

– **Time-series correlation**: generally, samples are generated from measurements over time and the data are presented in chronological order. As a result of the underlying nature of processes, feedback control systems and time correlated disturbances, dynamic behaviours will arise between process variables, and different samples of each variable are autocorrelated with each other (Ge et al., 2013).

2.3.3. Methods of data labelling

Data labelling has been addressed in many contexts and by researchers in many disciplines. The objective of data labelling is to assign unique labels to data groups within each of which the observations are matched with certain predefined characteristics. For example, labels for data consisting of various faults would be fault types. After labelling, the differences among groups become distinct while the common characteristics within each group emerge. Thus, users can have insights into process performance and operation, then put forward effective solutions to the problems encountered. Chegini et al. (2019) overviewed the methods supporting the labelling of multivariate records, and categorised these methods into visual clustering, clustering, classification and active learning. Yet, since active learning is more of a mechanism to improve the labelling results, this thesis only discusses the labelling methods from following three categories.

Visual clustering

Visual clustering is to visually inspect and explore the similarities/dissimilarities in the data. For example, scatter plots are commonly used to display two-dimensional data where clusters can be discerned by spacial proximities of samples, and labels can be assigned accordingly. To present multivariate data in scatter plots, dimensionality reduction steps are performed. Other approaches for visualising data in higher dimensions, such as parallel coordinates (Wang et al., 2004) and high density plots (Thornhill et al., 2006), can be used. Nevertheless, the criteria and process of label assignment proceeds fully under user governance and requires the input of user knowledge and experience.

Classification

Classification-based methods is a mapping operation, from a set of unlabelled data $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots$ to a finite set of J discrete class labels $c_1, c_2, \ldots, c_J$, written as $f(\boldsymbol{x}_i) \in \{c_1, c_2, \ldots, c_J\}$ where $f()$ is a mapping function. The prerequisite of data labelling using classification methods is to have a comprehensive understanding of the process performance and operation as well as an accurate and robust mapping function. Given this information, classification is a convenient and effective tool to partition raw data into several groups.

Clustering

In clustering analysis, there are no class labels available but raw data $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots$. The goal of clustering is to separate a set of unlabelled data into an appropriate number of subsets. The clustering methods can operate either with the number of clusters specified or without. In the former manner, the fundamental problem is to determine the number of clusters, J. For some applications, J can be provided by expertise of users. Under some circumstances, the estimate of J are exclusively from the data themselves. For example, a heuristic scheme (Tseng and Yang, 2001) and a Monte-Carlo cross validation method (Smyth, 1997) were proposed for estimating J. Many clustering algorithms require J to be provided as a-prior,

and the quality of resulting clusters largely dependent on the estimation of J (Rui Xu and Wunsch, 2005). To minimize the influence from pre-specified J, some clustering algorithms are developed to adaptively and dynamically adjust the number of clusters. For instance, adaptive resonance theory networks create a new cluster only when the characterisation match between the data and their expectations is below some given confidence value (Carpenter and Grossberg, 1987). Dirichlet Process (DP)-based clustering analysis, starting with a large specification of the number of clusters, works in an iteration way, and gradually converges to an appropriate small value (Escobar, 1988, 1994).

2.3.4. Monitoring models

Traditional multivariate statistical monitoring approaches apply many assumptions to data (Joe Qin, 2003; Ge et al., 2013; Kruger and Xie, 2012; Zhao and Gao, 2014), for example, process variables of linear, deterministic, normally distributed and operated under single mode. These assumptions will cause a decrease in the accuracy of process monitoring. Hence, monitoring approaches designed for multimode processes need to consider all the above-mentioned characteristics. There are three frequently used methods for building monitoring models:

- **Local model-based method** builds several sub-models corresponding to the monitored operating modes. Some examples can be found in the literature of Zhao et al. (2004), Zhao et al. (2006) and Natarajan and Srinivasan (2010).
- **Global model-based method** builds a unified model to fit all the given operating modes. Hwang and Han (1999) proposed to apply PCA to build a global model which accounts for various modes, however, it requires that across all the modes, covariance structures share common process behaviour characteristics in the meanwhile nonlinearity characteristics are weak. Deng et al. (2017) and Zhang et al. (2017a) have applied Kernel PCA to account for multiple operating modes, and lead to a single monitoring model.
- **Adaptive model-based method** is to adaptively update the model according to the mode changes. Xie and Shi (2012) developed an adaptive monitoring scheme in which the real-time model update was performed by tracing process variations. Ma et al. (2014) introduced a two-step adaption monitoring approach to keep the monitoring model up-to-date. Rather than fixed monitoring models, adaptive methods behave more flexible to the changes in the process.

The above methods are also suitable for the situation where explicit mathematical models are required.

2.3.5. Monitoring indices

In this subsection, a number of monitoring indices are introduced. These indices are usually used to determine the health of plants.

Model-based indices

The most commonly followed Fault Detection and Isolation (FDI) algorithm structure for model-based PCM is to generate residuals, compute thresholds, and make decisions (Ding et al., 2009). Due to the development of advanced system and control theory, model-based process condition monitoring techniques are widely applied to highly dynamic systems and control loops which are typically located at

the process level, aiming to provide an efficient and powerful tool to detect faults and to diagnose faults (Ding, 2014). A highly effective FDI algorithm is often the result of highly complex process models (Ding, 2014). In particular, the model-based PCM for Linear Time Invariant (LTI) systems has been well-researched and established (Gertler, 1998; Blanke et al., 2006; Ding, 2008; Chen and Patton, 2012; Patton et al., 2013). Given process models, numerous standard methods are available for designing the fault detection and isolation systems (Frank and Ding, 1997; Venkatasubramanian et al., 2003).

Model-based FDI methods are mainly based on analytical residuals which describe the discrepancy between the real and estimated system information. Usually, the real system information are given (e.g. system parameters) or indirectly read from sensors (e.g. system measurements). The residual techniques have the advantage of dealing with process dynamics, robustness issues and structural fault isolation problems effectively and systematically (Ding et al., 2009). According to Calado et al. (2001) and Kothamasu et al. (2006), there are three ways to generate residuals:

– Observer-based methods compare the actual system state with those estimated by either Luenberger observers or Kalman filters (Simani et al., 2003). Lower residual values indicate healthy operation whereas higher values indicate the likely presence of a fault.

– Parameter estimation methods evaluate and analyse the changes in system parameters with measure inputs and outputs. For example, variation in reaction rate of a chemical process may indicate the occurrence of a fault. However, the reaction rate is a system parameter that is difficult to measure directly. To deal with this issue, the measurements, such as pressure and temperature, can be used for estimating the reaction rate.

– Parity space methods rely on the measurements from the system, generating residuals by comparing the model and the system behaviour when the explicit system models are known.

Statistics-based indices

Generally, the multivariate statistical approaches are designed to perform process monitoring in static or dynamic processes in the steady-state, and are able to deliver optimal performance for high-level fault detection and diagnosis in large-scale systems (Ding, 2014).

Statistics-based methods aim to use statistics to indicate the variability in multivariate processes. Assuming that data of normal operation are available and subject to a specific probability distribution, a statistic for a given data sample can be calculated using distribution properties. Other most commonly used statistics are Hotelling's T^2 and Squared Prediction Error (SPE) (Joe Qin, 2003). Hawkins' T_H^2 statistic is a symmetric implementation of T^2 in the residual subspace (Hawkins, 1974). The sum of T^2 and T_H^2 is the Mahalanobis distance (Joe Qin, 2003) which is also a widely used detection index. A combined use of T^2 and SPE was proposed by Yue and Qin (2001) as a fault detection index.

2.4. Decision-making process

2.4.1. Multiple-monitoring-model scheme

Commonly, local model-based methods generate a monitoring indicator for each of operating modes/ faulty operations. An additional decision-level fusion step, which maps the multiple indicators into a

single, consensus monitoring decision, is often performed to identify the current operating mode or diagnose faults. Monitoring performance benefits from this decision-making process. For example, given a data sample, its indicator values of two or more fault types may have similar values. Without further analysis, it is hard to determine which type of fault the data sample belongs to. After implementing decision-level fusion, correct values are amplified, and incorrect ones are reduced (Ghosh et al., 2011). Hence, the final decision is easy to be determined, and would be more reasonable and convincing.

More general, a decision-fusion framework usually comprises a set of classifiers. Each classifier is able to recognise several class labels, and choose one of them as its sub-decision. The aim of a fusion framework is to integrate the sub-decisions of all the classifiers and to conclude the final class decision. A review of decision-level fusion in the applications of condition monitoring can be found in (Stief, 2019). Decision-level fusion can be broadly categorised as utility-based and evidence-based methods (Ghosh et al., 2011; Tidriri et al., 2016). For each category, one representative fusion method is introduced.

Utility-based fusion

Voting-based fusion is a utility-based method, which is easy and simple to implement. Aggregating all of the sub-decisions, the final decision is drawn through a voting fashion. Based on the voting strategies, voting-based fusion has following variants (Ghosh et al., 2011):

- **Unanimous voting**: there are two status of the final decision, accept or reject. If all sub-decisions are in agreement, the final decision accepts them; otherwise, rejects.

- **Simple majority voting**: the final decision is determined by at least one more than 50% the number of sub-decisions.

- **Plurality/Majority voting**: the sub-decision with highest vote counts is the final decision, whether or not the agreement exceeding 50%.

Evidence-based fusion

Bayesian fusion is an efficient evidence-based method, applied to the classifier where each class label is estimated in a posterior probabilistic way. Bayes' theorem is used to calculate the posterior probability. The final label is determined based on the class with the maximal posterior value. Ghosh et al. (2011) summarised that the Bayesian fusion method follows a four-step process: compute individual probability of each class; compute overall probability of each class; compute Bayesian probability value of each class; and apply decision rule.

Bayesian fusion has been successfully and widely used in diverse fields, ranging from medical testing (e.g. disease diagnosis based on image processing (Zheng et al., 2005)), to machine condition monitoring (e.g. fault diagnosis in a power transformer (McArthur et al., 2004) and in a motor (Niu and Li, 2017)) and to pattern recognition (e.g. Fingerprints and handwritten signature recognition (Yang et al., 2013))

For multimode PCM within a probability framework, a mixture modelling approaches have been proposed in many research works (Ge and Song, 2010a; Yang et al., 2015; Zhu et al., 2015). As a result, fusing mixture models in a bayesian way is able to naturally tackle noisy data of industrial processes (Ge and Song, 2010b; Ge, 2018) and implement robust process monitoring (Zhu et al., 2014).

2.4.2. Single-monitoring-model scheme

The development of a single monitoring model for FDI has attracted attention because of the reduced modelling effort. This effort becomes prominent in the cases where data are of multiple operating modes. The traditional multiple-monitoring-model scheme generates several indicators, a single model scheme usually has one indicator. The most simple and easy way of its application to FDI is to set a threshold for testing the indicator to identify if a data sample is normal or abnormal. For a continuous and stationary process, the threshold can be a predefined fixed value, and the determination of it can be from expert experience or from statistical knowledge. Alternatively, in the monitoring progress, the threshold value can be adjusted automatically to adapt to the operating conditions (Isermann, 2006). However, it is difficult to build a single monitoring model incorporating all of the static and dynamic characteristics across various process conditions.

2.5. Opportunities and challenges in industrial practice

The industrial processes have been hugely increasing their degree of automation ever since the 1960's owing to both more demanding performance requirements and the need to reduce human exposure to repetitive, tedious and often dangerous tasks (Dos Reis and Costa, 2013). In the meantime, the evolution in industrial automation also requires CBM to detect and diagnose abnormal operation in order to ensure system reliability, productivity and safety. Furthermore, the industrial use of CBM has its economic feasibility. Rastegari and Bengtsson (2014) analysed a pilot case study of a major manufacturing site in Sweden and found that great paybacks can be gained after the implementation of CBM. There are also many other research efforts on economic feasibility analysis of implementing CBM, and Al-Najjar and Alsyouf (2004) have long promoted the idea that CBM can convert maintenance to a profit centre.

The performance of CBM substantially relies on the accuracy of selected PCM algorithms. According to Kline (1991), there are two types of errors in detection algorithms, namely false alarms and missed alarms. The difference between these two kinds of errors is that false alarms occur in normal operation while missed alarms appear in faulty operating conditions (Zhang et al., 2017b).

In normal operation, the operator might be too distracted or overwhelmed by false alarms to properly carry out work duty. Some literatures investigated the causes of this alarm overload. Borowski et al. (2011) noted that the occurrence of false alarms might be caused by irrelevant noise and outliers in measured data. In addition, false alarms may result from the way that an adaptive monitoring method is implemented. One frequent occasion for such false alarms to occur is during the transition between two operating modes (Ge and Song, 2012). Tan et al. (2019) discussed the impact of overfitting issue on the number of false alarms. The presence of missed alarms might originate in following causes. Due to the nature of a given decision-making mechanism, a faulty data sample might be considered as normal. For example, unanimous voting strategy for drawing detection conclusion is prone to missing faults (Stief, 2019). Another cause of the missed alarm might be that, the effect of the fault is minuscule (Stief et al., 2019, Tan et al., 2019) or at a comparable level to noise. Moreover, a detection algorithm with a relaxed monitoring threshold tends to have a rate of missed alarms (Yang et al., 2009). In practice, an overlooked fault owing to missed alarms might develop into failure, further causing loss of production, equipment, and impacting the safety of the process.

To exploit the advantages of CBM, it is critical to design algorithms with acceptable levels of errors, including both false and missed alarms. Additionally, the design of algorithms should be accountable for the complexities of industrial processes, such as the noises and the multimodal operation. Normally, noises come from processes themselves and sensors, the levels of which might be varying in distinct applications (Akhlaghi et al., 2017). Thus, PCM algorithms should be able to adapt to the various scenarios, such that CBM might be more reliable and robust. On the other hand, for a specific application, there can be multiple modes. Within each mode, the correlations between variables might be unique. Monitoring models that can not fully capture these correlations might lead to poor detection results. Moreover, the performance of data-driven monitoring approaches might be hindered by the training data that are not representative of the true behaviours of processes. To address the issues of unrepresentative data, adaptability should be considered.

2.6. Summary

In this chapter, an introduction to Process Condition Monitoring (PCM) has been given conceptually and technically. Also, this chapter has presented the opportunities and challenges of the application of PCM in industry. According to the investigation of the aforementioned knowledge, the following requirements have been identified:

- The monitoring system/algorithms should be sufficiently adaptive to be used on processing systems with a variety of complexity. The complexity might arise from the varying production demand and loading conditions on equipments, which leads to multimodality issues. In addition, the highly connected physical components and parts in industrial plants might also increase the complexity, making first-principle modelling non-trivial. Thus, general solutions enable monitoring algorithms to be easily implemented in practice.

- Monitoring algorithms should work regardless of the size of data, availability of data labels and existence of mathematical process models.

- Fault detection algorithms should have acceptable levels of false alarms and missed alarms so as to minimise downtime and maximise production efficiency.

- It is required that monitoring algorithms with classification ability should distinguish various operation behaviours with low misclassification rate. This means that the classification is robust to the noise and modelling uncertainties.

- The monitoring results has to be intuitive and interpretable for support engineers and operators to make decisions regarding to Condition-Based Maintenance (CBM) with more confidence.

3. Binary Classifier for Fault Detection (BaFFle) algorithm

Historical data might contain the measurements from start-up, transient or other unsteady phases. To mitigate the influence of these measurements on the quality of monitoring models, monitoring algorithms are required to accommodate the presence of unrepresentative data.

To this end, a fault detection algorithm, called the Binary Classifier for Fault Detection (BaFFle) algorithm, is introduced in this chapter. The design of BaFFle aims to reduce false alarms and missed alarms so as to improve the detection performance. To evaluate the detection performance, concepts, such as sensitivity, specification and accuracy, are given in Section 3.1. The adaptability of fault detection algorithms is discussed in Section 3.2. In Section 3.3, the uni- and multivariate Shewhart Control Charts (SCC) are reviewed. Furthermore, a framework of applying multiple univariate control charts to monitor a multivariate process is proposed. Principal Component Analysis (PCA) is reviewed and used in the proposed framework. Section 3.4 introduces three widely-used univariate density estimation methods. The details of the BaFFle algorithm are presented in Section 3.5. The chapter ends with a summary. This chapter is developed based on (Cong and Baranowski, 2018a) and (Cong and Baranowski, 2018b).

3.1. Evaluation of the performance of fault detection algorithms

In fault detection algorithms, the detection outcomes are usually binary: healthy and faulty. When comparing the true health state of a system against the state identified by a fault detection algorithm, there are four possible outcomes, as shown in Table 3.1.

Table 3.1: Four possible cases of comparing the true health state of a system and the identified health state (Márquez-Flores, 2010)

Cases	True health state	Identified health state
True Positive (TP)	Anomaly	Anomaly
True Negative (TN)	Normal operation	Normal operation
False Positive (FP)[a]	Normal operation	Anomaly
False Negative (FN)[b]	Anomaly	Normal operation

[a] FP is also called false alarm in fault detection;
[b] FN is also called missed alarm in fault detection.

In order to quantify the performance and reliability of fault detection algorithms, statistics, such as sensitivity, specificity and accuracy, are often used. Zhu et al. (2010) and Baratloo et al. (2015) gave the concepts and calculations of these statistics:

– Sensitivity evaluates the performance of a monitoring algorithm in detecting anomalies. The calculation is:

$$\text{Sensitivity} = \frac{n_{\text{TP}}}{n_{\text{TP}} + n_{\text{FN}}} \tag{3.1}$$

where n_{TP} is the number of samples correctly identified as anomalies and n_{FN} is the number of samples incorrectly identified as normal operation.

– Specification evaluates the performance of a monitoring algorithm in recognising normal operation. The calculation is:

$$\text{Specification} = \frac{n_{\text{TN}}}{n_{\text{TN}} + n_{\text{FP}}} \tag{3.2}$$

where n_{TN} is the number of samples correctly identified as normal operation and n_{FP} is the number of samples incorrectly identified as anomalies.

– Accuracy evaluates the performance of a monitoring algorithm in identifying normal and abnormal samples correctly. The calculation is:

$$\text{Accuracy} = \frac{n_{\text{TP}} + n_{\text{TN}}}{n_{\text{TP}} + n_{\text{TN}} + n_{\text{FP}} + n_{\text{FN}}} \tag{3.3}$$

Generally, fault detection algorithms which are performing well should have high sensitivity, specification and accuracy. This indicates that the number of false alarms (FP) and missed alarms (FN) should be low. In practice, false alarms usually lead to unnecessary maintenance actions while missed alarms may result in severe failures. In order to avoid the economic losses due to equipment downtime, and to reduce the time required to fix equipment, Orkisz (2017) suggested that monitoring systems should be prone to tolerant of false alarms, but be less tolerant of missed alarms.

3.2. Motivations of the adaptability of fault detection approaches

In industry, due to the plant-wide installed sensors, multiple process-condition-related variables can be collected, and be used to analyse the health state of plants. However, independently monitoring these variables might result in misleading interpretation of the health index of plants (Bersimis et al., 2007). On the contrary, Multivariate Statistic Process Monitoring (MSPM) techniques, for instance, Hotelling's T^2, aims to treat variables collectively and draw a unified monitoring limit. In such way, plant operators do not need to inspect multiple variables simultaneously. Particularly, MSPM techniques are appropriate for monitoring cases with massive amounts of measurement data. Nevertheless, high correlation between variables might cause biased monitoring results. To deal with this problem, methods, such as Principal Component Analysis (PCA), Partial Least Squares (PLS) and Canonical Variate Analysis (CVA), can be adopted to minimise the correlation, as well as to extract features and reduce data dimensionality. Furthermore, Jiang et al. (2015) pointed that given independent and identically distributed process noises, PCA and PLS methods are suitable for process measurements. Due to the fact that the performance of PLS method is largely dependent on the number of selected features (Guo et al., 2020), in this thesis, PCA is selected for pre-processing measurement data.

In fault detection applications, typically, PCA models for feature extraction and control limits for distinguishing faults from normal operation are trained in an off-line manner, then applied to on-line

detection tasks. The monitoring models comprising of PCA models and control limits are fixed, thus are suited to perform fault detection for time-invariant processes. However, in industrial practice, there are many normal process changes over time, such as batch processes and changing external loading conditions. The characteristics of time-varying industrial processes summarised by Li et al. (2000) include changes in the mean, changes in the variance and changes in the correlation structure among variables. Often, monitoring models which do not consider these time-varying characteristics will be prone to false or missed alarms. To address this problem, various authors have investigated adaptive monitoring methods such that monitoring models can be adjusted on-line so as to reflect the real-time process behaviours. Wold (1994) proposed the exponentially weighted moving average based-PCA method. Li et al. (2000) used a recursive approach to update PCA-based monitoring models. Liu et al. (2009) applied a moving window into recursive PCA methods.

In the online monitoring case, the number of collected data grows in the course of operation. The newly acquired measurements can be used to improve the monitoring performance. However, it might be unnecessary to continuously update PCA models, especially for a steady-state process. Also, frequently updating PCA models online will decrease computation efficiency. Given a steady-state process, one fixed monitoring model may be able to account for the entire process if the training data are sufficient. Nevertheless, a process might be unsteady at the start-up phase and gradually evolve into steady state. For such a process, the problem is that the control limits obtained using data including start-up will be higher than the ones with only steady-state data. This problem may potentially be addressed by training a monitoring model which is representative of the steady-state by excluding the data from start-up stage. However, in multivariate data cases, due to signal delays and highly correlated relations between variables, data examination to remove the start-up data requires engineers and operators to be very familiar with the monitored systems. Adaptive fault detection algorithms are promising to fit monitoring models to normal changes in the signal measurements and to detect anomalies. In Alkaya and Eker (2011), a threshold adjustment mechanism was designed to enable the monitoring models applicable for fault detection in transient operating conditions. In this thesis a heuristic fault detection algorithm, called BaFFle, is proposed to adaptively update monitoring models, particularly for control limits.

3.3. Process control charts

In order to distinguish abnormal operation in industrial processes, control charts are widely used. According to the number of inspected variables, monitoring approaches can be generally divided into univariate and multivariate categories. Shewhart Control Charts (SCC) are often used in online process monitoring due to their low computational complexity and performance in detecting severe faults (Chaabane et al., 2018). In this section, the univariate and multivariate SCC are reviewed. Moreover, the flow diagram showing how univariate control charts may be applied to inspect multiple variables is proposed.

3.3.1. Univariate Shewhart chart

In the 1920s, Walter A. Shewhart devised the univariate SCC (Shewhart, 1926) which was often applied to monitoring stable processes (Hryniewicz and Kaczmarek-Majer, 2018). There are two key elements in the univariate SCC, the baseline and the control limits. The baseline represents the mean value of a normal process (Hryniewicz and Kaczmarek-Majer, 2018). The control limits depict the boundary

of a process under normal operation, which can be bilateral (two limits, the Upper Control Limit (UCL) and the Lower Control Limit (LCL)) or unilateral (either UCL or LCL is selected for monitoring). Faults defined in the univariate SCC are those samples falling outside the limits while samples within given boundaries are normal operation.

An assumption for applying the univariate SCC is that the healthy samples are independent, and follow a Gaussian distribution. Given univariate process measurements $x_1, x_2, \ldots, x_n$, the baseline and the control limits are calculated by:

$$\begin{aligned} \text{BL} &= \frac{1}{n}\sum_{i=1}^{n} x_i \\ \text{UCL}_\text{u} &= \text{BL} + c\hat{\sigma} \\ \text{LCL}_\text{u} &= \text{BL} - c\hat{\sigma} \end{aligned} \tag{3.4}$$

where BL represents the baseline, UCL_u and LCL_u are respectively the UCL and LCL of a univariate control chart. c is a coefficient defining the level of deviation, usually set to 3 corresponding to a confidence level of value 99.97% . The univariate SCC allows the detection of a possible mean shift above UCL_u or below LCL_u.

3.3.2. Multivariate Shewhart control chart

In bivariate cases, when process data follow a Gaussian distribution, probability elliptical contours can be derived. The probability elliptical contours are lines connecting points of equal probability, which can be used as monitoring limits (Alt and Smith, 1988). However, when the number of inspected variables is more than 3, the elliptical contours are not applicable. This is because it becomes increasingly difficult to visualise the contour plots (Liu, 1995). Thus, a multivariate SCC needs to be designed, which is applicable to high dimensional (3+) cases. The basic idea of the multivariate SCC is to employ a univariate monitoring statistics to extract the information in multiple variables. Assuming a random sample $\boldsymbol{x}_* \in \mathbb{R}^m$ subject to a multivariate Gaussian distribution parameterised by $\boldsymbol{\mu} \in \mathbb{R}^m$ and $\Sigma \in \mathbb{R}^{m\times m}$, Bersimis et al. (2005) gave the derivations of the monitoring statistics of $\boldsymbol{x}_*$ as follows :

– If $\boldsymbol{\mu}$ and Σ are known,

$$\mathcal{C}(\boldsymbol{x}_*) = (\boldsymbol{x}_* - \boldsymbol{\mu})^\top \Sigma^{-1} (\boldsymbol{x}_* - \boldsymbol{\mu}). \tag{3.5}$$

where $\mathcal{C}(\boldsymbol{x}_*)$ is the monitoring statistics of $\boldsymbol{x}_*$, following a Chi-square distribution with m degrees of freedom.

– If $\boldsymbol{\mu}$ and Σ are unknown,

$$\mathcal{C}(\boldsymbol{x}_*) = (\boldsymbol{x}_* - \hat{\boldsymbol{\mu}})^\top \hat{\Sigma}^{-1} (\boldsymbol{x}_* - \hat{\boldsymbol{\mu}}) \tag{3.6}$$

where $\hat{\boldsymbol{\mu}}$ and $\hat{\Sigma}$ are the estimates of $\boldsymbol{\mu}$ and Σ using historical data $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$. $\mathcal{C}(\boldsymbol{x}_*)$ in Eq. (3.6) follows the distribution (Odiowei and Cao, 2009):

$$\frac{n(n-m)}{(n-1)(n+1)m}\mathcal{C}(\boldsymbol{x}_*) \sim F(m, n-mx) \tag{3.7}$$

where F denotes the Snedecor's distribution.

The multivariate SCC is the plot of monitoring statistics against time, along with control limits which are determined by the chosen statistical significance level (e.g. 95%).

3.3.3. Apply univariate control charts to multivariate cases using PCA

The multivariate SCC might be inapplicable for high-dimension systems with collinearities (Bersimis et al., 2005) since collinearities might result in model parameters with high uncertainty levels, and the increase of inaccuracy in statistics (De Marco and Nóbrega, 2018). To address the collinearity issue, a common method is the use of projection methods, such as Principal Component Analysis (PCA) and Partial Least Sqaures (PLS). In this work, PCA is selected due to its abilities of extracting uncorrelated features and reducing dimensionality (Jolliffe, 2011).

Feature extraction: Principal Component Analysis (PCA)

PCA has been successfully used in MSPM. The main objective of PCA is to derive a few independent components from high-dimension data. The specific operation is to linearly project highly correlated multivariate data to a lower dimension space in which variables are uncorrelated. The extracted components are also known as features.

Let $\tilde{X} = [\tilde{\boldsymbol{x}}_1, \tilde{\boldsymbol{x}}_2, \ldots, \tilde{\boldsymbol{x}}_n] \in \mathbb{R}^{m \times n}$ denote a normalised measurement matrix with zero mean and unit variance, where $\tilde{\boldsymbol{x}}_i$ represents the i-th measurement vector of m-dimension. To linearly project $\tilde{X}$ from a m-dimension space to a v-dimension principal component space, eigenvectors of $\tilde{X}$ are required. There are two popular methods to calculate the eigenvectors:

– Eigenvalue Decomposition (ED):

$$\tilde{\Sigma} = \mathbf{V}\mathbf{L}\mathbf{V}^\top \tag{3.8}$$

where $\mathbf{V} \in \mathbb{R}^{m \times m}$ is a matrix of eigenvectors. Each column of $\mathbf{V}$ is an eigenvector. $\mathbf{L} \in \mathbb{R}^{m \times m}$is a diagonal matrix with eigenvalues in descending order. $\tilde{\Sigma} \in \mathbb{R}^{m \times m}$ is the sample covariance matrix calculated by

$$\tilde{\Sigma} = \frac{\tilde{X}\tilde{X}^\top}{n-1}. \tag{3.9}$$

– Singular Value Decomposition (SVD): $\tilde{X}$ can be factorised as

$$\tilde{X}^\top = \mathbf{U}\mathbf{S}\mathbf{L}^\top \tag{3.10}$$

where $\mathbf{U} \in \mathbb{R}^{n \times n}$ is a matrix containing orthogonal eigenvectors of $\tilde{X}^\top\tilde{X}$, $\mathbf{L} \in \mathbb{R}^{m \times m}$ is a matrix containing orthogonal eigenvectors of $\tilde{X}\tilde{X}^\top$ and $\mathbf{S} \in \mathbb{R}^{n \times m}$ is a rectangular diagonal matrix with square roots of eigenvalues in descending order (Baker, 2005) .

Denote v as the first v eigenvectors with a certain accumulated explained variance. v is selected by

$$\frac{\sum_{i=1}^{v} l_i}{\sum_{i=1}^{m} l_i} \geq \text{Var}_{acc} \tag{3.11}$$

where l_i is the ith eigenvalue corresponding to eigenvalues in the descending order and Var_{acc} is the value of accumulated explained variance. A reduced matrix $\mathbf{V}$ by selecting the first v columns of it is

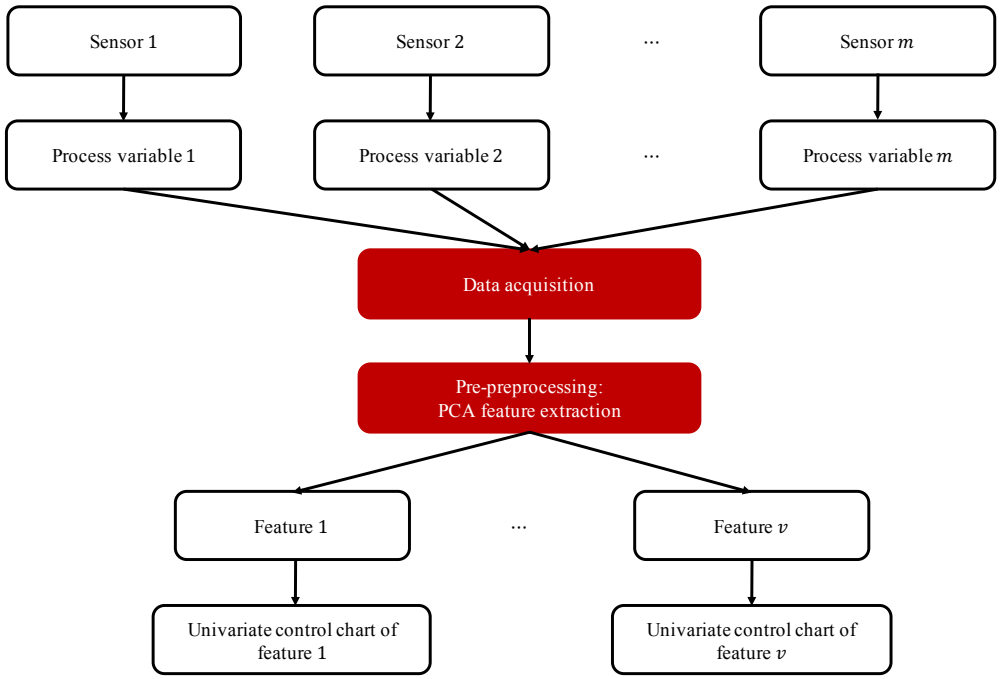


Figure 3.1: The diagram of generating multiple univariate control charts: data are recorded from various sensors, each sensor documenting a specific process variable. Given the collected data, PCA is applied to extract uncorrelated features. As features are independent to each other, univariate control charts can be obtained according to Section 3.3.1, furthermore can be used for monitoring processes.

written as $\mathbf{V}_v \in \mathbb{R}^{m\times v}$ which is also called the projection matrix. The projection is performed by:

$$Y = \tilde{X}^\top \mathbf{V}_v \tag{3.12}$$

where $Y = [\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_n]^\top \in \mathbb{R}^{n\times v}$ is the projection of $\tilde{X}^\top$ in a v-dimension space. $\boldsymbol{y}_i \in \mathbb{R}^v$ is the projection of $\boldsymbol{x}_i$. Let $y_{\{j,i\}}$ denote the j-th feature of $\boldsymbol{y}_i$ where $j = 1, \ldots, v$. The j-th feature of Y is written as $Y_j = \{y_{\{j,1\}}, y_{\{j,2\}}, \ldots, y_{\{j,n\}}\}$. $Y_j, \forall j$ are independent to each other. Another way to calculate Y is the product of $\mathbf{U}_v \in \mathbb{R}^{n\times v}$ and $\mathbf{S}_v \in \mathbb{R}^{v\times v}$. Although both methods can obtain the projection matrix, when $\tilde{\Sigma}$ is either singular or numerically very close to singular, Erichson et al. (2016) pointed out that SVD takes advantage of its numerically stable matrix decomposition.

Multiple univariate control charts

The objective of using control charts is to visualise the variation of a process with time stamps displayed on the horizontal axis and monitoring statistics on the vertical axis (Stijn, 2018). When there are a large quantity of inspected variables, to enable the visualisation, it is necessary to convert multiple variables into a univariate monitoring statistics. The conversion can be achieved using the multivariate SCC. However, if variables are correlated to each other, the monitoring statistics might bring bias to detection results. Therefore, often pre-processing steps, for example, feature extraction, are employed to mitigate the bias. In this thesis, due to the use of PCA, extracted features can be treated independently. With this property, different from generating a univariate monitoring statistics via the multivariate SCC,

this work retains the use of univariate control charts.

The diagram in Fig. 3.1 illustrates how univariate control charts might be generated. For a system equipped with numerous sensors, there are a wide range of process variables generated, leading to high dimensionality of process data. When the process is operating, sensor readings are collected in data acquisition systems. After PCA is applied to process data, only features retaining most variances are kept. The number of remaining features is less than the number of process variables. Thus, dimension reduction is achieved. Also, since the features are independent to each other, the detection result given by each feature should not be influenced by others. In this sense, it is possible to implement univariate control charts over individual features instead of the original process variables, and to obtain individual detection results.

3.4. Density estimation approaches for univariate data

The use of univariate control charts requires the probability distribution of each feature to be known. In this section, three density estimation methods for univariate data are introduced and discussed.

3.4.1. Histograms

The histogram was first introduced by Pearson (1895), and considered one of the most simple and widely used density estimators (Bedoui, 2013). To plot the histogram, it is needed to specify an origin y_0 and a bin width b. Given this information, bins of the histograms are defined as $B_k = [y_0 + kb, y_0 + (k+1)b)$ where $k = \ldots, -1, 0, 1, \ldots$. For a random feature sample $y_* \in Y_j$, the estimated density $\hat{P}(y_*)$ at interval B_k, follows:

$$\hat{P}(y_*) = \frac{n(y_{\{j,i\}} \in B_k)}{nb} \tag{3.13}$$

where $n(y_{\{j,i\}} \in B_k)$ denotes the number of $y_{\{j,i\}}$ in the interval B_k.

The histogram is a useful tool to present the density of a set of univariate data. However, the choices of the origin point y_0 and the bin width may have quite an effect on the density estimation (Silverman, 1986). Moreover, any discontinuity in the histogram indicates that the derivatives of the Probability Density Function (PDF) are not obtainable (Silverman, 1986).

3.4.2. Parameter estimation for Gaussian distributions

The Gaussian distribution is a continuous distribution and is an approximation of the probability distribution of a large quantity of independent random samples according to the Central Limit Theorem. The PDF is (Gubner, 2006):

$$P(y_*) = \frac{1}{\sqrt{2\pi\sigma_j^2}} \exp\left[-\frac{1}{2}\frac{(y_* - \mu_j)^2}{\sigma_j^2}\right] \tag{3.14}$$

where μ_j and σ_j are the Gaussian parameters for the j-th feature, and can be estimated from Y_j according to Lee et al. (2015):

$$\begin{aligned}\hat{\mu}_j &= \frac{1}{n}\sum_{i=1}^{n} y_{\{j,i\}} \\ \hat{\sigma}_j &= \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(y_{\{j,i\}} - \hat{\mu}_j\right)^2}\end{aligned} \tag{3.15}$$

Given Gaussian distributions, the univariate SCC can be applied.

3.4.3. Nonparametric method: Kernel Density Estimation

The PDF estimation solution given in Section 3.4.2 is suitable in the cases where univariate data follow Gaussian distributions. To deal with the non-Gaussian cases, Kernel Density Estimation (KDE) can be applicable, particularly for univariate random processes (Bowman and Azzalini, 1997). The estimated PDF at point y_* is

$$\begin{aligned}z_j &= \frac{y_* - y_{\{j,i\}}}{h} \\ \hat{P}(y_*) &= \frac{1}{nh}\sum_{i=1}^{n} K(z_j)\end{aligned} \tag{3.16}$$

where h is the bandwidth and $K(\cdot)$ is a kernel function. The overall performance of the distribution estimation is significantly associated with the selection of bandwidth. Chen (2017) illustrated that an excessively small value of h will result in a rough distribution plot whereas an excessively large h will over-smooth the plot. There are many equally valid ways of determining the bandwidth (Odiowei and Cao, 2009). The optimal bandwidth h_{opt} can be roughly estimated by (Bowman and Azzalini, 1997):

$$h_{\text{opt}} \approx 1.06\sigma_j n^{-0.2}. \tag{3.17}$$

The choice of the kernel is less crucial for density estimation of independently and identically distributed random variables (Shen and Agrawal, 2006). In this thesis, the Gaussian kernel is used:

$$K(z_j) = \frac{e^{\frac{-z_j^2}{2}}}{\sqrt{2\pi}}. \tag{3.18}$$

3.5. BaFFle algorithm

3.5.1. Nomenclature

Table 3.2 lists the mathematical symbols used in the BaFFle algorithm as well as their definitions. Please note that these symbols and definitions are only relevant for this chapter.

3.5.2. Fault detection across individual features

In traditional anomaly detection algorithms, any incoming samples exceeding control limits are considered as anomalies. Nevertheless, within such a detection mechanism, false alarm rates might be high. For example, even if a process is under control, some individual samples falling outside the control limits

will be identified as abnormal operation, due to random fluctuations. Reducing false alarms by setting wide control limits might cause high numbers of missed alarms. To balance the false and missed alarms, a mechanism of warning and detection is introduced in this work.

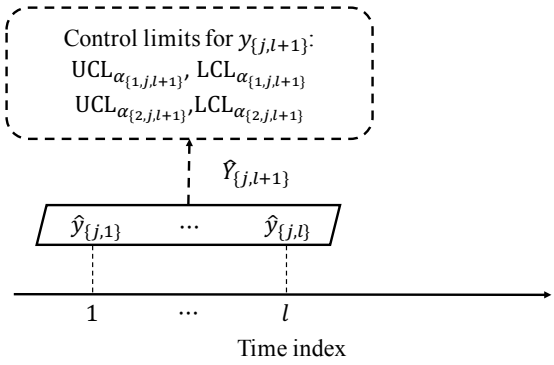
Table 3.2: Nomenclature for the BaFFle algorithm

Symbol	Description
$\boldsymbol{y}_t = [y_{\{1,t\}}, \ldots, y_{\{v,t\}}]^\top$	A vector containing v extracted features of a multivariate data sample at time t
$y_{\{j,t\}}$	The j-th feature of $\boldsymbol{y}_t$
$\hat{y}_{\{j,t\}}$	A sample in the moving window associated with $y_{\{j,t\}}$
l	The width of the moving window
$\hat{Y}_{\{j,t\}} = \{\hat{y}_{\{j,t-l\}}, \hat{y}_{\{j,t-l+1\}}, \ldots, \hat{y}_{\{j,t-1\}}\}$	The data set for deriving control limits for $y_{\{j,t\}}$
$\alpha_{\{1,j,t\}}$	The confidence level for calculating the control limits to alert if $y_{\{j,t\}}$ is an anormaly
$\alpha_{\{2,j,t\}}$	The confidence level for calculating the control limits to determine $y_{\{j,t\}}$ is normal or abnormal
$\text{UCL}_{\alpha_{\{1,j,t\}}}$	Given $\alpha_{\{1,j,t\}}$, the UCL for $y_{\{j,t\}}$
$\text{LCL}_{\alpha_{\{1,j,t\}}}$	Given $\alpha_{\{1,j,t\}}$, the LCL for $y_{\{j,t\}}$
$\text{UCL}_{\alpha_{\{2,j,t\}}}$	Given $\alpha_{\{2,j,t\}}$, the UCL for $y_{\{j,t\}}$
$\text{LCL}_{\alpha_{\{2,j,t\}}}$	Given $\alpha_{\{2,j,t\}}$, the LCL for $y_{\{j,t\}}$
λ	A contribution coefficient
$s_\uparrow$	The step rate of increasing $\alpha_{\{2,j,t\}}$
$s_\downarrow$	The step rate of reducing $\alpha_{\{2,j,t\}}$
$W_{\{j,t\}}$	Warning result of $y_{\{j,t\}}, \in \{0, 1\}$
$D_{\{j,t\}}$	Detection result of $y_{\{j,t\}}, \in \{0, 1\}$
$n_{D_{\{j,t\}}=1}$	The counts of $D_{\{j,t\}}$ equal to 1, $j = 1, \ldots, v$
$n_{D_{\{j,t\}}=0}$	The counts of $D_{\{j,t\}}$ equal to 0, $j = 1, \ldots, v$
$D_{\boldsymbol{y}_t}$	The detection result of $\boldsymbol{y}_t, \in \{0, 1\}$

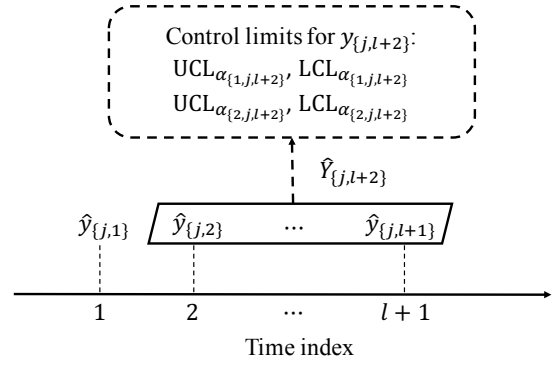
Determination of control limits

As discussed in Section 3.3.3, multiple univariate control charts are adopted in the BaFFle algorithm. Moreover, the control charts are designed to be adaptive to account for the changes over time.

Fig. 3.2 illustrates a sliding window may be used to calculate the adaptive control limits. The BaFFle algorithm starts to function at the time stamp $t = l + 1$ due to the requirement of collecting l samples for each extracted feature. When $t = l + 1$, the control limits for monitoring $y_{\{j,l+1\}}$ are derived from a training set $\hat{Y}_{\{j,l+1\}}$ as the initial window shown in Fig. 3.2. The samples in $\hat{Y}_{\{j,l+1\}}$ are initialised by $\hat{y}_{\{j,1\}} = y_{\{j,1\}}, \hat{y}_{\{j,2\}} = y_{\{j,2\}}, \ldots, \hat{y}_{\{j,l\}} = y_{\{j,l\}}$. Given $\hat{Y}_{\{t,j\}}$, the control limits for a confidence level



(a) Sliding window of the j–th feature at time $l+1$



(b) Sliding window of the j–th feature at time $l+2$

Figure 3.2: Illustration of sliding window: to generate adaptive control limits, a sliding window is used in the BaFFle algorithm. For each feature, a window of length l is created. Sub-figure (a) gives the window of the j-th feature at time $l+1$. The control limits for monitoring $y_{j,l+1}$ is derived from $\hat{Y}_{\{j,l+1\}} = \{\hat{y}_{\{j,1\}}, \hat{y}_{\{j,2\}}, \ldots, \hat{y}_{\{j,l\}}\}$ which are initialised with $y_{j,1}, y_{j,2}, \ldots, y_{j,l}$. $y_{j,t}$ is the j-th feature of $\boldsymbol{y}_t$. Given $\hat{Y}_{\{j,l+1\}}$, the control limits $\text{UCL}_{\alpha_{\{1,j,l+1\}}}$, $\text{LCL}_{\alpha_{\{1,j,l+1\}}}$, $\text{UCL}_{\alpha_{\{2,j,l+1\}}}$ and $\text{LCL}_{\alpha_{\{2,j,l+1\}}}$ are derived using Eq. 3.19. To generate $\hat{Y}_{\{j,l+2\}}$ for calculating $\text{UCL}_{\alpha_{\{1,j,t\}}}$, $\text{LCL}_{\alpha_{\{1,j,t\}}}$, $\text{UCL}_{\alpha_{\{2,j,t\}}}$ and $\text{LCL}_{\alpha_{\{2,j,t\}}}$, the window slides to the next time stamp, resulting in $\hat{y}_{\{j,2\}}, \hat{y}_{\{j,3\}}, \ldots, \hat{y}_{\{j,l+1\}}$, as shown in sub-figure (b). The value of $\hat{y}_{\{j,l+1\}}$ is determined according to Eq. 3.22, 3.23 and 3.24.

$\alpha_{\{1,j,t\}}$ are calculated by:

$$P(\text{LCL}_{\alpha_{\{1,j,t\}}} < \hat{y}_* < \text{UCL}_{\alpha_{\{1,j,t\}}}) = \int_{\text{LCL}_{\alpha_{\{1,j,t\}}}}^{\text{UCL}_{\alpha_{\{1,j,t\}}}} p(\hat{y}_*) d\hat{y}_* = 1 - \alpha_{\{1,j,t\}} \tag{3.19}$$

where $p(\hat{y}_*)$ is the PDF of $\hat{Y}_{\{t,j\}}$. In the same way, the control limits, $\text{UCL}_{\alpha_{\{2,j,t\}}}$ and $\text{LCL}_{\alpha_{\{2,j,t\}}}$, for a given confidence level $\alpha_{\{2,j,t\}}$, can be derived. When $p(\hat{y}_*)$ is subject to a Gaussian distribution, control limits can be directly obtained by the univariate SCC in Eq. (3.4); otherwise, the KDE can be applied to estimate the PDF, then find the confidence intervals for given $\alpha_{\{1,j,t\}}$ and $\alpha_{\{2,j,t\}}$. In this thesis, $\alpha_{\{1,j,l+1\}}$ and $\alpha_{\{2,j,l+1\}}$ are initialised with 99.7% and 99.9% corresponding to the deviation of $3\sigma_j$ and $4\sigma_j$ in the Gaussian distribution.

Fault detection

The alert system for $y_{\{j,t\}}$ works as

$$W_{\{j,t\}} = \begin{cases} 0, & \text{LCL}_{\alpha_{\{1,j,t\}}} < y_{\{j,t\}} < \text{UCL}_{\alpha_{\{1,j,t\}}} \\ 1, & \text{otherwise.} \end{cases} \tag{3.20}$$

where value 0 represents normal operation whereas 1 represents abnormal operation. The determination of if $y_{\{j,t\}}$ is abnormal is according to

$$D_{\{j,t\}} = \begin{cases} 0, & \text{LCL}_{\alpha_{\{2,j,t\}}} < y_{\{j,t\}} < \text{UCL}_{\alpha_{\{2,j,t\}}} \\ 1, & \text{otherwise.} \end{cases} \tag{3.21}$$

where the values of 0 and 1 have the same meanings as in Eq. (3.20).

Update of monitoring models

For a new time stamp (see Fig. 3.2), the window of length l slides one sample to have $\hat{Y}_{\{j,l+2\}} = \{\hat{y}_{\{j,2\}}, \hat{y}_{\{j,3\}}, \ldots, \hat{y}_{\{j,l+1\}}\}$ in which $\hat{y}_{\{j,l+1\}}$ is created according to the detection results of $y_{\{j,l+1\}}$ (see Eq. (3.20) and Eq. (3.21)). Also, to increase the sensitivity of the detection algorithm, $\alpha_{\{2,j,t\}}$ is adjusted over time, depending on $W_{\{j,t\}}$ and $D_{\{j,t\}}$. The update of monitoring models are shown as follows:

– $W_{\{j,t\}} = 1$ and $D_{\{j,t\}} = 1$. When $y_{\{t,j\}}$ is identified as an anomaly given $\alpha_{\{1,j,t\}}$ and $\alpha_{\{2,j,t\}}$, the calculations of $\alpha_{\{2,j,t+1\}}$ and $\hat{y}_{\{j,t\}}$ follows:

$$\begin{aligned} \alpha_{\{2,j,t+1\}} &= \alpha_{\{2,j,t\}} \\ \hat{y}_{\{j,t\}} &= y_{**}. \end{aligned} \tag{3.22}$$

where y_{**} denotes a random sample from $\hat{Y}_{\{j,l+1\}}$.

– $W_{\{j,t\}} = 1$ and $D_{\{j,t\}} = 0$. When $y_{\{j,t\}}$ is indicated as an anomaly by $\alpha_{\{1,j,t\}}$, but recognised as a normal sample by $\alpha_{\{2,j,t\}}$, it is necessary to narrow the normal operation region defined by $\alpha_{\{2,j,t\}}$. To this end, the value of $\alpha_{\{2,j,t+1\}}$ should be adjusted smaller relative to the value of $\alpha_{\{2,j,t\}}$ while its minimum value is restricted to $\alpha_{\{1,j,l+1\}}$.

$$\begin{aligned} \alpha_{\{2,j,t+1\}} &= \max(\alpha_{\{1,j,l+1\}}, \alpha_{\{2,j,t\}} - s_{\downarrow}) \\ \hat{y}_{\{j,t\}} &= \lambda y_{\{j,t\}} + (1-\lambda) y_{**}. \end{aligned} \tag{3.23}$$

where $s_{\downarrow}$ is the step rate. The value of $s_{\downarrow}$ is in the range $[0, \alpha_{\{2,j,l+1\}} - \alpha_{\{1,j,l+1\}}]$. When $s_{\downarrow}$ takes its maximum value, $\text{UCL}_{\alpha_{\{2,j,t\}}}$ and $\text{LCL}_{\alpha_{\{2,j,t\}}}$ are invariant along time. In this work, $s_{\downarrow} = 0.01\%$. $\lambda \in [0, 1]$ is a coefficient which is used to determine the portion of $y_{\{j,t\}}$ contributing to $\hat{y}_{\{j,t\}}$. λ is set with the value of 0.1 in order to mitigate the impact of a false negative $y_{\{j,t\}}$ on $\hat{y}_{\{j,t\}}$

– $W_{\{j,t\}} = 0$ and $D_{\{j,t\}} = 0$. In this case,

$$\begin{aligned} \alpha_{\{2,j,t+1\}} &= \min(\alpha_{\{2,j,l+1\}}, \alpha_{\{2,j,t\}} + s_{\uparrow}) \\ \hat{y}_{\{j,t\}} &= \lambda y_{\{j,t\}} + (1-\lambda) y_{\{j,t-1\}}. \end{aligned} \tag{3.24}$$

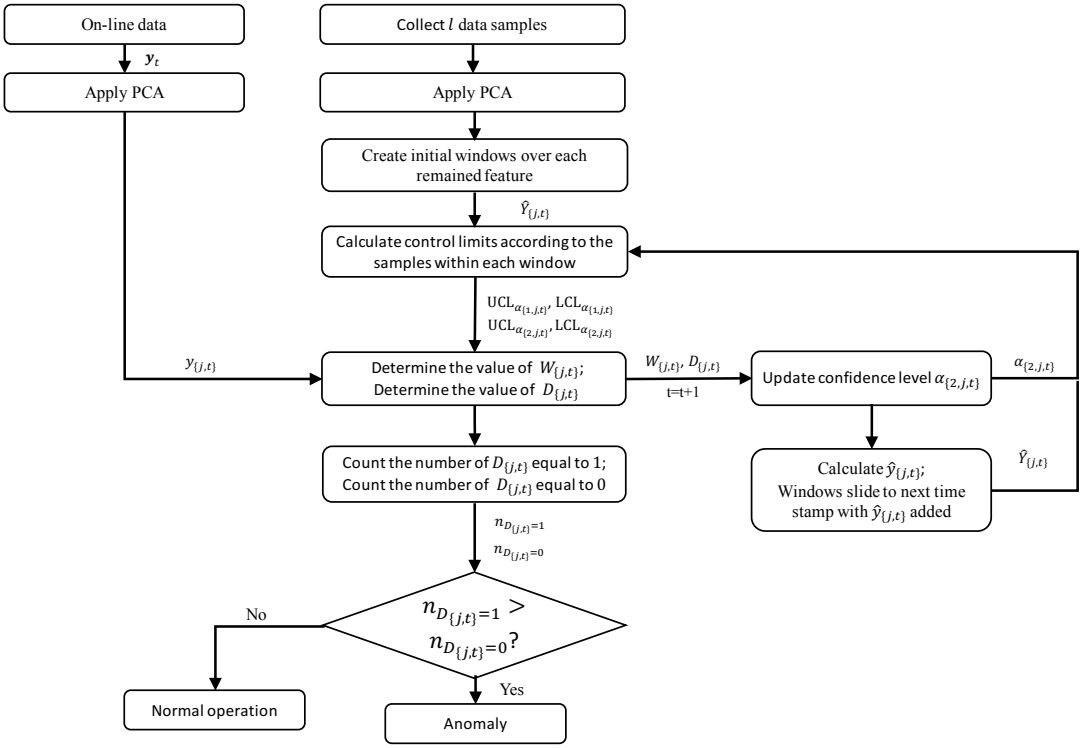


Figure 3.3: Flowchart of the BaFFle algorithm: the initial monitoring model is trained with the first l samples. PCA is applied to extract independent features $\hat{Y}_{j,t}, \forall j$. Then control limits over each feature are calculated using Eq. 3.19 to have $\text{UCL}_{\alpha\{1,j,t\}}$, $\text{LCL}_{\alpha\{1,j,t\}}$, $\text{UCL}_{\alpha\{2,j,t\}}$ and $\text{LCL}_{\alpha\{2,j,t\}}$. Given these control limits, the binary monitoring indicators $W_{j,t}$ and $D_{j,t}$ are determined for $y_{j,t}$. $y_{j,t}$ is the j-th feature of sample $\boldsymbol{y}_t$. If there are more $D_{j,t}$ of value 1 than of value 0, sample $\boldsymbol{y}_t$ is identified as abnormal. In addition, the combinations of the values of $W_{j,t}$ and $D_{j,t}$ guide how to update the confidence level and calculate $\hat{y}_{j,t}$ for the next time stamp (see Eq. 3.22, 3.23 and 3.24).

where $s_\uparrow \in [0, \alpha_{\{2,j,l+1\}} - \alpha_{\{1,j,l+1\}}]$ is the step rate of increasing $\alpha_{\{2,j,t\}}$. Analogous to $s_\downarrow$, for $s_\uparrow$ with its maximum value, the control limits corresponding to $\alpha_{\{2,j,t\}}$ are unchanged over time. In practice, $s_\uparrow$ is set according to user requirements. It is suggested to let $s_\uparrow > s_\downarrow$, given which, control limits would return back to previous values faster so as to limit the number of false alarms. In this thesis, $s_\uparrow = 0.05\%$.

Decision-making in the BaFFle algorithm

In Chapter 2, three voting-based decision-making strategies have been briefly introduced, respectively, unanimous voting, simple majority voting and plurality/majority voting. Since in multi-label classification, the majority voting is advantageous due to its low missed alarm rate comparing with the

unanimous voting and simple majority voting. In the BaFFle algorithm, the majority voting strategy is employed for determining if $\boldsymbol{y}_t$ is normal. Nevertheless, for binary classification cases, simple majority and majority votings are one and the same. The fault detection in the BaFFle is expressed as:

$$D_{\boldsymbol{y}_t} = \begin{cases} 0, & n_{D_{\{j,t\}=1}} < n_{D_{\{j,t\}=0}} \\ 1, & n_{D_{\{j,t\}=1}} > n_{D_{\{j,t\}=0}}. \end{cases} \tag{3.25}$$

where values 0 and 1 respectively represents normal and abnormal. When $n_{D_{\{j,t\}=1}} = n_{D_{\{j,t\}=0}}$, it means that $\boldsymbol{y}_t$ is undetermined .

3.5.3. Workflow of the BaFFle algorithm

The flowchart shown in Fig. 3.3 summarises the procedure of anomaly detection using the BaFFLe algorithm. The PCA-based feature extraction method is applied to a small amount of multivariate data samples collected on-line. Over each retained feature, an initial window is created, the length of which is same as the number of collected samples. Control limits for warning ($\text{UCL}_{\alpha\{1,j,t\}}$ and $\text{LCL}_{\alpha\{1,j,t\}}$) and identifying ($\text{UCL}_{\alpha\{2,j,t\}}$ and $\text{LCL}_{\alpha\{2,j,t\}}$) anomalies are calculated based on the samples within each window (see Eq. (3.19)). Subsequently, the values of warning and detection indicators are determined by Eq. (3.20) and (3.21), respectively. The identification of whether or not a sample $y_{j,t}$ is anomalous is performed by a majority voting strategy. If there are more features supporting the identify of normal, the process operation of time t is considered as normal; otherwise, abnormal. In addition, the control charts are variant along time, determined by the samples within monitoring windows and confidence levels.

3.6. Summary

Chapter 3 has investigated the statistics for evaluating the fault detection algorithms. The formulation of these statistics have shown that reducing the number of false and missed alarms would contribute to improving monitoring performance. Through literature review, it has been found that adaptability could be a direction to work on. A Binary Classifier for Fault Detection (BaFFle) algorithm have been proposed in this chapter, which can adaptively update monitoring thresholds. To derive unbiased multiple univariate control charts, Principal Component Analysis (PCA) has been adopted. Kernel Density Estimation (KDE) has been used to estimate the distribution of non-Gaussian data. A moving window has been involved in the BaFFle to dynamically incorporate and discard data. A flowchart has been given to demonstrate the use of the BaFFle algorithm.

4. Dirichlet Process-Gaussian Mixture Models (DP-GMMs)

Production demand and loading conditions might be varying in the production course, resulting in various operating modes. The characteristics of data generated by each operating mode might be different from each other. Thus, rather than taking all the data as a whole, data of individual modes should be treated as a cluster, and analysed individually.

In order to design cluster-based monitoring algorithms, it is necessary to partition unlabelled historical data into several groups. In this chapter, the focus is the Dirichlet Process-Gaussian Mixture Models (DP-GMMs) as well as their application to the clustering analysis. At the beginning of the chapter, the problem regarding DP-GMMs based clustering is stated. After that, several typical distributions, such as the Multinomial distribution and the Dirichlet distribution, are revisited. Also, the concept of the *conjugate prior* is introduced in the context of Bayes' theorem. The relationships between the Dirichlet distribution and other distributions are highlighted. Next, finite GMMs are reviewed and the DP-GMMs are introduced. A discussion on how to optimally choose the hyper-parameters in the DP-GMMs is given. The derivation of the Normal Inverse Wishart (NIW) distribution is shown. The NIW distribution is commonly used as the prior knowledge in DP-GMMs. The chapter continues with the review of Gibbs sampling. Bayesian inference in the framework of the finite GMMs and the DP-GMMs are presented. The implementation of the DP-GMMs clustering is demonstrated via a multimode simulation model. Moreover, the parameter initialisation of the NIW distribution is analysed. An initialisation step of data normalisation and specific parameter settings are proposed. The impact of various settings on the parameter estimation of Gaussian distributions as well as on clustering results is demonstrated. In addition, a monitoring framework incorporating the DP-GMMs clustering is proposed. Finally, the chapter ends with a summary. This chapter includes work which has previously been reported in (Tan et al., 2019, 2020).

4.1. Problem statement

Multimodality may exist in data recorded from multimode processes. To analyse the properties and characteristics of data in each mode, it is necessary to partition multimodal data corresponding to their modes. To this end, data clustering analysis based on the DP-GMMs is a prominent partition approach without specifying the number of clusters/modes. Before implementing the DP-GMMs, there are two hyper-parameters, concentration parameter α and base function G_0, to be assigned. The value of α and the form of the base distribution are important which will affect the clustering performance. Many research efforts have focused on their assignments. Escobar and West (1995) proposed that learning about α from the to-be-clustered data may be addressed by incorporating α into the clustering analysis. In terms of G_0, its most widely used form is the NIW distribution which comprised four parameters, respectively, mean vector $\boldsymbol{u}_0$, a positive scalar κ_0, the number of degrees of freedom ν_0 and covariance matrix Λ_0.

The value choices of $\boldsymbol{u}_0$, ν_0 and Λ_0 have been mentioned and well-researched in literature (for example, see (Görür and Rasmussen, 2010; Nydick, 2012; Alvarez et al., 2014; Schuurman et al., 2016)). A guide of choosing κ_0 is to assign an extremely small, approaching zero value (Gelman et al., 2013). However, how to determine a specific value of κ_0 as well as its effect on clustering performance have been afforded less attention. Without this knowledge, the clustering results may be inaccurate and unreliable, the use of which may also hinder the effectiveness of monitoring algorithms designed based on clustering results. To solve this problem, the main contributions of this chapter are as follows:

- Investigate the influence of parameters on the accuracy of DP-GMMs clustering, particularly parameter κ_0;
- Propose a method to improve the accuracy of the DP-GMMs clustering, in which the determination of κ_0 is addressed;
- Validate the proposed method in a multimode simulation model.

4.2. Preliminary

4.2.1. Discrete distributions

Grinstead and Snell (1998) gave the definition of *Bernoulli trials process* as follows: suppose that an experiment has two potential outcomes, which can be **success** and **failure**. A *Bernoulli trials process* is a sequence of n such experiments. The probability of **success** is p, and p is the same in each experiment, not affected by any previous outcomes. The probability of **failure** is given by $1 - p$.

Bernoulli distribution

Let x be an independently and identically distributed random variable. The random binary outcome x of a single experiment in a Bernoulli trials process follows a Bernoulli distribution. The Probability Mass Function (PMF) of the Bernoulli distribution is (Bernoulli, 1713):

$$P(x) = \begin{cases} p, & \text{when } x = 1 \text{ stands for } \textbf{success} \\ 1 - p = q, & \text{when } x = 0 \text{ stands for } \textbf{failure}. \end{cases} \tag{4.1}$$

Binomial distribution

x_i is the outcome in the i-th experiment. In a binary-outcome situation, $x_i = 1$ if the outcome is **success** and $x_i = 0$ otherwise. The outcomes of n chance experiments can be expressed in the form $S = x_1 + x_2 + \cdots + x_n$. S is a random variable, counting the number of outcomes that are **success**, subject to the Binomial distribution, written as $S \sim \text{Binomial}(n, p)$. The probability of having s **success**es in n experiments is (Gubner, 2006)

$$P(S = s) = \binom{n}{s} p^s (1 - p)^{n-s} \quad \text{for } s = 0, 1, \ldots, n \tag{4.2}$$

where $\binom{n}{s} = \frac{n!}{s!(n-s)!}$. $n = 1$ leads to the Bernoulli distribution which is a special case of the Binomial distribution.

Category distribution

When a single experiment has more than two mutually exclusive outcomes (e.g. $x \in \{1, 2, \ldots, J\}$), the Bernoulli distribution turns into the Category distribution Cat$(p_1, p_2, \ldots, p_J)$ where J is the number of possible outcomes and p_j is a fixed probability value corresponding to the outcome $x = j$. The PMF is (Murphy, 2012):

$$P(x = j) = p_j \quad \forall j \tag{4.3}$$

where $\sum_{j=1}^{J} p_j = 1$

Multinomial distribution

The Multinomial distribution arises from a generalisation of the Binomial distribution to the situations where each experiment is subject to the Category distribution. It models the probability of a count of observations in a sequence of independent experiments. Suppose that there are J possible outcomes and their corresponding fixed probabilities are $p_1, p_2, \ldots, p_J$. $s_1, s_2, \ldots, s_J$ are used for representing the counts result after n experiments where s_j is the counts of the j-th outcome. Following the Multinomial distribution, the probability of the results $s_1, s_2, \ldots, s_J$ is (Murphy, 2012):

$$P(s_1, s_2, \ldots, s_J) = \frac{n!}{s_1! \ldots s_J!} p_1^{s_1} p_2^{s_2} \ldots p_J^{s_J} \tag{4.4}$$

where $\sum_{j=1}^{J} p_j = 1$ and $\sum_{j=1}^{J} s_j = n$.

4.2.2. Continuous distributions

In Section 4.2.1, $p, p_1, p_2, \ldots, p_J$ are fixed values. In this section, cases where $p, p_1, p_2, \ldots, p_J$ are flexible are taken into consideration.

Beta distribution

p is assumed to be a random variable on the interval $(0, 1)$, subject to the Beta distribution parameterised by a and b, written as $p \sim$ Beta(a, b). The Probability Density Function (PDF) of the Beta distribution is (Murphy, 2012):

$$P(p) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} p^{a-1}(1-p)^{b-1} \tag{4.5}$$

where $\Gamma()$ denotes the gamma function, $\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}$ is a constant term, $a > 0$ and $b > 0$.

Dirichlet distribution

Assume a random PMF Q with J components having variables $p_1, p_2, \ldots, p_J$ accordingly:

$$\sum_{j=1}^{J} p_j = 1, \quad p_j > 0, \quad \forall j. \tag{4.6}$$

In addition, let $\alpha_1, \alpha_2, \ldots, \alpha_J$ have

$$\sum_{j=1}^{J} \alpha_j = \alpha, \quad \alpha_j > 0, \quad \forall j. \tag{4.7}$$

As the Dirichlet distribution can be considered as a distribution over PMFs (Frigyik et al., 2010), the PMF Q of length J can be sampled from the Dirichlet distribution parameterised by α, denoted as $Q \sim \text{Dir}(\alpha)$ or $(p_1, p_2, \ldots, p_J) \sim \text{Dir}(\alpha_1, \alpha_2, \ldots, \alpha_J)$, with the probability

$$P(Q) = \frac{\Gamma(\alpha)}{\prod_{j=1}^{J} \Gamma(\alpha_j)} \prod_{j=1}^{J} p_j^{\alpha_j - 1} \tag{4.8}$$

If $(p_1, p_2, \ldots, p_J) \sim \text{Dir}(\alpha_1, \alpha_2, \ldots, \alpha_J)$, some key properties of the Dirichlet distribution are summarised as follows (Xing, 2014):

– Coalesce rule for reducing the components of the Dirichlet distribution :

$$(p_1 + p_2, p_3, \ldots, p_J) \sim \text{Dir}(\alpha_1 + \alpha_2, \alpha_3, \ldots, \alpha_J) \tag{4.9}$$

– Expansion rule for increasing the components of the Dirichlet distribution: if $(\gamma_1, \gamma_2, \ldots, \gamma_{n_+}) \sim \text{Dir}(\alpha_1\eta_1, \alpha_1\eta_2, \ldots, \alpha_1\eta_{n_+})$ and $\sum_{i=1}^{n_+} \eta_i = 1$ where n_+ denotes the number of the increased components, then

$$(p_1\gamma_1, p_1\gamma_2, \ldots, p_1\gamma_{n_+}, p_2, p_3, \ldots, p_J) \sim \text{Dir}(\alpha_1\eta_1, \alpha_1\eta_2, \ldots, \alpha_1\eta_{n_+}, \alpha_2, \alpha_3, \ldots, \alpha_J) \tag{4.10}$$

The density plots of the Dirichlet distribution in a three-dimension space obtained by setting various values of the parameter α are demonstrated in (Frigyik et al., 2010). For detailed derivations and interpretations of the Dirichlet distribution, readers are guided to references, for example (Frigyik et al., 2010; Gelman et al., 2013; Paisley, 2015; Lin, 2016)

When $J = 2$, the Dirichlet distribution is equivalent to the Beta distribution. To make the connection clear, note that if a random variable p has a Beta distribution $\text{Beta}(a, b)$, the Beta distribution can be rewritten as $(p, 1-p) \sim \text{Dir}(a, b)$(Frigyik et al., 2010) .

4.2.3. Relationships between distributions

The probabilities $p, p_1, \ldots, p_J$ in the discrete distributions described in 4.2.1 are fixed values. However, the values of $p, p_1, \ldots, p_J$ are generally unknown. A common way to describe the uncertainties concerning $p, p_1, \ldots, p_J$ is the usage of a family of probability density distributions, which constitute an approach to estimate $p, p_1, \ldots, p_J$. Section 4.2.2 gives two potential distributions.

Bayes' theorem

Bayes' theorem is formulated as (see, e.g., Stuart and Ord (1994)):

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}. \tag{4.11}$$

where A could be a parameter, hypothesis or assumption and B is an evidence. $P(A)$ is a probability measure about A, called the *prior* probability. $P(B|A)$ is the *likelihood* probability, representing the probability of evidence B being observed under hypothesis A. $P(A|B)$ is the *posterior* probability of hypothesis A is true given evidence B. $P(B)$ is the probability of observing evidence B, also called *marginal likelihood*. Bayes' theorem describes how the probability of A changes when evidence B is available.

Conjugate prior

The term, *conjugate prior*, was first introduced in the work (Schlaifer and Raiffa, 1961). A *prior* distribution is the *conjugate prior* distribution for a *likelihood* distribution when the *prior* and the *posterior* distributions are from the same family. The advantage of the use of *conjugate prior* distributions is that the calculation of the *posterior* probability in Eq. (4.11) will have an analytical closed-form (Orloff and Bloom, 2018; Jordan, 2010; Fink, 1997).

Relationships between the Beta distribution and the Binomial distribution

Assuming that $p \in (0, 1)$ is a random variable and is modelled by the Beta distribution, from the Bayesian perspective, the Binomial distribution is:

$$\begin{aligned} p|a, b &\sim \text{Beta}(a, b) \\ s|p &\sim \text{Binomial}(n, p). \end{aligned} \tag{4.12}$$

According to Bayes' theorem, the posterior probability $P(p|s, a, b)$ of variable p given s, a, b is (Murphy, 2012):

$$\begin{aligned} P(p|s, a, b) &\propto P(s|p)P(p|a, b) \\ &\propto (p^s(1-p)^{n-s})(p^{a-1}(1-p)^{b-1}) \\ &\propto p^{(a+s)-1}(1-p)^{(b+n-s)-1}. \end{aligned} \tag{4.13}$$

where $P(s|p)$ and $P(p|a, b)$ are respectively the *likelihood* probability and the *prior* probability.

Comparing Eq. (4.13) and Eq. (4.5), it can be seen that the *posterior* probability $P(p|s, a, b)$ takes the same form of the Beta distribution, written as $\text{Beta}(a + s, b + n - s)$. Since the *prior* probability and the *posterior* probability in Eq. (4.13) are from the same probability family, the Beta distribution is the *conjugate prior* probability distribution for the Binomial distribution.

Fig. 4.1 also shows the relationships among the aforementioned distributions.

4.3. Gaussian Mixture Models (GMMs)

In practical problems, when the samples are generated from different subpopulations that can be described by relatively simple models, this will result in mixture models (Gelman et al., 2013). Gaussian Mixture Models (GMMs) are a kind of mixture model in which the simple model is assumed to be Gaussian distribution.

In statistics, the Gaussian distribution is a commonly used continuous probability distribution, parameterised with mean vector $\boldsymbol{\mu} \in \mathbb{R}^m$ and covariance matrix $\Sigma \in \mathbb{R}^{m \times m}$. In the Gaussian distribution,

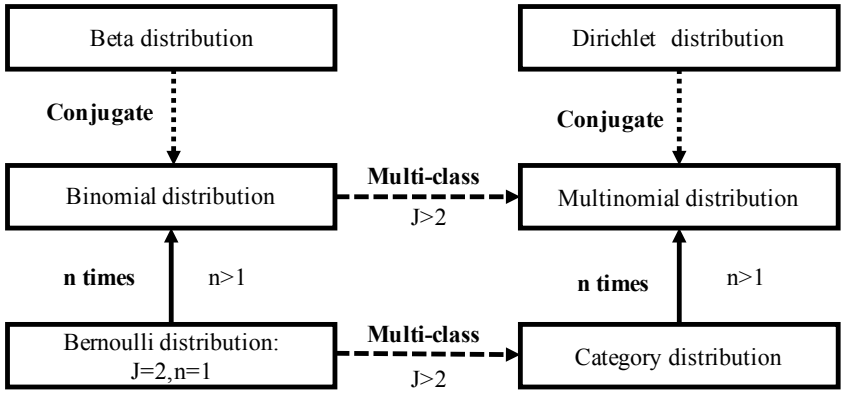


Figure 4.1: Relationships between the Bernoulli, Binomial ,Category, Multinomial, Beta and Dirichlet distributions: J is the number of outcomes in one single experiment, n is the number of experiments, **n times** denotes more than one experiment considered and **Multi-class** denotes more than two outcomes considered (Jung, 2019).

the probability of drawing a random variable $\boldsymbol{x} \in \mathbb{R}^m$ can be obtained by:

$$P(\boldsymbol{x}|\boldsymbol{\mu}, \Sigma) = \frac{1}{(2\pi)^{m/2}\det(\Sigma)^{1/2}}\exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right] \tag{4.14}$$

where det is a determinant operation.

4.3.1. Finite GMMs

Finite mixture models refer to the mixture models with finite mutually exclusive components. A random variable $\boldsymbol{x}$ follows a finite GMM with J components:

$$\begin{aligned} &\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}_j, \Sigma_j) \quad \text{with probability } \pi_j \\ &\text{s.t.} \quad \pi_j > 0 \; \forall j, \quad \sum_{j=1}^{J} \pi_j = 1. \end{aligned} \tag{4.15}$$

where $\boldsymbol{\mu}_j \in \mathbb{R}^m$, $\Sigma_j \in \mathbb{R}^{m\times m}$ are the mean vector and the covariance matrix of the j-th Gaussian component, respectively. The mixing proportion π_j is the probability of drawing from the j-th component.

4.3.2. Infinite GMMs

In contrast to finite GMMs, infinite GMMs explore the GMMs in the limit where $J \to \infty$ (Neal, 1992, Rasmussen, 2000). Typically, the parameters for finite GMMs can be estimated and illustrated component by component. However, when there are an uncountable number of components, the illustration of these components becomes cumbersome, and nearly impossible. In this sub-section, the Dirichlet Process (DP)-GMMs are introduced to cope with the specification of the priori over infinite components. In addition, a method, the Chinese Restaurant Process (CRP), is described for constructing the DP.

Dirichlet Process (DP)-GMMs

Assuming a collection of data samples that can be separated in any finite cluster, in GMMs, each cluster is a Gaussian component. Note that as the number of collected data can be infinitely large, the number of clusters could be countably infinite ($J \to \infty$). $G(\cdot)$ is a discrete distribution assigning the probabilities $\pi_1, \ldots, \pi_J$ of these individual Gaussian components, $(\mu_1, \Sigma_1), \ldots, (\mu_J, \Sigma_J)$. According to Escobar (1988, 1994), when the joint probability $P(\pi_1, \ldots, \pi_J)$ follows the Dirichlet distribution (Forbes et al., 2011), the priori over $G(\cdot)$ can be described with the base function G_0 and the concentration parameter α. The expression of the DP-GMMs can be written as:

$$\begin{aligned} G(\cdot) &\sim \text{DP}(\alpha, G_0) \\ (\boldsymbol{\mu}_j, \Sigma_j) &\sim G(\cdot) \quad \text{for } j = 1, \ldots, J \\ \boldsymbol{\pi} &\sim \text{Dir}(\underbrace{\frac{\alpha}{J}, \ldots, \frac{\alpha}{J}}_{J}) \end{aligned} \tag{4.16}$$

where J is the number of components in GMMs, $\boldsymbol{\pi} = \{\pi_1, \ldots, \pi_J\}$ is a vector of mixing proportions, DP and Dir stand for the Dirichlet Process and the Dirichlet distribution, respectively. To sum up, the DP-GMMs provides a way for defining the prior distribution of $G(\cdot)$ on the settings of mixture Gaussian models, even in the situation that $J \to \infty$. One of the applications of the DP-GMMs is data clustering. The idea is to find a set of Gaussian components with a certain mixing proportion that can mimic the to-be-clustered data. The DP-GMMs clustering is further detailed in Section 4.5 and 4.6.

Chinese Restaurant Process (CRP): a representation of DP

Aldous (1985) introduced an intuitive and efficient way for representing the DP, which is called Chinese Restaurant Process (CRP). Assuming there is a Chinese restaurant which can contain as many tables as possible, the CRP is about how to allocate a new customer to one of these tables. Denote that z_i is the table number selected by the i-th customer. Specifically, in this restaurant, the first customer will always sit on the 1st table, marked as $z_1 = 1$. For the incoming i-th customer, the table selection follows the rule (Lu et al., 2018):

$$P(z_i = j | z_{-i}, \alpha) \begin{cases} \frac{n_j}{N+\alpha-1}, j \in 1, 2, 3 \ldots, J \\ \frac{\alpha}{N+\alpha-1}, j \notin 1, 2, 3 \ldots, J, j = J+1 \end{cases} \tag{4.17}$$

where N is the total number of customers seated in the restaurant. z_{-i} is a collection of the table numbers for the seated customers. n_j is the number of customers already seated in table j. J is the number of occupied tables. The occupied table means $n_j > 0$. As presented in Eq. 4.17, the probability of the i-th customer being assigned to one of the occupied tables is $\frac{n_j}{N+\alpha-1}$, while the probability of being assigned to a new table is $\frac{\alpha}{N+\alpha-1}$.

The construction of the CRP takes $N \to \infty$, $J \to \infty$, however in practice as N is reasonably large, only a finite of components can be observed (Navarro and Perfors, 2014). As the table assignment is a random partition of customers $1, 2, \ldots, N$ to J clusters, the CRP gives a prior distribution on partitions of customers. This is analogous to partitioning data into clusters with the DP-GMMs.

4.4. Hyper-parameters in DP-GMMs

An intuitive interpretation of α and G_0 in the DP given by (Teh, 2010) is that G_0 is the mean of the DP and α is an inverse variance:

$$\begin{aligned} \mathbb{E}[G] &= G_0 \\ \text{var}[G] &= \frac{G_0(1-G_0)}{\alpha+1} \end{aligned} \tag{4.18}$$

where $\mathbb{E}$ and var respectively represent the expectation and variance in probability theory.

4.4.1. Concentration parameter α

From Eq. (4.18), it can be seen that a smaller α results in a larger variance, meaning that $G(\cdot)$ is a sparser PMF, each component in which is more dispersive from one other. Given a set of observations of size n, Escobar and West (1995) states that in practice, a suitable value of α will typically be smaller than n. West (1992) and Escobar and West (1995) proposed that α might be inferred in tandem with Gibbs sampling algorithms (the implementation of Gibbs sampling is discussed in Section 4.5.1). In this thesis, α is set to a constant value 1.

4.4.2. Base function G_0

The base function G_0 is a hyper-parameter on which the Gaussian components from $G(\cdot)$ is centred(Antoniak, 1974). The means and covariance matrices of these Gaussian components are treated as unknown, being modelled by placing prior distributions over them. There are many choices of the base function. For the problem of DP-GMMs we are interested in, the prior knowledge concerning the mean of the Gaussian distribution can be modelled by a Gaussian distribution, and the Inverse Wishart (IW) distribution (Barnard et al., 2000) is used to model the covariance of the Gaussian distribution. The distribution of the mean is dependent on the covariance such that the joint *prior* distribution of the mean and the covariance is conjugate to the *likelihood* distribution (Görür and Rasmussen, 2010).

To proceed the specification of the base function G_0 in the DP-GMMs, the detailed prior knowledge formulation in term of the IW distribution and the Gaussian distribution are introduced. Sequentially, the derivation of the Normal Inverse Wishart (NIW) distribution is given. Furthermore, the NIW distribution is proved that it is the *conjugate prior* of the Gaussian distribution, and can be used as the base function in the DP-GMMs.

Inverse Wishart (IW) distribution

The covariance Σ of the m-dimension Gaussian distribution can be modelled with

$$\Sigma \sim \text{IW}(\nu_0, \Lambda_0) \tag{4.19}$$

where IW is the Inverse Wishart distribution, ν_0 is the number of the degrees of freedom with the restriction that $\nu_0 > m-1$ (Nydick, 2012; Alvarez et al., 2014; Schuurman et al., 2016) for ensuring an invertible Σ, and $\Lambda_0 \in \mathbb{R}^{m\times m}$ is a positive definite matrix. The PDF of Σ following the IW distribution

(Anderson, 2003) is

$$\begin{aligned} P(\Sigma|\nu_0, \Lambda_0) &= \frac{2^{\frac{\nu_0 m}{2}} \Gamma_m(\frac{\nu_0}{2})}{\det(\Sigma)^{\frac{\nu_0+m+1}{2}} |\Lambda_0|^{\nu_0/2}} \exp\left[-\frac{1}{2}\text{tr}(\Lambda_0 \Sigma^{-1})\right] \\ &\propto \det(\Sigma)^{-\frac{\nu_0+m+1}{2}} \exp\left[-\frac{1}{2}\text{tr}(\Lambda_0 \Sigma^{-1})\right] \end{aligned} \tag{4.20}$$

where $\Gamma_m()$ is the m-dimension generalisation of the gamma function and tr() stands for the trace operation. The mean of IW(ν_0, Λ_0) for $\nu_0 > m + 1$ (Alvarez et al., 2014) is

$$\mathbb{E}(\Sigma) = \frac{\Lambda_0}{\nu_0 - m - 1}. \tag{4.21}$$

Gaussian distribution

The prior knowledge for the mean $\boldsymbol{\mu}$ of the Gaussian distribution is a Gaussian distribution linked with Σ:

$$\boldsymbol{\mu}|\Sigma \sim \mathcal{N}(\boldsymbol{u}_0, \frac{\Sigma}{\kappa_0}) \tag{4.22}$$

where $\mathcal{N}$ represents the Gaussian distribution, $\boldsymbol{u}_0$ is the expectation of $\boldsymbol{\mu}$ and κ_0 is a positive value. The PDF of $\boldsymbol{\mu}$ given Eq. (4.22) is:

$$\begin{aligned} P(\boldsymbol{\mu}|\boldsymbol{u}_0, \frac{\Sigma}{\kappa_0}) &= \frac{1}{(2\pi)^{m/2} \det(\frac{\Sigma}{\kappa_0})^{1/2}} \exp\left[-\frac{1}{2}(\boldsymbol{\mu} - \boldsymbol{u}_0)^\top (\frac{\Sigma}{\kappa_0})^{-1} (\boldsymbol{\mu} - \boldsymbol{u}_0\right] \\ &\propto \det(\Sigma)^{-1/2} \exp\left[-\frac{\kappa_0}{2}(\boldsymbol{\mu} - \boldsymbol{u}_0)^\top (\Sigma)^{-1} (\boldsymbol{\mu} - \boldsymbol{u}_0)\right] \\ &\propto \det(\Sigma)^{-1/2} \exp\left[-\frac{\kappa_0}{2}\text{tr}(\Sigma^{-1}(\boldsymbol{\mu} - \boldsymbol{u}_0)(\boldsymbol{\mu} - \boldsymbol{u}_0)^\top)\right]. \end{aligned} \tag{4.23}$$

NIW distribution

Eq. (4.19) and Eq. (4.22) lead to the joint distribution of $\boldsymbol{\mu}$ and Σ being the NIW distribution (Murphy, 2007) parametrised with $\boldsymbol{u}_0, \kappa_0, \nu_0$ and Λ_0:

$$(\boldsymbol{\mu}, \Sigma) \sim \text{NIW}(\boldsymbol{u}_0, \kappa_0, \nu_0, \Lambda_0). \tag{4.24}$$

The PDF of the NIW distribution is the product of Eq. (4.20) and (4.23) :

$$\begin{aligned} P(\boldsymbol{\mu}, \Sigma|\boldsymbol{u}_0, \kappa_0, \nu_0 \Lambda_0) &= P(\Sigma|\nu_0, \Lambda_0) P(\boldsymbol{\mu}|\boldsymbol{u}_0, \frac{\Sigma}{\kappa_0}) \\ &= \det(\Sigma)^{-\frac{\nu_0+m+2}{2}} \exp\left[-\frac{1}{2}\text{tr}(\Lambda_0 \Sigma^{-1})\right] \exp\left[-\frac{\kappa_0}{2}\text{tr}(\Sigma^{-1}(\boldsymbol{\mu} - \boldsymbol{u}_0)(\boldsymbol{\mu} - \boldsymbol{u}_0)^\top)\right] \end{aligned} \tag{4.25}$$

Relationships between the NIW distribution and Gaussian distribution

Let a set of observations $X_n = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\} \in \mathbb{R}^{m \times n}$ be drawn independently and identically from a Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$. Eq. (4.14) gives the *likelihood* probability of observing a random sample from a Gaussian distribution. The *likelihood* probability of observing $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ when Gaussian parameters $\boldsymbol{\mu}$ and Σ are given is the product of the *likelihood* probabilities of individual samples, thus

$P(X_n|\boldsymbol{\mu}, \Sigma)$ is calculated by:

$$P(X_n|\boldsymbol{\mu}, \Sigma) = \prod_{i=1}^{n} P(\boldsymbol{x}_i|\boldsymbol{\mu}, \Sigma). \tag{4.26}$$

According to Eq. (4.14), Eq. (4.26) can be expanded as

$$P(X_n|\boldsymbol{\mu}, \Sigma) = \prod_{i=1}^{n} \frac{1}{(2\pi)^{m/2}\det(\Sigma)^{1/2}} \exp\left[-\frac{1}{2}(\boldsymbol{x}_i - \boldsymbol{\mu})^\top \Sigma^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu})\right]. \tag{4.27}$$

Extracting the common term in the multipliers, then applying the product rule of exponentiation deduces:

$$P(X_n|\boldsymbol{\mu}, \Sigma) = \frac{1}{(2\pi)^{nm/2}\det(\Sigma)^{n/2}} \exp\left\{-\frac{1}{2}\mathrm{tr}\left[\Sigma^{-1}\sum_{i=1}^{n}(\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^\top\right]\right\}, \tag{4.28}$$

in which $\sum_{i=1}^{n}(\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^\top$ can be rewritten as $n(\boldsymbol{\mu} - \overline{X})(\boldsymbol{\mu} - \overline{X})^\top + \sum_{i=1}^{n}(\boldsymbol{x}_i - \overline{X})(\boldsymbol{x}_i - \overline{X})^\top$ (Murphy, 2007) where $\overline{X} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i$. Futhermore, Eq. (4.28) can be expressed as:

$$\begin{aligned} P(X_n|\boldsymbol{\mu}, \Sigma) &= 2\pi^{-\frac{nm}{2}}\det(\Sigma)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2}\mathrm{tr}\left\{\Sigma^{-1}\left[n(\boldsymbol{\mu} - \overline{X})(\boldsymbol{\mu} - \overline{X})^\top + \sum_{i=1}^{n}(\boldsymbol{x}_i - \overline{X})(\boldsymbol{x}_i - \overline{X})^\top\right]\right\}\right\} \\ &\propto \det(\Sigma)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2}\mathrm{tr}\left\{\Sigma^{-1}\left[n(\boldsymbol{\mu} - \overline{X})(\boldsymbol{\mu} - \overline{X})^\top + \sum_{i=1}^{n}(\boldsymbol{x}_i - \overline{X})(\boldsymbol{x}_i - \overline{X})^\top\right]\right\}\right\}. \end{aligned} \tag{4.29}$$

Following Bayes' theorem, the *posterior* probability of the Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$ given observations X_n is

$$P(\boldsymbol{\mu}, \Sigma|X_n) \propto P(X_n|\boldsymbol{\mu}, \Sigma)P(\boldsymbol{\mu}, \Sigma) \tag{4.30}$$

where $P(\boldsymbol{\mu}, \Sigma) = P(\boldsymbol{\mu}, \Sigma|\boldsymbol{u}_0, \kappa_0, \nu_0\Lambda_0)$. In this thesis, both $\boldsymbol{\mu}$ and Σ are unknown, and parameterised by the NIW distribution, the PDF of which is shown in Eq. (4.25). Substitution of Eq. (4.25) and Eq. (4.29) into Eq. (4.30) gives

$$\begin{aligned} P(\boldsymbol{\mu}, \Sigma|X_n) \propto \det(\Sigma)^{-\frac{\nu_0+m+n+2}{2}} &\exp\left[-\frac{1}{2}\mathrm{tr}(\Lambda_0\Sigma^{-1})\right] \exp\left[-\frac{\kappa_0}{2}\mathrm{tr}(\Sigma^{-1}(\boldsymbol{\mu} - \boldsymbol{u}_0)(\boldsymbol{\mu} - \boldsymbol{u}_0)^\top)\right] \\ &\times \exp\left\{-\frac{1}{2}\mathrm{tr}\left\{\Sigma^{-1}\left[n(\boldsymbol{\mu} - \overline{X})(\boldsymbol{\mu} - \overline{X})^\top + \sum_{i=1}^{n}(\boldsymbol{x}_i - \overline{X})(\boldsymbol{x}_i - \overline{X})^\top\right]\right\}\right\}. \end{aligned} \tag{4.31}$$

For Eq. 4.31, integrating terms containing $\boldsymbol{\mu}$ together obtains:

$$\begin{aligned} P(\boldsymbol{\mu}, \Sigma|X_n) \propto \det(\Sigma)^{-\frac{\nu_0+m+n+2}{2}} &\exp\left\{\Sigma^{-1}\left[\Lambda_0 + \sum_{i=1}^{n}(\boldsymbol{x}_i - \overline{X})(\boldsymbol{x}_i - \overline{X})^\top\right)\right]\right\} \\ &\times \exp\left\{-\frac{1}{2}\mathrm{tr}\left\{\Sigma^{-1}\left[\kappa_0(\boldsymbol{\mu} - \boldsymbol{u}_0)(\boldsymbol{\mu} - \boldsymbol{u}_0)^\top + n(\boldsymbol{\mu} - \overline{X})(\boldsymbol{\mu} - \overline{X})^\top\right]\right\}\right\}. \end{aligned} \tag{4.32}$$

The term $\mathrm{tr}\left\{\Sigma^{-1}\left[\kappa_0(\boldsymbol{\mu} - \boldsymbol{u}_0)(\boldsymbol{\mu} - \boldsymbol{u}_0)^\top + n(\boldsymbol{\mu} - \overline{X})(\boldsymbol{\mu} - \overline{X})^\top\right]\right\}$ in Eq. (4.32), marked as $\mathbb{A}$, can be rewritten as follows:

$$\mathbb{A} = \kappa_0(\boldsymbol{\mu} - \boldsymbol{u}_0)^\top\Sigma^{-1}(\boldsymbol{\mu} - \boldsymbol{u}_0) + n(\boldsymbol{\mu} - \overline{X})^\top\Sigma^{-1}(\boldsymbol{\mu} - \overline{X}) \tag{4.33}$$

$$= \kappa_0 \boldsymbol{\mu}^\top \Sigma \boldsymbol{\mu} - \kappa_0 \boldsymbol{u}_0^\top \Sigma \boldsymbol{\mu} - \kappa_0 \boldsymbol{\mu}^\top \Sigma \boldsymbol{u}_0 + \kappa_0 \boldsymbol{u}_0^\top \Sigma \boldsymbol{u}_0 + n\boldsymbol{\mu}^\top \Sigma^{-1} \boldsymbol{\mu} - n\boldsymbol{\mu}^\top \Sigma^{-1} \overline{X} - n\overline{X}^\top \Sigma^{-1} \boldsymbol{\mu} + n\overline{X}^\top \Sigma^{-1} \overline{X} \tag{4.34}$$

$$= (\kappa_0 + n)\boldsymbol{\mu}^\top \Sigma^{-1} \boldsymbol{\mu} - \boldsymbol{\mu}^\top \Sigma^{-1}(\kappa_0 \boldsymbol{u}_0 + n\overline{X}) - (\kappa_0 \boldsymbol{u}_0^\top + n\overline{X}^\top)\Sigma^{-1} \boldsymbol{\mu} + \kappa_0 \boldsymbol{u}_0^\top \Sigma^{-1} \boldsymbol{u}_0 + n\overline{X}^\top \Sigma^{-1} \overline{X} \tag{4.35}$$

Adding terms $\frac{1}{\kappa_0+n}(\kappa_0 \boldsymbol{u}_0 + n\overline{X})^\top \Sigma^{-1}(\kappa_0 \boldsymbol{u}_0 + n\overline{X}) - \frac{1}{\kappa_0+n}(\kappa_0 \boldsymbol{u}_0 + n\overline{X})^\top \Sigma^{-1}(\kappa_0 \boldsymbol{u}_0 + n\overline{X})$ to the right side of Eq. (4.33) gives

$$\begin{aligned}\mathbb{A} = {} & (\kappa_0 + n)\boldsymbol{\mu}^\top \Sigma^{-1} \boldsymbol{\mu} - \boldsymbol{\mu}^\top \Sigma^{-1}(\kappa_0 \boldsymbol{u}_0 + n\overline{X}) - (\kappa_0 \boldsymbol{u}_0^\top + n\overline{X}^\top)\Sigma^{-1} \boldsymbol{\mu} + \kappa_0 \boldsymbol{u}_0^\top \Sigma^{-1} \boldsymbol{u}_0 + n\overline{X}^\top \Sigma^{-1} \overline{X} \\ & + \frac{1}{\kappa_0 + n}(\kappa_0 \boldsymbol{u}_0 + n\overline{X})^\top \Sigma^{-1}(\kappa_0 \boldsymbol{u}_0 + n\overline{X}) - \frac{1}{\kappa_0 + n}(\kappa_0 \boldsymbol{u}_0 + n\overline{X})^\top \Sigma^{-1}(\kappa_0 \boldsymbol{u}_0 + n\overline{X})\end{aligned} \tag{4.36}$$

which can be simplified to

$$\mathbb{A} = (\kappa_0 + n)\text{tr}\left[\Sigma^{-1}\left(\boldsymbol{\mu} - \frac{\kappa_0 \boldsymbol{u}_0 + n\overline{X}}{\kappa_0 + n}\right)\left(\boldsymbol{\mu} - \frac{\kappa_0 \boldsymbol{u}_0 + n\overline{X}}{\kappa_0 + n}\right)^\top\right] + \frac{n\kappa_0}{\kappa_0 + n}\text{tr}\left[\Sigma^{-1}(\boldsymbol{u}_0 - \overline{X})(\boldsymbol{u}_0 - \overline{X})^\top\right]. \tag{4.37}$$

Therefore, the re-formulated Eq. (4.32) is

$$\begin{aligned}P(\boldsymbol{\mu}, \Sigma | X_n) \propto & |\Sigma|^{-(\frac{\nu_0+m+n+2}{2})} \\ & \times \exp\left\{-\frac{1}{2}\text{tr}\left\{\Sigma^{-1}\left[\Lambda_0 + \sum_{i=1}^{n}(\boldsymbol{x}_i - \overline{X})(\boldsymbol{x}_i - \overline{X})^\top + \frac{n\kappa_0}{\kappa_0 + n}(\boldsymbol{u}_0 - \overline{X})(\boldsymbol{u}_0 - \overline{X})^\top\right]\right\}\right\} \\ & \times \exp\left\{-\frac{\kappa_0 + n}{2}\text{tr}\left[\Sigma^{-1}\left(\boldsymbol{\mu} - \frac{\kappa_0 \boldsymbol{u}_0 + n\overline{X}}{\kappa_0 + n}\right)\left(\boldsymbol{\mu} - \frac{\kappa_0 \boldsymbol{u}_0 + n\overline{X}}{\kappa_0 + n}\right)^\top\right]\right\}.\end{aligned} \tag{4.38}$$

Substitution of

$$\begin{aligned}\kappa_1 &= \kappa_0 + n \\ \nu_1 &= \nu_0 + n \\ \boldsymbol{u}_1 &= \frac{\kappa_0 \boldsymbol{u}_0 + n\overline{X}}{\kappa_0 + n} \\ \Lambda_1 &= \Lambda_0 + \sum_{i=1}^{n}(\boldsymbol{x}_i - \overline{X})(\boldsymbol{x}_i - \overline{X})^\top + \frac{n\kappa_0}{\kappa_0 + n}(\boldsymbol{u}_0 - \overline{X})(\boldsymbol{u}_0 - \overline{X})^\top\end{aligned} \tag{4.39}$$

into Eq. (4.38) gives:

$$p(\boldsymbol{\mu}, \Sigma | X_n) \propto |\Sigma|^{-(\frac{\nu_1+m+2}{2})}\exp\{-\frac{1}{2}\text{tr}(\Sigma^{-1}\Lambda_1)\}\exp\{-\frac{\kappa_1}{2}\text{tr}(\Sigma^{-1}(\boldsymbol{\mu} - \boldsymbol{u}_1)(\boldsymbol{\mu} - \boldsymbol{u}_1)^\top))\}. \tag{4.40}$$

It can be seen that the *posterior* PDF in Eq. (4.40) and the *prior* PDF in Eq. (4.25) are from the same probability distribution family. Hence, the NIW distribution is the *conjugate prior* of the Gaussian distribution.

NIW distribution as the base function G_0 in DP-GMMs

From a Bayesian perspective, the base distribution G_0 is the *prior* distribution of $\boldsymbol{\mu}_j$ and Σ_j while $G(\cdot)$ is the *posterior* distribution to be updated using observations. The NIW distribution is selected for G_0, formulated as Eq. 4.41.

$$\begin{aligned} \Sigma_j &\sim IW(\nu_0, \Lambda_0) \\ \boldsymbol{\mu}_j|\Sigma_j &\sim N(\boldsymbol{u}_0, \frac{\Sigma_j}{\kappa_0}) \\ (\boldsymbol{\mu}_j, \Sigma_j) &\sim NIW(\boldsymbol{u}_0, \kappa_0, \nu_0, \Lambda_0) \\ \text{for } j &= 1, 2, \ldots, J. \end{aligned} \tag{4.41}$$

The selection of the NIW distribution as the base function is that due to the conjugacy property, the *posterior* distribution of $G(\cdot)$ can be expressed in close-form. In addition, the updated close-form of $G(\cdot)$ using observations can have the same formulation as Eq. 4.41, but with $\boldsymbol{u}_0, \kappa_0, \nu_0, \Lambda_0$ updated to Eq. 4.40.

4.5. Computation of finite GMMs and DP-GMMs

4.5.1. Review of Gibbs sampling

Markov Chain Monte Carlo (MCMC) methods are a computer-driven sampling method (Green et al., 1966; Gamerman and Lopes, 2006). A pronounced benefit from MCMC is to sidestep the analytic expression of a probability distribution. When it is difficult to summarise the PDF in a closed-form, the desired distribution may be approximated by the samples drawn from it.

Gibbs sampling (Geman and Geman, 1984) is an MCMC technique for drawing samples from multivariate distributions. The idea is that for each variable, samples are generated from its conditional distribution while the other variables are fixed to their current values (Yildirim, 2012). For example, if a random three-dimension sample $\boldsymbol{x} = [x_1, x_2, x_3]$ is assumed to follow a joint distribution $P(x_1, x_2, x_3)$, then the conditional distributions $P(x_1|x_{-1})$, $P(x_2|x_{-2})$ and $P(x_3|x_{-3})$ can be easily obtained and sampled. x_{-1} denotes all variables excluding x_1. At iteration t, samples are $x_1^{(t)} \sim P(x_1|x_2^{(t-1)}, x_3^{(t-1)})$, $x_2^{(t)} \sim P(x_2|x_1^{(t)}, x_3^{(t-1)})$ and $x_3^{(t)} \sim P(x_3|x_1^{(t)}, x_2^{(t)})$.

Algorithm 1 illustrates the procedure of a generic Gibbs sampling for m-dimension data is given. Assisted with distributions $P(x_1|x_{-1})$, $P(x_2|x_{-2})$, $\ldots$, $P(x_m|x_{-m})$, instant sampling from a multivariate distribution $P(x_1, x_2, \ldots, x_m)$ is converted to conduct m samplings from univariate conditional distributions $P(x_1|x_{-1})$, $P(x_2|x_{-2})$, $\ldots$, $P(x_m|x_{-m})$. The evidence of convergence in Gibbs sampling is that the conditional distributions (e.g. $P(x_m|x_{-m})$) remains unchanged. In other words, for each variable, the empirical distribution of the obtained samples is the true marginal distribution (e.g. $P(x_m)$).

However, convergence is difficult to be verified. Gilks et al. (1995) pointed out that the samples generated by MCMC methods have a distribution that approaches the target joint distribution. In addition, the empirical distribution of each variable converges to its true distribution as the number of iterations approaches infinity (Casella and George, 1992; Tierney, 1994). However, it is impractical to let algorithms enter an infinite loop. Usually in practice, a relatively large number of iterations is set to enable the empirical distributions as close as possible to true distributions.

Algorithm 1: Gibbs sampling algorithm for an m-dimension distribution (Gilks et al., 1995)

Initialise $\boldsymbol{x} = [x_1^{(0)}, x_2^{(0)}, \ldots, x_m^{(0)}]$, maximum number of iterations $MaxItn$ and $t = 1$
while $t \leq MaxItn$ **do**
  Sample $x_1^{(t)} \sim P(x_1|x_2^{(t-1)}, x_3^{(t-1)}, \ldots, x_m^{(t-1)})$
  Update current variable values $(x_1^{(t)}, x_2^{(t-1)}, \ldots, x_m^{(t-1)})$
  Sample $x_2^{(t)} \sim P(x_2|x_1^{(t)}, x_3^{(t-1)}, \ldots, x_m^{(t-1)})$
  Update current variable values $(x_1^{(t)}, x_2^{(t)}, \ldots, x_m^{(t-1)})$
  $\vdots$
  Sample $x_{m-1}^{(t)} \sim P(x_{m-1}|x_1^{(t)}, x_2^{(t)}, \ldots, x_{m-2}^{(t)}, x_m^{(t-1)})$
  Update current variable values $(x_1^{(t)}, x_2^{(t)}, \ldots, x_{m-2}^{(t)}, x_{m-1}^{(t)}, x_m^{(t-1)})$
  Sample $x_m^{(t)} \sim P(x_m|x_1^{(t)}, x_2^{(t)}, \ldots, x_{m-2}^{(t)}, x_{m-1}^{(t)})$
  Update current variable values $(x_1^{(t)}, x_2^{(t)}, \ldots, x_{m-2}^{(t)}, x_{m-1}^{(t)}, x_m^{(t)})$
  $t = t + 1$
end

4.5.2. Bayesian inference in finite GMMs

Assuming a set of data $X_N = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N\}$ generated from a GMM model within which each Gaussian component θ_j is parameterised by μ_j and Σ_j, a sample $\boldsymbol{x}_* \in X$ is classified to the j-th component with the probability calculated by Bayes' theorem which is introduced in Section 4.2.3

$$P(\theta_j|\boldsymbol{x}_*) = \frac{P(\boldsymbol{x}_*|\theta_j)P(\theta_j)}{P(\boldsymbol{x}_*)} \tag{4.42}$$

where $P(\theta_j)$ is the *prior* probability of the j-th Gaussian component, $P(\boldsymbol{x}_*|\theta_j)$ is the *likelihood* probability of observing $\boldsymbol{x}_*$ in the j-th component, $P(\theta_j|\boldsymbol{x}_*)$ is the *posterior* probability of the j-th component when $\boldsymbol{x}_*$ is observed. $P(\boldsymbol{x}^*)$ is the *marginal likelihood* probability

$$P(\boldsymbol{x}_*) = \int P(\boldsymbol{x}_*|\theta)P(\theta)d\theta. \tag{4.43}$$

In the case of discrete components, Eq. (4.43) turns into a summation of the form:

$$P(\boldsymbol{x}_*) = \sum_{j=1}^{J} P(\boldsymbol{x}_*|\theta_j)P(\theta_j). \tag{4.44}$$

Substitution of $P(\boldsymbol{x}_*)$ in Eq. (4.42) into Eq. (4.44) gives

$$P(\theta_j|\boldsymbol{x}_*) = \frac{P(\boldsymbol{x}_*|\theta_j)P(\theta_j)}{\sum_{j=1}^{J} P(\boldsymbol{x}_*|\theta_j)P(\theta_j)}. \tag{4.45}$$

Since the denominator $\sum_{j=1}^{J} P(\boldsymbol{x}_*|\theta_j)P(\theta_j)$ in Eq. 4.45 is a constant, the *posterior* probability can be simplified to

$$P(\theta_j|\boldsymbol{x}_*) \propto P(\boldsymbol{x}_*|\theta_j)P(\theta_j) \tag{4.46}$$

Based on the *posterior probabilities* $P(\theta_1|\boldsymbol{x}_*), \ldots, P(\theta_J|\boldsymbol{x}_*)$, the most likely component index is determined by

$$\arg\max_j P(\theta_j|\boldsymbol{x}_*). \tag{4.47}$$

4.5.3. Bayesian inference in DP-GMMs

In the rest of this chapter, the index i always runs over observations. The index j runs over Gaussian components/clusters. A cluster indicator variable $I_i \in \{1, 2, \ldots, J\}$ is introduced for encoding which Gaussian component/cluster $\boldsymbol{x}_i$ belongs to. Instead of the mixing proportions $\boldsymbol{\pi}$, the clustering analysis employs a vector $\boldsymbol{I} = \{I_1, \ldots, I_N\} \in \mathbb{R}^N$ denoting the cluster indicators for each sample in X_N. The objective of the use of DP-GMMs is to find the number of Gaussian components/clusters underlying X_N. To this end, the inference in DP-GMMs is conducted using MCMC methods while relying on the Gibbs sampling for updating the cluster indicator I_i.

Let $\boldsymbol{I}^{(t)} \in \mathbb{R}^N$ be a vector of the updated cluster indicators at iteration t of the Markov chain. According to Theorem 2 given by Escobar (1994), the distribution $P(\boldsymbol{I}|X_N)$ is the stationary distribution of the Markov chain and $\boldsymbol{I}^{(t)}$ converges to the stationary distribution, $P(\boldsymbol{I}|X_N)$, regardless of the initial values of the Markov chain. Therefore, MCMC can be utilised for inference on the number of Gaussian components/clusters underlying X_N. Gibbs sampling can be used for updating the cluster indicators $\boldsymbol{I}^{(t)}$ in turn (Neal, 2000).

In the context of DP-GMMs, the base function G_0 is the NIW distribution parameterised with $\Phi = \{\boldsymbol{u}_0, \kappa_0, \nu_0, \Lambda_0\}$. In the initialisation step, the NIW parameter Φ and a vector of cluster indicators $\boldsymbol{I}$ are assigned, respectively marked with $\Phi^{(0)}$ and $\boldsymbol{I}^{(0)}$. Each sample may be allocated to an individual cluster. Usually, $I_i^{(0)} = i$. This implies that the initial guess of the number of the clusters is equivalent to the number of samples. The initial parameter estimates of the Gaussian component associated with the j-th cluster, $\boldsymbol{\mu}_j^{(0)}$ and $\Sigma_j^{(0)}$, are sampled from NIW($\Phi^{(0)}$) given $\boldsymbol{x}_i$.

Over the procedure of Markov chain, the number of unique values in $\boldsymbol{I}^{(t)}$ will reduce to typically much fewer than the number of samples due to the clustering effect. Specifically, only a few clusters will contain samples while the others become empty.

Görür and Rasmussen (2010) summarised the Gibbs updates for $I_i^{(t)}$ through all the X_N as (the update of α is omitted since α is set constant):

– Update the identity vector $\boldsymbol{I}^{(t)}$. It is equivalent to clustering the samples. According to Algorithm 2 given by Neal (1992), each sample $\boldsymbol{x}_*$ is assigned to a cluster with respect to the current clustering result of all other samples, i.e. each identity $I_*^{(t)}$ is sampled given the existing clusters. If the j-th cluster contains additional samples except for x_*, the probability of a sample x_* being assigned to this cluster is

$$\begin{aligned} &p(I_* = j|\boldsymbol{I}_{-*}^{(t)}, \Phi = \{\boldsymbol{u}_0, \kappa_0, \nu_0, \Lambda_0\}, \boldsymbol{\mu}_1^{(t)}, \Sigma_1^{(t)}, \ldots, \boldsymbol{\mu}_J^{(t)}, \Sigma_J^{(t)}, \alpha) \\ &\qquad = \frac{n_j p(\boldsymbol{x}_*|\boldsymbol{\mu}_j^{(t)}, \Sigma_j^{(t)})}{\alpha G_0(\boldsymbol{x}_*) + \sum_{j=1}^{J} n_j p(\boldsymbol{x}_*|\boldsymbol{\mu}_j^{(t)}, \Sigma_j^{(t)})} \end{aligned} \tag{4.48}$$

where for the j-th cluster, $\boldsymbol{\mu}_j^{(t)}$ and $\Sigma_j^{(t)}$ are the mean vector and covariance matrix of the Gaussian distribution associated with the j-th cluster at iteration t and n_j is the number of samples in the

j-th cluster. $G_0(\boldsymbol{x}_*)$ is the marginal likelihood of $\boldsymbol{x}_*$ evaluated at a Gaussian distribution whose prior information is modelled using a NIW distribution:

$$G_0(\boldsymbol{x}_*) = \int P(x_*|\boldsymbol{\mu}, \Sigma) P(\boldsymbol{\mu}, \Sigma|\boldsymbol{u}_0, \kappa_0, \nu_0, \Lambda_0) d\boldsymbol{\mu} d\Sigma. \tag{4.49}$$

It should be noted that the clusters without samples in are treated equally. The probability of x_* being assigned to any empty cluster is

$$\begin{aligned} p(I_* \in j_\varnothing \}|\boldsymbol{I}_{-*}^{(t)}, \Phi = \{\boldsymbol{u}_0, \kappa_0, \nu_0, \Lambda_0\}, \boldsymbol{\mu}_1^{(t)}, \Sigma_1^{(t)}, \ldots, \boldsymbol{\mu}_J^{(t)}, \Sigma_J^{(t)}, \alpha) \\ = \frac{\alpha G_0(\boldsymbol{x}_*)}{\alpha G_0(\boldsymbol{x}_*) + \sum_{j=1}^{J} n_j p(\boldsymbol{x}_*|\boldsymbol{\mu}_j^{(t)}, \Sigma_j^{(t)})} \end{aligned} \tag{4.50}$$

where $j_\varnothing$ denotes the index of an empty cluster.

– Update the parameters $\Phi_j^{(t)}$ for $j \notin j_\varnothing$. $\Phi_j^{(t)}$ is the parameter of the *posterior* NIW distribution associated with the j-th cluster at iteration t. Let $X^{(j)} = \{\boldsymbol{x}_1^{(j)}, \boldsymbol{x}_2^{(j)}, \ldots, \boldsymbol{x}_{n_j}^{(j)}\}$ be a set of samples allocated to the j-th cluster. $\Phi_j^{(t)} = \{\boldsymbol{u}^{(j)}, \kappa^{(j)}, \nu^{(j)}, \Lambda^{(j)}\}$ is updated using

$$\begin{aligned} \boldsymbol{u}^{(j)} &= \frac{\kappa_0}{\kappa_0 + n_j}\boldsymbol{u}_0 + \frac{n_j}{\kappa_0 + n_j}\bar{X}_j \\ \kappa^{(j)} &= \kappa_0 + n_j \\ \nu^{(j)} &= \nu_0 + n_j \\ \Lambda^{(j)} &= \Lambda_0 + \sum_{i=1}^{n_j} (\boldsymbol{x}_i^{(j)} - \bar{X}_j)(\boldsymbol{x}_i^{(j)} - \bar{X}_j)^T + \\ &\quad \frac{\kappa_0 n_j}{\kappa_0 + n_j}(\bar{X}_j - \boldsymbol{u}_0)(\bar{X}_j - \boldsymbol{u}_0)^T \\ \text{where} \quad \bar{X}_j &= \frac{1}{n_j}\sum_{i=1}^{n_j} \boldsymbol{x}_i^{(j)}. \end{aligned} \tag{4.51}$$

– Update the parameter $\boldsymbol{\mu}_j^{(t)}, \Sigma_j^{(t)}$ of non-empty clusters. The posterior distribution of $\boldsymbol{\mu}_j^{(t)}$ and $\Sigma_j^{(t)}$ is derived based on Bayes' theorem:

$$\begin{aligned} p(\boldsymbol{\mu}_j^{(t)}, \Sigma_j^{(t)}|X^{(j)}, \Phi_0) &\propto p(X^{(j)}|\boldsymbol{\mu}_j^{(t)}, \Sigma_j^{(t)}) p(\boldsymbol{\mu}_j^{(t)}, \Sigma_j^{(t)}|\Phi_0) \\ &= p(\boldsymbol{\mu}_j^{(t)}, \Sigma_j^{(t)}|\Phi_j^{(t)}) \end{aligned} \tag{4.52}$$

Therefore, it can be implemented by sampling $\boldsymbol{\mu}_j^{(t)}$ and $\Sigma_j^{(t)}$ from the updated posterior NIW distribution NIW($\Phi_j^{(t)}$).

4.6. An investigation into the influence of parameters on the accuracy of DP-GMMs clustering

There are four hyper parameters in the DP-GMMs. Inappropriate assignments of these hyper parameters might cause a failure in partitioning data as their natural clusters. This section investigates how

the hyper parameters influence the clustering results. To clearly demonstrate these influences, simulation data are used for illustration. Section 4.6.1 gives the expression of a four-mode simulation model. Section 4.6.2 discusses the Jeffrey's parameter initialisation and gives the commonly used parameter settings. Section 4.6.3 presents the clustering results of data generated from the four-mode simulation model, with the commonly used parameter settings. In addition, Section 4.6.3 evaluated the estimation of Gaussian parameters associated to each cluster. The estimation is based on the DP-GMMs. Furthermore, Section 4.6.4 analyses how to improve the clustering results and the DP-GMMs-based Gaussian parameter estimation. The analysis also incorporates a new mode into the 4-mode simulation model for highlighting the improvements using the proposed hyper parameter settings.

4.6.1. Simulation model

The performance of clustering data using the DP-GMMs approach is validated on a bivariate simulated model which is comprised of four modes. This simulated model was previously described by Tan et al. (2019, 2020):

Mode 1:

$$\begin{aligned} x_1 &= e_{21} \\ x_2 &= 1.5x_1 + e_{22} \end{aligned} \tag{4.53}$$

where $e_{21} \sim \mathcal{N}(0, 1)$ and $e_{22} \sim \mathcal{N}(0, 9)$.

Mode 2:

$$\begin{aligned} x_1 &= e_{11} + 8 \\ x_2 &= -0.2x_1 + 5 + e_{12} \end{aligned} \tag{4.54}$$

where $e_{11} \sim \mathcal{N}(0, 2.25)$ and $e_{12} \sim \mathcal{N}(0, 0.25)$.

Mode 3:

$$\begin{aligned} x_1 &= e_{41} + 15 \\ x_2 &= -x_1 + 20 + e_{42} \end{aligned} \tag{4.55}$$

where $e_{41} \sim \mathcal{N}(0, 0.25)$ and $e_{42} \sim \mathcal{N}(0, 0.09)$.

Mode 4:

$$\begin{aligned} x_1 &= e_{31} + 9 \\ x_2 &= \frac{1}{3}x_1 - 4 + e_{32} \end{aligned} \tag{4.56}$$

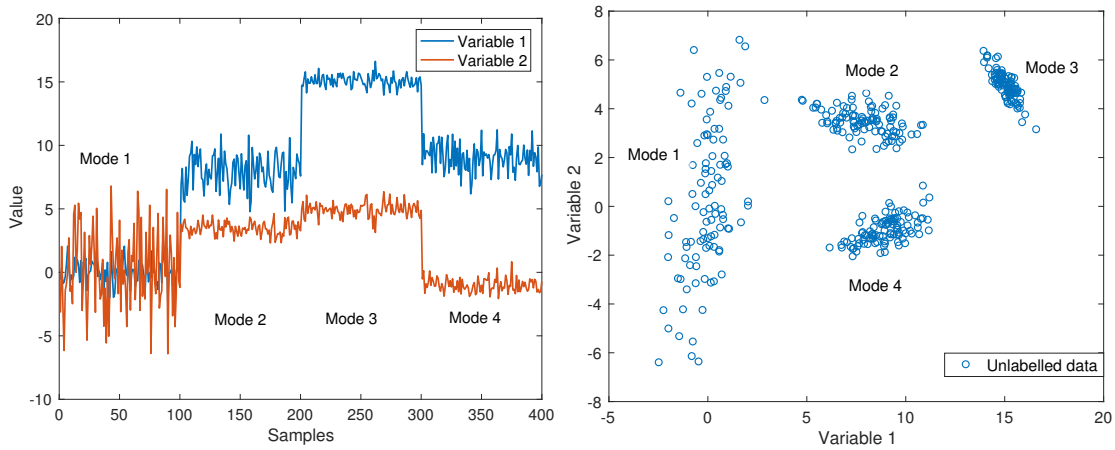
where $e_{31} \sim \mathcal{N}(0, 1)$ and $e_{32} \sim \mathcal{N}(0, 0.25)$.

Each mode generates 100 samples. Fig. 4.2(a) shows the trend plot of x_1 and x_2. The corresponding scatter plot is presented in Fig. 4.2(b).

4.6.2. Parameter initialisation for the NIW distribution

Similar to the introduction to the base function in the Section 4.4.2, the discussion of the parameter initialisation in the NIW distribution develops based on a single Gaussian component $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$. $X_n = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ are samples drawn independently and identically from $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$. The *prior* and *posterior* NIW distribution are NIW$(\boldsymbol{u}_0, \kappa_0, \nu_0, \Lambda_0)$ and NIW$(\boldsymbol{u}_1, \kappa_1, \nu_1, \Lambda_1)$, respectively.

As in Bayesian inference, parameters $\boldsymbol{\mu}$ and Σ are sampled from the *posterior* NIW$(\boldsymbol{u}_1, \kappa_1, \nu_1, \Lambda_1)$, the most often appearing values of $\boldsymbol{\mu}$ and Σ are of interest to be known. These values can be calculated



(a) Trend plot of the simulated data

(b) Scatter plot of the simulated data

Figure 4.2: Illustration of data generated from a bivariate simulated model (Tan et al., 2019): sub-figure (a) gives the trend plot of two variables, showing that there are four distinct modes with different levels for variable x_1 and x_2. The relative positions between samples are demonstrated in Sub-figure (b) in the form of scatter plot. It is apparent that the data consists of four clusters.

by (Fraley and Raftery, 2007):

$$\begin{aligned} \boldsymbol{\mu}_{\text{sampled}} &= \boldsymbol{u}_1 = \frac{\kappa_0 \boldsymbol{u}_0 + n\overline{X}}{\kappa_0 + n} = \frac{\kappa_0}{\kappa_0 + n}\boldsymbol{u}_0 + \frac{n}{\kappa_0 + n}\overline{X} \\ \Sigma_{\text{sampled}} &= \frac{\Lambda_1}{\nu_1 + m + 2} = \frac{\Lambda_0 + \sum_{i=1}^{n}(\boldsymbol{x}_i - \overline{X})(\boldsymbol{x}_i - \overline{X})^\top + \frac{n\kappa_0}{\kappa_0+n}(\boldsymbol{u}_0 - \overline{X})(\boldsymbol{u}_0 - \overline{X})^\top}{\nu_0 + n + m + 2} \end{aligned} \tag{4.57}$$

where $\overline{X}$ is the sample mean of X_n, and $\boldsymbol{\mu}_{\text{sampled}}$ and Σ_{sampled} respectively denote the most often sampled $\boldsymbol{\mu}$ and Σ.

Jeffrey's parameter initialisation

The base function G_0 expresses an centroid around which the component parameters should be centred. In the DP-GMMs where G_0 is the NIW distribution, we expect that a draw $(\boldsymbol{\mu}, \Sigma)$ from $\text{NIW}(\boldsymbol{u}_1, \kappa_1, \nu_1, \Lambda_1)$ will reflect the sample mean and sample covariance of $X_n = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\}$. A non-informative assignment to $\boldsymbol{u}_0, \kappa_0, \nu_0, \Lambda_0$ is selected so as to have a negligible influence on the estimated parameters (Schuurman et al., 2016). Such an assignment given by Jeffrey (Gelman et al., 2013) is $\kappa_0 \to 0$, $\nu_0 \to -1$ and $\det(\Lambda_0) \to 0$. Then, the parameters in the *posterior* NIW distribution become

$$\boldsymbol{u}_{1,\text{Jeff}} = \bar{X} \tag{4.58}$$

$$\kappa_{1,\text{Jeff}} = n \tag{4.59}$$

$$\nu_{1,\text{Jeff}} = n - 1 \tag{4.60}$$

$$\Lambda_{1,\text{Jeff}} = \sum_{i=1}^{n}(\boldsymbol{x}_i - \bar{X})(\boldsymbol{x}_i - \bar{X})^\top. \tag{4.61}$$

where $\boldsymbol{u}_{1,\text{Jeff}}$, $\kappa_{1,\text{Jeff}}$, $\nu_{1,\text{Jeff}}$ and $\Lambda_{1,\text{Jeff}}$ denote the parameters of the posterior NIW distribution given Jeffrey's assignments.

However, the Jeffrey's parameter initialisation may be problematic in the application of Bayesian inference. According to Murphy (2007), the marginal likelihood of measurements X_n evaluated at a Gaussian distribution parameterised by NIW$(\boldsymbol{u}_0, \kappa_0, \nu_0, \Lambda_0)$ can be calculated by:

$$G_0(X_n) = \frac{1}{\pi^{nm/2}} \frac{\Gamma_m(\nu_n/2)}{\Gamma_m(\nu_0/2)} \frac{\det(\Lambda_0)^{\nu_0/2}}{\det(\Lambda_n)^{\nu_n/2}} (\frac{\kappa_0}{\kappa_n})^{m/2}. \tag{4.62}$$

As mentioned in Section 4.5.3, $G_0(X_n)$ in clustering analysis is the probability of observations assigned to an empty cluster. Given either $\kappa_0 = 0$ or $\det(\Lambda_0) = 0$, $G_0(X_n)$ will be a constant value equal to zero. This means that there is no change of sampling a cluster difference from current clustering result. In addition, even if not taking the extreme values of Jeffrey's parameter initialisation, it is difficult to determine how small these values should be to guarantee the results in Eq. (4.58)-(4.61).

Default parameter settings

A default setting for Λ_0 and ν_0 is $\Lambda_0 = \boldsymbol{I} \in \Re^{m \times m}$ and $\nu_0 = m + 1$ where $\boldsymbol{I}$ is an identity matrix. (Alvarez et al., 2014). The term $\sum_{i=1}^{n} (\boldsymbol{x}_i - \overline{X})(\boldsymbol{x}_i - \overline{X})^\top$ in Eq. (4.57) is the sample covariance, which can be denoted as Σ_{sc}. $\boldsymbol{u}_0 - \overline{X}$ is a measure on the error between the guess of the $\boldsymbol{u}_0$ and the sample mean $\overline{X}$. Therefore, $(\boldsymbol{u}_0 - \overline{X})(\boldsymbol{u}_0 - \overline{X})^\top$ is the residual sum of squares, denoted using RSS. Then Eq. (4.57) can be rewritten as follows:

$$\begin{aligned} \boldsymbol{\mu}_{\text{sampled}} &= \frac{\kappa_0}{\kappa_0 + n} \boldsymbol{u}_0 + \frac{n}{\kappa_0 + n} \overline{X} \\ \Sigma_{\text{sampled}} &= \frac{\Lambda_0}{\nu_0 + n + m + 2} + \frac{n-1}{\nu_0 + n + m + 2} \Sigma_{sc} + \frac{n\kappa_0(\nu_0 + n + m + 2)}{\kappa_0 + n} RSS. \end{aligned} \tag{4.63}$$

According to Eq. (4.63), $\boldsymbol{\mu}_{\text{sampled}}$ can be interpreted as the weighted summation of $\boldsymbol{u}_0$ and $\overline{X}$. In addition, from Bayesian perspective, $\boldsymbol{\mu}$ following the Gaussian distribution in Eq. (4.22) can be written in a conditional form

$$P(\boldsymbol{\mu}|\Sigma) = \mathcal{N}(\boldsymbol{\mu}|\boldsymbol{u}_0, \frac{\Sigma}{\kappa_0}) \tag{4.64}$$

where $\boldsymbol{\mu}$ can be interpreted as κ_0 observations of $\boldsymbol{u}_0$ on Σ (Murphy, 2007). Given observations X, the *posterior* $\boldsymbol{\mu}$ written in conditional form is

$$P(\boldsymbol{\mu}|X_n) = \mathcal{N}(\boldsymbol{\mu}|\boldsymbol{u}_1, \frac{\Sigma}{\kappa_1}). \tag{4.65}$$

Recalling $\kappa_1 = \kappa_0 + n$, κ_0 is equivalent to the prior sample size, playing a role analogous to n. To give a non-informative *prior* distribution, $\kappa_0 = 0$ (Murphy, 2007). However, to make the fraction $\frac{\Sigma}{\kappa_0}$ have meaning, κ_0 should not be of value zero, in the meanwhile to represent the prior sample size of less informative, $\kappa_0 = 1$. To make $\boldsymbol{\mu}_{\text{sampled}}$ equal to the sample mean, $\boldsymbol{u}_0 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$. The inaccuracy in the mode of $\boldsymbol{\mu}$ caused by $\kappa_0 = 1$ may be negligible and eliminated as $n \to \infty$.

Table 4.1: Comparison of parameter estimations of Gaussian distributions: the sample means and sample covariances are used as the true Gaussian parameters associated to each cluster. The estimates of Gaussian parameters are derived using the DP-GMMs algorithm with the frequently used parameter settings. The comparison shows that the sample means and estimated means are very close to each other, while there are significant differences between sample covariances and estimated covariances, particularly in the covariances of cluster 3.

Cluster index	Estimation based on samples		Estimation with $\boldsymbol{u}_0 = \begin{bmatrix}0\\0\end{bmatrix}, \kappa_0 = 1, \nu_0 = 3, \Lambda_0 = \boldsymbol{I}$	
	Mean	Covariance	Mean	Covariance
1	$\begin{bmatrix}0.01\\0.28\end{bmatrix}$	$\begin{bmatrix}0.99; 1.57\\1.57; 9.74\end{bmatrix}$	$\begin{bmatrix}0.01\\0.27\end{bmatrix}$	$\begin{bmatrix}0.92; 1.45\\1.45; 9.02\end{bmatrix}$
2	$\begin{bmatrix}7.95\\3.45\end{bmatrix}$	$\begin{bmatrix}1.87; -0.30\\-0.30; 0.27\end{bmatrix}$	$\begin{bmatrix}7.87\\3.42\end{bmatrix}$	$\begin{bmatrix}2.32; -0.02\\-0.02; 0.37\end{bmatrix}$
3	$\begin{bmatrix}15.09\\4.95\end{bmatrix}$	$\begin{bmatrix}0.23; -0.22\\-0.22; 0.32\end{bmatrix}$	$\begin{bmatrix}14.94\\4.90\end{bmatrix}$	$\begin{bmatrix}2.33; 0.48\\0.48; 0.53\end{bmatrix}$
4	$\begin{bmatrix}8.99\\-1.00\end{bmatrix}$	$\begin{bmatrix}1.22; 0.35\\0.35; 0.29\end{bmatrix}$	$\begin{bmatrix}8.90\\-0.99\end{bmatrix}$	$\begin{bmatrix}1.89; 0.24\\0.24; 0.29\end{bmatrix}$

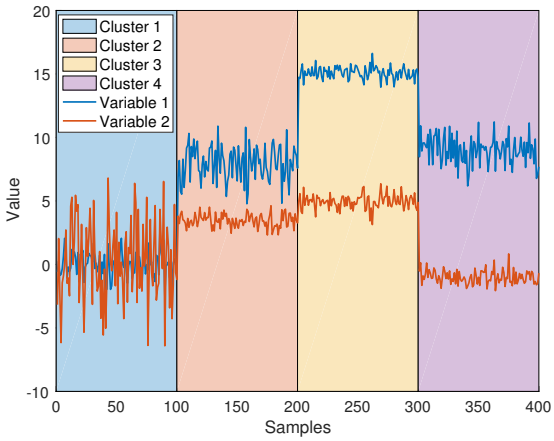
4.6.3. Clustering results and parameter estimation for Gaussian distributions

This section conducts the clustering results using the DP-GMMs method, also evaluates the parameter estimation of Gaussian components associated to each cluster. The DP-GMMs is implemented in a Gibbs sampling manner. In this implementation, the initial assignment to $\Phi^{(0)}$ is $\boldsymbol{u}_0 = \boldsymbol{0}, \kappa_0 = 1, \nu_0 = 3, \Lambda_0 = \begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}$, which is the default parameter settings (see Section 4.6.2). Applied with the Gibbs sampling in the framework of DP-GMMs, the simulated data are clustered according to their modes. The clustering results are shown in Fig. 4.3(a) and (b). It can be seen that without specifying the number of modes, DP-GMMs-based clustering algorithm is able to successfully partition the unlabelled data.

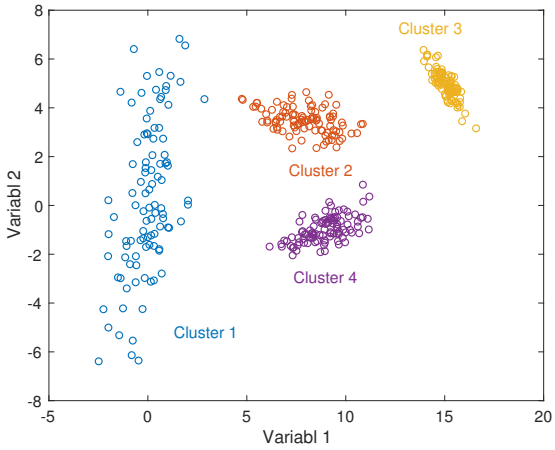
Table 4.1 shows the parameter estimations of the Gaussian components using the samples themselves and the given parameter settings. Taking the sample means and sample covariances as the true parameters of the obtained four Gaussian distributions, it can be seen that the estimated means using the given parameter settings are accurate. However, the estimated covariance matrices, particularly for cluster 3, are different from the sample covariance matrices. The differences between them are further illustrated in Fig. 4.3(c) and (d). For cluster 2, the ellipse contour against 95% confidence level in Fig. 4.3(d) is relaxed relative to the one in Fig. 4.3(c). For cluster 3, the directions of the long-axes of the ellipses in Fig. 4.3(c) and (d) are opposite.

This is because for $n \to \infty$ and $\nu_0 \ll n$, $\boldsymbol{\mu}_{\text{sampled}}$ and Σ_{sampled} have the following results

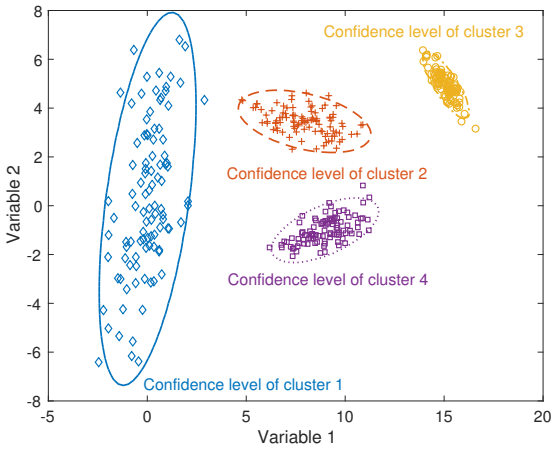
$$\begin{aligned}\lim_{n\to\infty} \boldsymbol{\mu}_{\text{sampled}} &= \overline{X}\\ \lim_{\nu_0 \ll n, n\to\infty} \Sigma_{\text{sampled}} &= \Sigma_{sc} + \kappa_0 nRSS.\end{aligned} \tag{4.66}$$



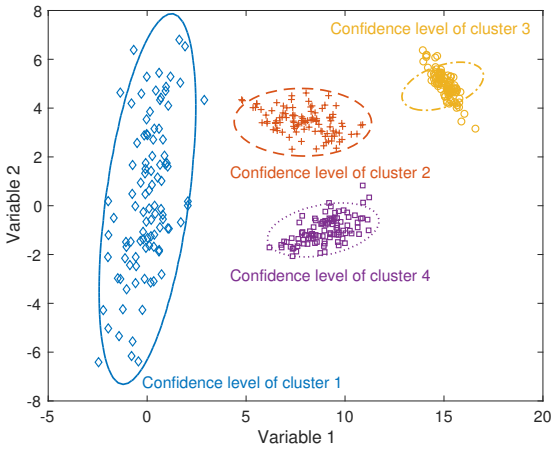
(a) Trend plot of the clustered data



(b) Scatter plot of the clustered data



(c) Plot of covariance ellipses based on the sample mean and sample covariance. The calculation of sample mean and sample covariance are using Eq. 4.67.



(d) Plot of covariance ellipses based on the estimation of DP-GMMs. The hyper parameters in DP-GMMs are initialised with $\boldsymbol{u}_0 = \mathbf{0}, \kappa_0 = 1, \nu_0 = 3, \Lambda_0 = \left[\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right]$.

Figure 4.3: Plots of clustered data based on the DP-GMMs algorithm and illustrations of covariance estimation of the Gaussian distributions associated to each cluster: the simulation data are automatically separated into four clusters using the DP-GMMs without specifying the number of clusters. These clusters are consistent with the nature of the data. The clustering results are presented in both trend and scatter plots as shown in sub-figures (a) and (b), respectively. To illustrate the covariance estimation, the ellipse contours at 95% confidence level are used. In sub-figure (c), the covariance ellipse of each cluster is based on the sample mean and sample covariance, while in sub-figure (d), the means and covariances are estimated using DP-GMMs algorithm, the initial parameters of which are $\boldsymbol{u}_0 = \mathbf{0}, \kappa_0 = 1, \nu_0 = 3, \Lambda_0 = \left[\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right]$. Comparing the covariance ellipses in (c) and (d), it can be observed that the DP-GMMs-based covariance estimation is biased. This estimation error is most apparent in cluster 3.

Because the mean of cluster 3 is far from the assigned $\boldsymbol{u}_0 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$, the RSS is large, leading to the estimate of a covariance with large error. Therefore, it is recommended to re-estimate the mean and covariance matrix of a Gaussian component by the sample mean and sample covariance of the cluster that is associated to each Gaussian component (Chang et al., 2018), as shown in Eq. (4.67).

$$\begin{aligned} \hat{\boldsymbol{\mu}}_j &= \bar{X}_j \\ \hat{\Sigma}_j &= \frac{1}{n_j - 1} \sum_{i=1}^{n_j} (\boldsymbol{x}_i^{(j)} - \bar{X}_j)(\boldsymbol{x}_i^{(j)} - \bar{X}_j)^\top. \end{aligned} \tag{4.67}$$

4.6.4. Improving clustering performance: data normalisation and determination of κ_0

The reason of having errors in covariance estimation is discussed and analysed in Section 4.6.3. Although the covariance estimated though the DP-GMMs can be replaced by sample covariances, the covariance error will lead to mis-clustering. An additional mode as follows is added to the simulated multimode model (see Section 4.6.1),

Mode 5:

$$\begin{aligned} x_1 &= e_{51} + 13 \\ x_2 &= \frac{8}{25} x_1 + e_{52} \end{aligned} \tag{4.68}$$

where $e_{51} \sim \mathcal{N}(0, 0.65)$ and $e_{52} \sim \mathcal{N}(0, 0.25)$. 100 samples are generated from this mode.

The trend plot and scatter plot of the data from 5 modes are illustrated in Fig. 4.4(a) and (b). Fig. 4.4(b) shows that in two-dimension panel, there is a close location relationship between modes 2, 3 and 5. If the covariance estimation of one of them is inaccurate, and with large error, the clustering results may be influenced. For example, due to the direction error in Fig. 4.3(d), some samples from mode 5 will be mis-classified to mode 3, and finally, this error will cause the failure in clustering mode 5 as shown in Fig. 4.4(c). It can be seen that mode 3 and 5 are treated as one cluster. In addition, all the covariance ellipses calculated by the given parameter initialisation are more relaxed than the ellipses in Fig. 4.3(c) calculated by the samples. These inaccuracies are caused by the RSS term in Eq. (4.66)

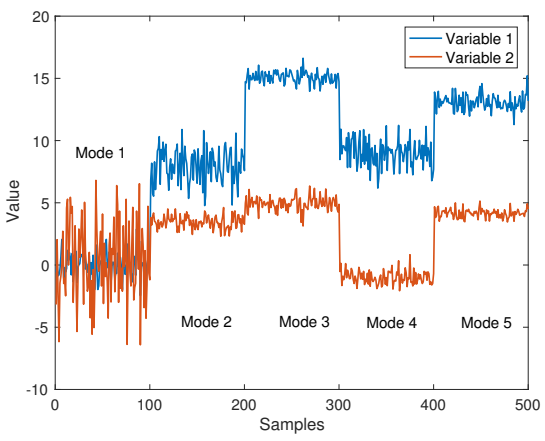
To narrow the error in RSS, it is proposed to normalise the unlabelled data with each variable centred to zero and with variance of one. An advantage of this normalisation way is that it can remain the probability distribution in the original data. The sample covariance ellipses of normalised data are plotted in Fig. 4.5(a), taken as the true covariances of these modes. Moreover, to avoid the RSS to be amplified by $\kappa_0 n$, an ideal κ_0 should be an extremely small value. The problem is how small κ_0 should be. In this thesis, the value of κ_0 is determined by

$$\kappa_0 = \frac{1}{N} \tag{4.69}$$

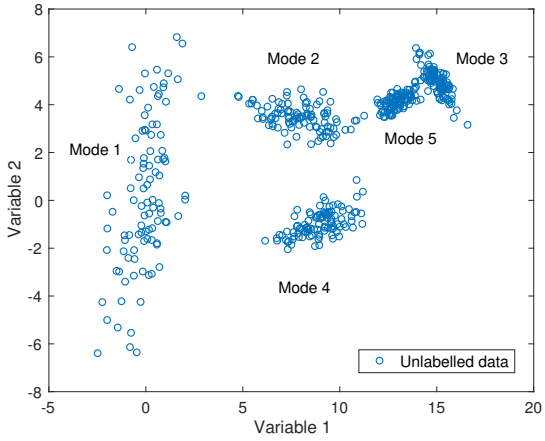
where N is the number of data to be clustered. Sequentially, given Eq. (4.69), Eq. (4.66) turns into

$$\begin{aligned} \lim_{n \to \infty} \boldsymbol{\mu}_{\text{sampled}} &= \overline{X} \\ \lim_{\nu_0 \ll n, n \to \infty} \Sigma_{\text{sampled}} &= \Sigma_{sc} + \frac{n}{N} RSS. \end{aligned} \tag{4.70}$$

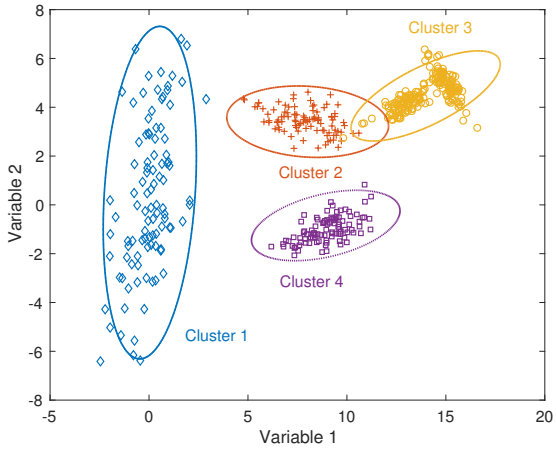
Since the number of samples in certain cluster should be less than N, the error caused by the RSS is diminished.



(a) Trend plot of simulation data



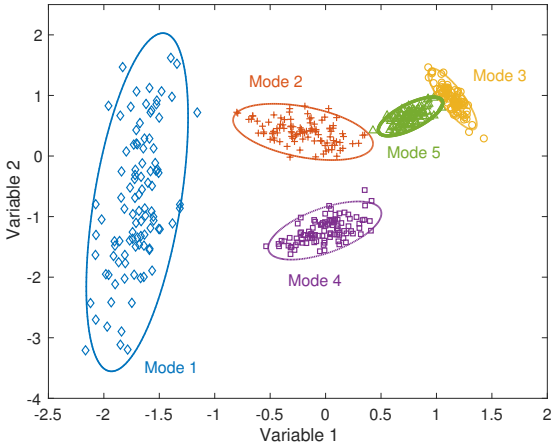
(b) Scatter plot of simulation data



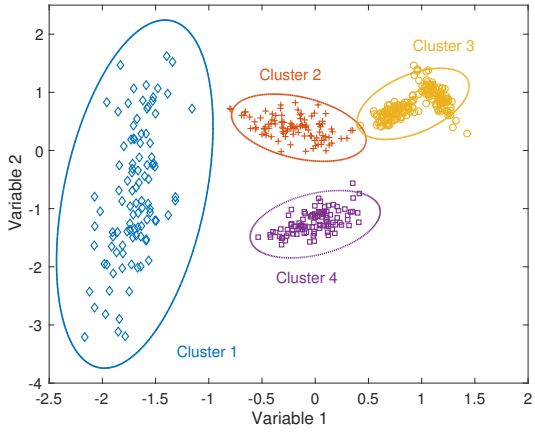
(c) Scatter plot of clustered simulation data, and the covariance ellipses associated to each cluster

Figure 4.4: Data clustering using the DP-GMMs algorithm. The parameter settings of the DP-GMMs are $\boldsymbol{u}_0 = \mathbf{0}, \kappa_0 = 1, \nu_0 = 3, \Lambda_0 = \left[\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right]$. The data to be clustered are generated from a 5-mode model. The expression of this model is given in Section 4.6.1 and 4.6.4. The unlabelled data are presented in sub-figures (a) and (b) in both trend and scatter plots. Sub-figure (c) highlights the labelled data, which does not match the nature of these data. The samples from mode 3 and 5 are mistakenly grouped together.

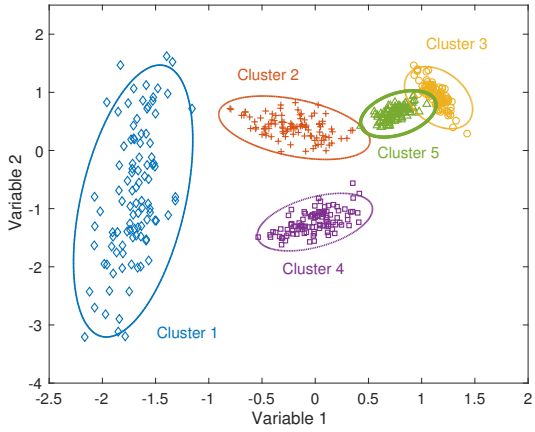
Fig. 4.5(b) demonstrates the clustering results of normalised data using parameter settings $\boldsymbol{u}_0 = \mathbf{0}, \kappa_0 = 1, \nu_0 = 3, \Lambda_0 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$, while Fig. 4.5(c) shows the clustered data and covariance plots when $\boldsymbol{u}_0 = \mathbf{0}, \kappa_0 = \frac{1}{500}, \nu_0 = 3, \Lambda_0 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$. Comparing the results in Fig. 4.5 (b) and (c), for cluster 3, the inaccuracies in the direction of the ellipse are significantly reduced. Moreover, the covariance ellipses of cluster 1, 2 and 4 calculated under the condition of $\kappa_0 = \frac{1}{500}$ are tighter than the ellipses calculated when $\kappa_0 = 1$. Most advantageous, mode 5 is successfully partitioned from mode 4. However, covariance ellipses are more relaxed relative to the true covariances in Fig. (4.5)(a). To guarantee the Gaussian statistics related to each mode are accurate, it is suggested to used the mean and covariance of samples.



(a) Plot of covariance ellipse for the samples of each mode: data are normalised to have zero mean and variable variance of one. The covariance is calculated based on samples of each mode. These covariances are treated as the ground truth.



(b) Data clustering based on the DP-GMMs and the plot of covariance ellipses. Parameter settings for clustering and covariance estimation: $\boldsymbol{u}_0 = \mathbf{0}, \kappa_0 = 1, \nu_0 = 3, \Lambda_0 = \left[\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right]$.



(c) Data clustering based on the DP-GMMs and the plot of covariance ellipses. Parameter settings for clustering and covariance estimation: $\boldsymbol{u}_0 = \mathbf{0}, \kappa_0 = \frac{1}{500}, \nu_0 = 3, \Lambda_0 = \left[\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right]$.

Figure 4.5: Clustering results comparison: the unlabelled data are normalised with zero mean and variable variance of one. Sub-figure (a) gives the true labels and covariance ellipses. Given different parameter settings, the DP-GMMs algorithm leads to distinct clustering results and covariance estimation. Sub-figure (b) shows that with inappropriate parameter settings, mode 3 and 5 still are identified as one cluster. However, using the proposed assignment of κ_0, the clustering results in sub-figure (c) is similar to the truth presented in sub-figure (a), having 5 clusters.

4.7. The application of DP-GMMs in a monitoring framework

Varying production demand and loading conditions on equipment often result in multiple operating modes in process operation. Typically, multimodality exists in the data recorded from such processes, and poses challenges to building monitoring models which may be trained from labelled historical data. An anomaly detected using such monitoring models might be either a symptomatic of a developing fault, or indicates the emergence of a new operating mode. Particularly, the appearance of a new operating mode means that the trained monitoring models are not adequate to account for the normal operation.

This situation necessitates the incorporation of the new healthy data into the monitoring model, thus minimising future false and missed alarms.

A key step in the monitoring framework proposed in joint work Tan et al. (2019, 2020) is to handle data multimodality in off-line training while being implemented on-line where the active update of a monitoring model is also taken into consideration. Fig. 4.6 presents a flowchart summarising the framework.

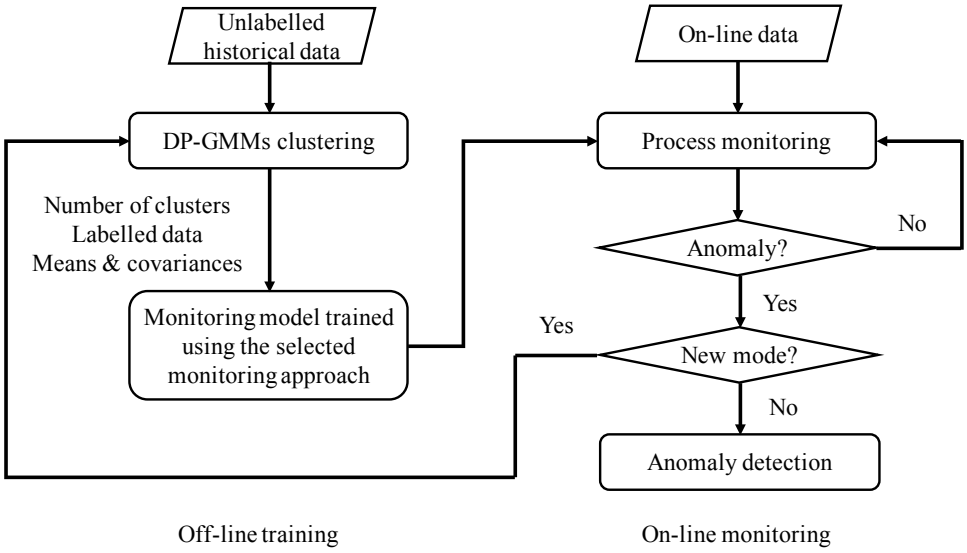


Figure 4.6: A monitoring framework for on-line anomaly detection and monitoring model update: the off-line training procedure of monitoring model is shown on the left of the diagram while the on-line monitoring is on the right part. For the off-line training, firstly, unlabelled historical data are partitioned into several clusters using the DP-GMMs algorithm. Secondly, the derived number of clusters, labelled data, means and covariances are passed to the selected monitoring algorithm for training monitoring model. The MSPM algorithm candidate could be the Non-stationary Discrete Convolution (NSDC) kernel-based method (Tan et al., 2019, 2020), which requires the clusters of data to be known and is suitable to account for multimode processes. Given the trained monitoring model, the incoming data can be identified as normal or faulty operation. If the data are recognised as abnormal, and from new modes, the data will be incorporated into the DP-GMMs clustering step along with historical data. The training model will be updated using the new clustering results and Gaussian parameters associated to each cluster.

4.7.1. Off-line training

The left side of Fig. 4.6 is the work flow of the off-line training. The DP-GMMs approach clusters the unlabelled data recorded from multiple normal operating modes. Then, the clustering results, such as the number of clusters, labelled data and means and covariances of each cluster, are used in training monitoring models.

Data labelling

The first step of most common frameworks for monitoring multimode processes is to explicitly label each mode data (Zhang et al., 2018). If the class labels are available, the historical data can be classi-

fied using classification methods. Widely used methods include Fisher discriminant analysis-based, support vector machine-based and Gaussian mixture model-based methods (Bishop, 2006; Ge et al., 2017). However, class labels are hard and expensive to obtain because manually labelling a large quantity of historical data requires massive time and manpower devotion. Therefore, clustering algorithms without the prior knowledge of class labels are considered. Existing clustering methods, such as K-means (Lloyd, 1982), expectation maximisation method (Dempster et al., 1977), and hierarchical clustering (Murtagh and Contreras, 2011), typically requires the number of clusters to be known in advance (Rui Xu and Wunsch, 2005). In contrast, the DP-GMMs can automatically determine the number of clusters when labelling the data.

Monitoring models

Various approaches in Multivariate Statistic Process Monitoring (MSPM) have been proposed for processes with multiple operating modes. In the joint paper with Tan et al. (2019, 2020), a Non-stationary Discrete Convolution (NSDC) kernel-based method was employed. This method is able to address the covariance structure of each mode in the kernel design, and has been applied to multimode process monitoring to build single monitoring models (Tan et al., 2019).

4.7.2. On-line monitoring

The right side of Fig. 4.6 is the work flow of on-line monitoring. In the on-line monitoring state, the on-line data is determined as either normal or anomalous using the NSDC kernel-based monitoring model. The detection of an anomaly indicates that the process has run at an operating mode that has not been seen in the training. At this stage, additional expert knowledge is required to categorise the anomaly as fault or a new mode. The confirmation of a new mode activates the clustering using the DP-GMMs clustering methods described in this chapter making use of historical data as well as the data from the new operating mode. Moreover, the monitoring model is re-trained using the updated number of clusters, labelled data and means and covariances. If there is no evidence to support the detected anomaly is a new operating mode, the end user will conclude that there is a fault in the process.

4.8. Summary

In this chapter, the Dirichlet Process-Gaussian Mixture Models (DP-GMMs) have been introduced. The DP-GMMs take advantage of partitioning a set of unlabelled data into clusters that follow multivariate Gaussian distributions without the number of clusters as a-prior. The Normal Inverse Wishart (NIW) distribution is the commonly used priori for multivariate Gaussian distributions due to its conjugacy. Its derivation has been presented. To implement the DP-GMMs clustering, Gibbs sampling was reviewed. The Bayesian inference in the DP-GMMs has been introduced.

A prominent value of DP-GMMs clustering relates to the fact that there is no requirement to specify the number of clusters. Such an approach may be used to automatically label historical data. However, the NIW distribution has issues with its parameter initialisation, which may impact the parameter estimation of the Gaussian distribution associated to each cluster, particularly for the covariance, and clustering results. In this chapter, the DP-GMMs clustering algorithm was implemented on a multimode simulation model. In addition, the potential impact with two kinds of parameter initialisation was

analysed, one of which was tested in a simulated example. The simulation experiment showed that improper parameter settings could lead to errors in covariance estimation; also, mis-clustering might be caused using these settings.

The main novelty of this chapter is to propose a method to minimise the impact from improper initialisation. Before clustering, it is advised that each variable of the unlabelled data are centred to zero and adjusted to variance of one. For clustering, a determination method of κ_0 was given. After clustering, the parameters of each cluster are re-estimated by the sample mean and sample covariance. The re-estimation operation aims to provide accurate statistical information related to each cluster. The proposed measures have been validated on a 5-mode simulation model. It showed that the clustering performance could be greatly improved by using the proposed normalisation and parameter initialisation.

A monitoring framework incorporating the clustering ability of DP-GMMs was proposed in collaboration work (Tan et al., 2019, 2020). Except for classifying the unlabelled historical data, the DP-GMMs are also used in the situation that on-line data are detected as new operating modes. In this case the historical data and on-line data are combined and clustered using the DP-GMMs approach in order to update the information of normal operation required in training monitoring models.

5. Field Kalman Filter (FKF) for process monitoring

In a process, various operating modes might have the same or similar steady-states, but with distinct dynamics. Monitoring algorithms that do not take dynamics into account would fail to distinguish such modes. In addition, there might be faulty operation occurring in a process. The ability of inferring the anomalies should also be included in the algorithm design.

The Field Kalman Filter (FKF) is a model-based Bayesian algorithm, being capable of simultaneously estimating the state, system parameter and noise parameter. This advantage can be applied to differentiating various operating modes both deterministically and stochastically. In addition, this chapter investigates the potential of extending the FKF for anomaly detection from multimode processes. The organisation of this chapter is as follows: Section 5.1 revisits the Kalman filter and introduces the FKF. The existing methods of model-based multimode process monitoring are reviewed and the challenges and opportunities are analysed in Section 5.2. The Multivariate Autoregression State-Space (MARSS) model is selected as the system identification method for modelling normal process operation (see Section 5.3). Section 5.4 develops the applications of the FKF to fault/mode classification and anomaly detection. The procedure of off-line training an FKF monitoring model and on-line monitoring is illustrated in Section 5.5. To validate the proposed PCM approaches, two simulated case studies are given in Section 5.6. A univariate model was used for a scenario of fault detection and fault isolation. A multivariate multimode model was used for evaluating the performance of MARSS models in anomaly detection and mode identification applications. This chapter builds upon the work of Baranowski et al. (2017) and Cong et al. (2020).

5.1. Introduction to the FKF

5.1.1. Review of the Kalman filter

Theoretical formulation of the Kalman filter

The Kalman filter was introduced by Kalman (1960). His idea was to construct a state estimator on the properties of conditional Gaussian random variables Roux (2003). Taking object tracking as an example, the velocity and position of the object at a certain time is the state of the system. The estimation criterion is that the current state is deduced from the previous state and current measurement. The Kalman filter is an optimal state estimator (Anderson and Moore, 1979), and guarantees the variance of the state estimation is minimal, when the following conditions hold (Matisko and Havlena, 2013):

- The system is Linear Time Invariant (LTI) and the mathematical model in state-space form of this system is exactly known;

- The state and the measurement noises are white processes;

- The covariance matrices of the state and the measurement noises are known.

Consider the discrete-time LTI system:

$$\boldsymbol{x}[t+1] = A[t]\boldsymbol{x}[t] + B[t]\boldsymbol{u}[t] + \boldsymbol{w}[t] \tag{5.1}$$

$$\boldsymbol{y}[t] = C[t]\boldsymbol{x}[t] + \boldsymbol{v}[t] \tag{5.2}$$

$$\boldsymbol{w}[t] \sim \mathcal{N}(0, W)), \boldsymbol{v}[t] \sim \mathcal{N}(0, V).$$

Eq. (5.1) is the state equation in which $\boldsymbol{x}[t] \in \mathbb{R}^m$ is the state to be estimated, $\boldsymbol{u}[t] \in \mathbb{R}^l$ is the system input and $\boldsymbol{w}[t] \in \mathbb{R}^m$ is the state noise independent of $\boldsymbol{x}[t]$ which follows a Gaussian distribution with zero mean and state noise covariance $W \in \mathbb{R}^{m\times m}$. $A[t] \in \mathbb{R}^{m\times m}$ and $B[t] \in \mathbb{R}^{m\times l}$ represent the state-transition matrix and the input-control matrix, respectively. Eq. (5.2) is the measurement equation in which $\boldsymbol{y}[t] \in \mathbb{R}^r$ is the measurement and $\boldsymbol{v}[t] \in \mathbb{R}^r$ is the measurement noise and follows a Gaussian distribution with zero mean and measurement noise covariance $V \in \mathbb{R}^{r\times r}$. $C[t] \in \mathbb{R}^{r\times m}$ represents the measurement-control matrix. $\mathcal{N}$ represents the Gaussian distribution.

Let $\boldsymbol{y}_t = \boldsymbol{y}[t]$ and $\boldsymbol{x}_t = \boldsymbol{x}[t]$. Denoting $\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots$ as a sequence of measurements, $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots$ as the corresponding states and Y_{t-1} as the past measurements in the time interval $(0, t]$. The expression of Y_{t-1} is:

$$Y_{t-1} = \begin{cases} \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_{t-1}\}, & \text{for} \quad \text{t = 2, 3, } \ldots \\ \varnothing, & \text{for} \quad \text{t = 1.} \end{cases} \tag{5.3}$$

The details of the Kalman filter are presented in Algorithm 2. To estimate the state, there are two main steps:

– Prediction of distribution $P(\boldsymbol{x}_t|Y_{t-1})$. The optimal prediction of $\boldsymbol{x}_t$ is described with $\boldsymbol{x}_t^-$ as Eq. (5.4c) and S_t^- as Eq. (5.4d).

– Correction of distribution $P(\boldsymbol{x}_t|Y_{t-1}, \boldsymbol{y}_t) = P(\boldsymbol{x}_t|Y_t)$. The corrected variables corresponding to $\boldsymbol{x}_t^-$ and S_t^- are $\boldsymbol{x}_t^+$ and S_t^+, respectively, shown in Eq. (5.4i) and Eq. (5.4j). To implement the Kalman filter in practice, for $t = 0$, $\boldsymbol{x}_0^+$ is initialised as a zero vector and S_0^+ is an identity matrix.

Limitations of the Kalman filter

When implementing the Kalman filter, the evaluation of the covariance matrices with respect to the physics of systems is non-trivial, especially for the state noise assessment (Formentin and Bittanti, 2014). The noise covariance matrices are typically identified using experimental data. Bishop et al. (2001) commented that in practice, it may be possible to evaluate the measurement noise covariance through recorded output measurements. However, the determination of the state noise covariance is challenging due to the non-observability of the process states.

To address the estimation problem of noise covariances, the adaptive Kalman filter (Mehra, 1970; Gao et al., 2012) was proposed. Odelson et al. (2006) introduced an Autocovariance Least Squares (ALS) method. Modified ALS methods can be found in (Rajamani and Rawlings, 2009; Åkesson et al., 2008). Matisko and Havlena (2013) put forward a Bayesian approach using Monte Carlo simulations. An maximum likelihood method (Bar-Shalom, 1972) and an ensemble Kalman filter (Zhou et al., 2012)

Algorithm 2: Kalman filter

Initialisation

$t = 1,$ (5.4a)

$\boldsymbol{x}_0^+,\ S_0^+$ (5.4b)

Prediction step

Predicted state:

$\boldsymbol{x}_t^- = A[t]\boldsymbol{x}_{t-1}^+ + B[t]\boldsymbol{u}[t]$ (5.4c)

Predicted noise covariance of state:

$S_t^- = A(t)S_{t-1}^+ A[t]^\top + W$ (5.4d)

Correction step (5.4e)

Predicted measurement:

$\boldsymbol{y}_t^- = C[t]\boldsymbol{x}_t^-$ (5.4f)

Predicted noise covariance of measurement:

$M_t^- = V + C[t]S_t^- C[t]^\top$ (5.4g)

Kalman gain:

$K_t = S_t^- C[t]^\top {M_t^-}^{-1}$ (5.4h)

Corrected state:

$\boldsymbol{x}_t^+ = K_t(\boldsymbol{y}_t - \boldsymbol{y}_t^-) + \boldsymbol{x}_t^-$ (5.4i)

Corrected noise covariance of state:

$S_t^+ = (\boldsymbol{I} - K_t C[t])S_t^- \times (\boldsymbol{I} - K_t C[t])^\top + K_t V {K_t}^\top$ (5.4j)

New time stamp:

$t = t + 1$ (5.4k)

Go to prediction step until process stops

Note: superscript $^+$ denotes the corrected variable and $^-$ denotes the predicted variable.

were proposed to simultaneously estimate the state and parameter. However, the concurrent estimation of state, parameter and noise covariance is not fully resolved.

5.1.2. Field Kalman Filter (FKF)

The FKF, rooted in a parameter-dependent state-space form, is an extension of the Kalman filter. The parameter is from a continuous space. Its purpose is to simultaneously estimate the state, parameter and noise covariance (Bania and Baranowski, 2016).

Theoretical formulation of the FKF

The state-space model defined in the FKF is :

$$\begin{aligned} \boldsymbol{x}[t+1] &= A(\theta)\boldsymbol{x}[t] + B(\theta)\boldsymbol{u}[t] + \boldsymbol{w}[t] \\ \boldsymbol{y}[t] &= C(\theta)\boldsymbol{x}[t] + \boldsymbol{v}[t] \\ \boldsymbol{w}[t] &\sim \mathcal{N}(0, W(\theta)), \boldsymbol{v}[t] \sim \mathcal{N}(0, V(\theta)) \end{aligned} \tag{5.5}$$

where $\theta \in \Omega \subset \mathbb{R}^p$ is a vector of parameters and the matrix functions A, B, C, W, V are of C^1 class [1] w.r.t. θ. Of appropriate dimensions, $A(\theta)$, $B(\theta)$, $C(\theta)$, $W(\theta)$ and $V(\theta)$ respectively denote the state transition matrix, input-control matrix, measurement-control matrix, state noise covariance and measurement noise covariance, and are associated with θ.

The main objective of the FKF is to simultaneously estimate the distribution of state $\boldsymbol{x}_t$ and parameter θ, $P(\boldsymbol{x}_t, \theta)$. The estimation of $P(\boldsymbol{x}_t, \theta)$ is a recursive process incorporating the information of past measurements Y_{t-1} as well as the current measurement $\boldsymbol{y}_t$. There are two main steps for each iteration in the recursive process summarised as:

- Prediction of joint distribution $P(\boldsymbol{x}_t, \theta | Y_{t-1})$;
- Correction of joint distribution $P(\boldsymbol{x}_t, \theta | Y_{t-1}, \boldsymbol{y}_t)$.

Readers who are interested in the detailed derivation of the FKF are guided to (Bania and Baranowski, 2016). The discrete FKF is elaborated in the Section 5.4.

5.2. Model-based multimode process monitoring

5.2.1. Existing methods

Most of modern industrial processes consists of multiple units and components to deliver the final product. To describe the inherent interactions between the inputs and outputs of such processes, physical or explicit mathematical models are applicable. Particularly, models in state-space form are practical in process monitoring (Jin and Shi, 1999), due to their convenient approximation of finite-order Linear Time-Invariant (LTI) systems in time-domain analysis (Yu and Falnes, 1995). Nevertheless, a single model can only describe specific temporal relationships and correlations between variables. Considering a process with multiple operating modes, there are a variety of connections between variables to be described. For such cases, multiple state-space models are required for monitoring a multimode process (Tan et al., 2012).

An example of multiple state-space models is a bank of Kalman filters which are able to formulate a system with finite dimensional, usually discrete-time, linear dynamic subsystems. Each dynamic subsystem can be expressed with a state-space model. Similar to traditional Kalman filter, a bank of Kalman filters can recursively estimate the current state using the previous states and current measurements, yet concurrently over multiple models. The discrepancies between the estimates and measurements are called residuals. In a bank of Kalman filters, each model has its own distinct residuals. In practical applications, the residuals from multiple models can be used for classification problems, such as differentiating various sensors and actuators in aircraft engines(Kobayashi and Simon, 2005; Xue et al., 2007), degrees of freedom of a quad rotor (Pebrianti et al., 2016), detecting fading channels in mobile communications(Rong

[1] C^1 class refers to all of the differentiable functions whose derivative is continuous.

Chen et al., 2000) and mechanical failure and sensor failure(Huang et al., 2012). In addition, these residuals can be further developed to statistical decisions by combining Bayes' theorem. Alkahe et al. (2002) and Meskin et al. (2013) used such statistical decisions for fault diagnosis.

The FKF can be treated as a bank of Kalman filters with infinite subsystems, the parameters and noise parameters of which are continuously distributed. In the FKF, the state, parameter, and noise covariance of a system with infinite subsystems are simultaneously estimated using Bayesian methods. As the estimation is a nonlinear problem, the FKF uses an appropriately selected bank of Kalman filters and a Bayesian updating scheme to approximate the joint posterior of state and parameters. When the system and noise parameters have discrete distributions, the nature of the FKF is a bank of Kalman filters. Thus, no approximation is needed to reconstruct the posterior distribution. In the application of the FKF to process monitoring, the FKF holds the following advantages over other comparable methods. Firstly, as the FKF takes the distributions of noise parameters into consideration, the FKF-based monitoring algorithm can differentiate various operating modes deterministically and stochastically. Secondly, a forgetting factor is formulated into the FKF such that when system parameters changes, the FKF can continue to perform monitoring reliably. Thirdly, the Bayesian monitoring statistics are interpretable and traceable for end users.

5.2.2. Challenges and opportunities

Multimode process monitoring based on Bayesian statistical decision methods is advantageous because the decision results are interpretable and visualisable. However, Bayesian methods for anomaly detection are limited to the known process operation. Therefore, unknown operation, such as the occurrence of faults or new operating modes, might be undetectable. Some research efforts have been paid to such a problem. Considering a process with multiple operating modes, Song et al. (2007) built statistical models and designed monitoring indicators for each mode. The operating mode is classified using Bayes' theorem while anomalies are distinguished using the monitoring indicator of the classified mode. Instead of working with multiple monitoring models and indicators, Yu and Qin (2008) and Ge and Song (2010a) proposed strategies of unifying Bayesian statistical decisions into a single monitoring indicator for anomaly detection.

However, it should be noted that in practice, the use of Bayes' theorem might encounter numerical problems. As introduced in Section 4.2.3 of Chapter 4, the Bayesian statistical decisions refer to posterior probabilities. The denominator of posterior probabilities is the sum of products of the prior and likelihood probabilities. The occurrence of anomalies will cause the denominator to approach to zero. As a result, posterior probabilities cannot be calculated. Nevertheless, the denominator can be utilised as a monitoring indicator, because it is the same across all the known operating modes. Additionally, such a monitoring indicator can also avoid the numerical problem. If the monitoring indicator is extremely small, the calculation of posterior probabilities for mode identification can be sidestepped.

5.3. System identification of process models of normal operation

In control engineering, system identification is a methodology to obtain the mathematical models of dynamic systems from the input and/or output measurements of a system (Hong et al., 2008; Ding et al., 2010; Westwick and Perreault, 2011; Dahunsi et al., 2010; Kolodziej and Mook, 2011). To imple-

ment model-based methods, mathematical models are required. Multivariate Autoregressive State-Space (MARSS) models are proposed to model normal operation. For linear systems, system identification has been well-researched and practically solved. Hence, the presentation of LTI systems in state-space form is available.

5.3.1. Motivations of the use of MARSS models

A large scale plant is complex due to a number of interacting subsystems. Given the complexity, it is challenging to derive an explicit model via first-principles. In cases where there is minimal or no prior knowledge of the physics of the system, having knowledge of the system inputs and outputs is a form of prior knowledge of the physics of the system, it is able to build data-driven system models. Particularly, from industrial plants, there are sufficient healthy data to be collected. In addition, the local dynamics existing in the healthy data may be well described by linear discrete models.

According to the data source category, there are two mainstream ways of applying historical data to system identification, respectively input-output data-based methods (Favoreel et al., 2000; Figwer, 2015; Larimore, 1990; Perreault et al., 1999; Risuleo et al., 2016) and output-only data-based methods(Bonciolini et al., 2017; Chang et al., 2017; Vicario et al., 2015; Holmes et al., 2012). When considering input-output system identification methods, it is necessary to clearly specify which variables are inputs to the model and which variables are outputs. However, it is difficult to distinguish input variables from a number of measured variables. Output-only system identification is a generalised way for historical data-based modelling. This method treats all the measured variables as output variables, thus removes the necessity of the input variable selection. In this thesis, since the model in the FKF is in state-space form, the output-only MARSS models is adopted. Autoregressive (AR) models have been used for modelling a steady-state process (Akaike, 1969). An AR model may be converted to a state-space form(Holmes et al., 2012).

5.3.2. Methodology of MARSS models

Data preparation

Typically, the first step of historical data-based approaches for monitoring a process with multiple operating modes is data partitioning (Zhang et al., 2018). Mixed data of various operation modes can be partitioned either manually by experts or by clustering algorithms such as the Dirichlet Process-Gaussian Mixture Models (DP-GMMs) introduced in Chapter 4. Assuming that Y_{H} is the mixed multimode historical data and $Y_{\mathrm{H}}^{(j)}$ is the historical data for the j-th operating mode, according to the operating modes, $Y_{\mathrm{H}} = \{Y_{\mathrm{H}}^{(j)}, j = 1, 2, \ldots, J\}$ is the resulting data partitioning, where J is the number of operating modes. For the j-th operating mode, historical data $Y_{\mathrm{H}}^{(j)}$ are further split into a training dataset $Y_{\mathrm{Tr}}^{(j)}$ and a validation dataset $Y_{\mathrm{Va}}^{(j)}$. It should be noted that, either in the data partitioning step or in the splitting step, the data samples are indexed in time order. $y_t^{(j)}$ denotes the sample in $Y_{\mathrm{H}}^{(j)}$ with time stamp t.

Multivariate Autoregressive (MAR) models

Having k autoregressive terms and a constant term, the MAR model of r-dimension takes the form

$$\boldsymbol{y}^{(j)}[t+1] = \Phi_1^{(j)}\boldsymbol{y}^{(j)}[t] + \cdots + \Phi_k^{(j)}\boldsymbol{y}^{(j)}[t-k+1] + \Phi_0^{(j)} \tag{5.6}$$

where $\boldsymbol{y}^{(j)}[t+1], \boldsymbol{y}^{(j)}[t], \ldots, \boldsymbol{y}^{(j)}[t-k+1] \in \mathbb{R}^r$ represents measurement vectors of the j-th mode equally spaced in time. The measurement value corresponding to $\boldsymbol{y}^{(j)}[t]$ is $y_t^{(j)}$. $\Phi_1^{(j)}, \ldots, \Phi_k^{(j)} \in \mathbb{R}^{r \times r}$ are coefficient matrices and $\Phi_0^{(j)} \in \mathbb{R}^r$ is a coefficient vector, of the j-th mode.

The coefficients $\Phi_1^{(j)}, \ldots, \Phi_k^{(j)}, \Phi_0^{(j)}$ can be obtained by a least squares fitting procedure (Weisberg, 1980). The procedure is to fit the model in Eq. (5.6) to an ensemble of measurements $Y_{\mathrm{H}}^{(j)} = \{y_1^{(j)}, y_2^{(j)}, \ldots\}$ (Thornhill et al., 1999).

Conversion of an MAR model to state-space form

For a process with J operating modes, the desired state-space form is

$$\begin{aligned} \boldsymbol{x}(\theta_j)[t+1] &= \hat{A}(\theta_j)\boldsymbol{x}(\theta_j)[t] + \hat{B}(\theta_j)\boldsymbol{u}[t] + \boldsymbol{w}(\theta_j)[t] \\ \boldsymbol{y}(\theta_j)[t] &= \hat{C}(\theta_j)\boldsymbol{x}(\theta_j)[t] + \boldsymbol{v}(\theta_j)[t] \\ \boldsymbol{w}(\theta_j)[t] &\sim \mathcal{N}(0, \hat{W}(\theta_j)), \boldsymbol{v}(\theta_j)[t] \sim \mathcal{N}(0, \hat{V}(\theta_j)) \\ j &= 1, \ldots, J \end{aligned} \tag{5.7}$$

where $\hat{A}(\theta_j)$, $\hat{B}(\theta_j)$ and $\hat{C}(\theta_j)$, $\hat{W}(\theta_j)$ and $\hat{V}(\theta_j)$ respectively are the estimates of state transition matrix, input-control matrix and measurement-control matrix, state noise covariance and measurement noise covariance of the j-th mode.

As Eq. (5.6) is to be converted to a state-space model, it is necessary to choose some states. There are many possible choices for the states (Holmes et al., 2012). In this thesis, the state are given by

$$\boldsymbol{x}(\theta_j)[t] = \begin{cases} \boldsymbol{x}_1[t] \\ \boldsymbol{x}_2[t] \\ \vdots \\ \boldsymbol{x}_k[t] \end{cases} \tag{5.8}$$

where

$$\boldsymbol{x}_i[t] \in \Re^r = \begin{cases} \boldsymbol{y}[t] \in \Re^r, & \text{for} \quad \text{i = 1} \\ \boldsymbol{x}_1[t-i+1] \in \Re^r, & \text{for} \quad \text{i = 2, 3, \ldots, k.} \end{cases} \tag{5.9}$$

With these definitions, $\hat{A}(\theta_j)$, $\hat{B}(\theta_j)$ and $\hat{C}(\theta_j)$ in Eq. (5.7) can be given as:

$$\begin{aligned} \hat{A}(\theta_j) &= \begin{pmatrix} \Phi_1^{(j)} & \Phi_2^{(j)} & \ldots & \Phi_k^{(j)} \\ \boldsymbol{I} & \mathbf{0} & \ldots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \ldots & \boldsymbol{I} & \mathbf{0} \end{pmatrix} \in \mathbb{R}^{rk \times rk} \\ \hat{B}(\theta_j) &= \begin{pmatrix} \Phi_0^{(j)} \\ 0 \\ \vdots \\ 0 \end{pmatrix} \in \mathbb{R}^{rk} \end{aligned} \tag{5.10}$$

$$\hat{C}(\theta_j) = \left(\boldsymbol{I} \,\vdots\, \mathbf{0} \,\vdots\, \dots \,\vdots\, \mathbf{0} \right) \in \mathbb{R}^{r \times rk}$$

where the dotted lines indicate that the vectors and matrices have sub-blocks, $\boldsymbol{I} \in \mathbb{R}^{r \times r}$ is an identity matrix and $\mathbf{0} \in \mathbb{R}^{r \times r}$ is a matrix with all elements of value 0. The variable $\boldsymbol{u}[t]$ is equal to 1, because there is a constant term in Eq. (5.6).

When state $\boldsymbol{x}_t(\theta_j)$ and measurement $\boldsymbol{y}_t(\theta_j)$, $t = 1, 2, \dots$, are available, the covariance matrices $\hat{W}(\theta_j)$ and $\hat{V}(\theta_j)$ can be derived through a nonlinear optimisation:

$$\begin{aligned} \min_{\hat{W}(\theta_j), \hat{V}(\theta_j)} \operatorname{tr} & \left[\frac{1}{n_j} \sum_{t=1}^{n_j} (\boldsymbol{x}_t(\theta_j) - \boldsymbol{x}_t^-(\theta_j))(\boldsymbol{x}_t(\theta_j) - \boldsymbol{x}_t^-(\theta_j))^\top \right] \\ & s.t. \text{ Eq.(5.7) hold}, \forall t, \text{ and } \hat{W}(\theta_j) > 0, \hat{V}(\theta_j) > 0 \end{aligned} \tag{5.11}$$

where n_j is the number of output measurements in the j-th mode, $\boldsymbol{x}_t(\theta_j)$ is the measured value of $\boldsymbol{x}(\theta_j)[t]$ and $\boldsymbol{x}_t^-(\theta_j)$ is the predicted value of $\boldsymbol{x}(\theta_j)[t]$ using the Kalman filter. Obtaining the global minima of Eq. (5.11) is non-trivial because noises are highly stochastic. Formentin and Bittanti (2014) pointed that $\hat{W}(\theta_j)$ and $\hat{V}(\theta_j)$ can be designed as positive semi-definite diagonal matrices, the elements on the diagonal of which are positive values. Besides, the diagonal form can significantly reduce the number of elements being estimated. Thus, this thesis adopts this design of $\hat{W}(\theta_j)$ and $\hat{V}(\theta_j)$.

5.4. FKF for process monitoring

Given $\theta \in \Omega = \{\theta_1, \dots, \theta_J\}$, the discretised Eq. (5.5), called the discrete FKF model, is

$$\begin{aligned} \boldsymbol{x}(\theta_j)[t+1] &= A(\theta_j)\boldsymbol{x}(\theta_j)[t] + B(\theta_j)\boldsymbol{u}[t] + \boldsymbol{w}(\theta_j)[t] \\ \boldsymbol{y}(\theta_j)[t] &= C(\theta_j)\boldsymbol{x}(\theta_j)[t] + \boldsymbol{v}(\theta_j)[t] \\ \boldsymbol{w}(\theta_j)[t] &\sim \mathcal{N}(0, W(\theta_j)), \boldsymbol{v}(\theta_j)[t] \sim \mathcal{N}(0, V(\theta_j)) \\ j &= 1, \dots, J \end{aligned} \tag{5.12}$$

where the definitions of A, B, C, W and V here are the same as in Eq. (5.5), but the associated parameter is θ_j. $\boldsymbol{x}(\theta_j)[t]$ and $\boldsymbol{y}(\theta_j)[t]$ respectively are the state and measurement given sub-model information $A(\theta_j)$, $B(\theta_j)$, $C(\theta_j)$, $W(\theta_j)$ and $V(\theta_j)$.

Under the assumption of $P(\theta_1|Y_t) \neq \dots \neq P(\theta_J|Y_t)$, $P(\boldsymbol{x}_t, \theta|Y_{t-1})$ and $P(\boldsymbol{x}_t, \theta|Y_t)$ are PMFs with J distinct components. Generally, the parameter matrices in Eq. (5.12) are unknown. In this thesis, the training of the discrete FKF model adopts the aforementioned MARSS method (see Section 5.3). It also should be noted that the actual dependence of MARSS models on parameters $\theta_1, \dots, \theta_j$ is strongly implicit. In the considered monitoring approach, the necessity of explicitly estimating these parameters is removed, as the interest is in the sub-model index j. More specifically, the discrete FKF model-based monitoring approach is essentially a marginalisation of parameters and an estimation of the sub-model index.

When the system information are fully known (e.g. for a J-mode process, $j \in \{1, \dots, J\}$), the monitoring model trained with these information can be used for classification, for example, fault diagnosis and mode identification. However, when the system information are limited, the monitoring techniques

are required to identify the unlearned operation of the plant (e.g. $j \notin \{1, \dots, J\}$ is related to the appearance of anomalies).

5.4.1. Apply the FKF for classification

Given $\forall j$, $A(\theta_j)$, $B(\theta_j)$, $C(\theta_j)$, $W(\theta_j)$ and $V(\theta_j)$, the FKF algorithm with $\theta \in \Omega = \{\theta_1, \dots, \theta_J\}$ is presented in Algorithm 3. The recursive process in the FKF includes two main steps:

– Prediction step determines the joint probability of $\boldsymbol{x}_t$ and θ using past measurements Y_{t-1}:

$$P(\boldsymbol{x}_t, \theta | Y_{t-1}) = P(\boldsymbol{x}_t | \theta, Y_{t-1}) P(\theta | Y_{t-1}) \tag{5.13}$$

where $\boldsymbol{x}_t \in \{\boldsymbol{x}(\theta_1)[t], \dots, \boldsymbol{x}(\theta_J)[t]\}$. $P(\boldsymbol{x}_t | \theta, Y_{t-1})$ is a PMF within which each component follows a Gaussian distribution. Each Gaussian component is parameterised with mean vector $\boldsymbol{x}_t^-(\theta_j)$ as Eq. (5.15f) in Algorithm 3 and covariance $S_t^-(\theta_j)$ as Eq. (5.15g) in Algorithm 3. $\boldsymbol{x}_t^-(\theta_j)$ and $S_t^-(\theta_j)$ are predicted in the same way as in a Kalman filter. $P(\theta | Y_{t-1})$ is the distribution of Gaussian components. At $t = 1$, $P(\theta | Y_{t-1}) = P(\theta | \varnothing) = P_0^+(\theta)$ is presumed uniformly distributed as Eq. (5.15e).

– Correction step determines the joint probability of $\boldsymbol{x}_t, \theta$ with Y_{t-1} and additional measurement $\boldsymbol{y}_t$:

$$\begin{aligned} P(\boldsymbol{x}_t, \theta | Y_{t-1}, \boldsymbol{y}_t) &= P(\boldsymbol{x}_t | \theta, Y_{t-1}, \boldsymbol{y}_t) P(\theta | Y_{t-1}, \boldsymbol{y}_t) \\ &= P(\boldsymbol{x}_t | \theta, Y_{t-1}, \boldsymbol{y}_t) \\ &\quad \times P(\boldsymbol{y}_t | \theta, Y_{t-1}) P(\theta | Y_{t-1}) \end{aligned} \tag{5.14}$$

where $P(\boldsymbol{x}_t | \theta, Y_{t-1}, \boldsymbol{y}_t)$ is subject to a mixture of corrected Gaussian distributions where each Gaussian component is parameterised by mean vector $\boldsymbol{x}_t^+(\theta_j)$ as Eq. (5.15j) and covariance $S_t^+(\theta_j)$ as Eq. (5.15k). Since $Y_{t-1} = \varnothing$ at $t = 1$, $P(\boldsymbol{x}_t | \theta, Y_{t-1}, \boldsymbol{y}_t) = P(\boldsymbol{x}_t | \theta, \boldsymbol{y}_t)$ in which case $\boldsymbol{x}_{t-1}^+(\theta_j) = \boldsymbol{x}_0^+(\theta_j)$ and $S_{t-1}^+(\theta_j) = S_0^+(\theta_j)$. Often $\boldsymbol{x}_0^+(\theta_j)$ and $S_0^+(\theta_j)$ are initialised as a zero vector and an identity matrix, respectively. $P(\boldsymbol{y}_t | \theta, Y_{t-1})$ can be indirectly estimated by $P(\boldsymbol{x}_t | \theta, Y_{t-1}, \boldsymbol{y}_t)$ because $\boldsymbol{y}_t$ is linearly linked to $\boldsymbol{x}_t$ and both noise covariances of $\boldsymbol{x}_t$ and $\boldsymbol{y}_t$ are Gaussian. Therefore, $P(\boldsymbol{y}_t | \theta, Y_{t-1})$ is also a mixture of Gaussian distributions with mean vector $\boldsymbol{y}_t^-(\theta_j)$ as Eq. (5.15l) and covariance $M_t^-(\theta_j)$ as Eq. (5.15m) for the j-th component.

Since $\theta \in \Omega = \{\theta_1, \dots, \theta_J\}$, there are J conditional probabilities yielded at each time instant with respect to $\boldsymbol{x}_t$ and θ. Assuming that at a given time, only a sub-model functions in the system, only one conditional probability is the best-fit to reflect the current system behaviour. The inference of the best-fit conditional probability is equivalent to finding the sub-model information $A(\theta_j)$, $B(\theta_j)$, $C(\theta_j)$, $W(\theta_j)$ and $V(\theta_j)$ at time t to give the best estimates of measurement $\boldsymbol{y}_t$.

Let $I_t \in \{1, \dots, J\}$ be the sub-model index at time t. $I_t = j$ implies that the system is running with the j-th sub-model and the estimation of $\boldsymbol{y}_t$ follows $\mathbb{E}(\boldsymbol{y}_t) = \boldsymbol{y}_t^-(\theta_j)$ and $\text{cov}(\boldsymbol{y}_t) = M_t^-(\theta_j)$ where $\mathbb{E}$ and cov denotes the expectation and covariance, respectively. The posterior probability $P(I_t = j | \boldsymbol{y}_t)$ of a system running with the j-th sub-model given $\boldsymbol{y}_t$ can be formulated according to Bayes' theorem (for brevity of the notation, we omit conditions on the past measurements Y_{t-1} and system information

Algorithm 3: Field Kalman Filter (FKF)

Initialisation

$$\theta \in \{\theta_1, \ldots, \theta_J\} \tag{5.15a}$$

$$t = 1, \forall j = 1, \ldots, J \tag{5.15b}$$

$$\boldsymbol{x}_0^+(\theta_j),\ S_0^+(\theta_j) \tag{5.15c}$$

$$P_0^+(\theta_1) = \ldots, P_0^+(\theta_J) = \frac{1}{J} \tag{5.15d}$$

$$P_0^+(\theta) = (P_0^+(\theta_1), \ldots, P_0^+(\theta_J))^\top \tag{5.15e}$$

Prediction step

Predicted state:

$$\boldsymbol{x}_t^-(\theta_j) = A(\theta_j)\boldsymbol{x}_{t-1}^+(\theta_j) + B(\theta_j)\boldsymbol{u}[t] \tag{5.15f}$$

Predicted noise covariance of state:

$$S_t^-(\theta_j) = A(\theta_j)S_{t-1}^+(\theta_j)A(\theta_j)^\top + W(\theta_j) \tag{5.15g}$$

Predicted prior distribution of $I_t = 1, \ldots, I_t = J$:

$$P_t^-(\theta) = F(\alpha)P_{t-1}^+(\theta) \tag{5.15h}$$

Correction step

Kalman gain:

$$K_t(\theta_j) = S_t^-(\theta_j)C(\theta_j)^\top M_t^-(\theta_j)^{-1} \tag{5.15i}$$

Corrected state:

$$\boldsymbol{x}_t^+(\theta_j) = K_t(\theta_j)(\boldsymbol{y}_t - \boldsymbol{y}_t^-(\theta_j)) + \boldsymbol{x}_t^-(\theta_j) \tag{5.15j}$$

Corrected noise covariance of state:

$$\begin{aligned} S_t^+(\theta_j) = &(\boldsymbol{I} - K_t(\theta_j)C(\theta_j))S_t^-(\theta_j) \\ &\times (\boldsymbol{I} - K_t(\theta_j)C(\theta_j))^\top + K_t(\theta_j)V(\theta_j)K_t(\theta_j)^\top \end{aligned} \tag{5.15k}$$

Predicted measurement:

$$\boldsymbol{y}_t^-(\theta_j) = C(\theta_j)\boldsymbol{x}_t^-(\theta_j) \tag{5.15l}$$

Predicted noise covariance of measurement:

$$M_t^-(\theta_j) = V(\theta_j) + C(\theta_j)S_t^-(\theta_j)C(\theta_j)^\top \tag{5.15m}$$

Posterior probability of $I_t = j$:

$$\begin{aligned} P_t^+(\theta_j) &= P(I_t = j|\boldsymbol{y}_t) \\ &= \frac{P(\boldsymbol{y}_t|I_t = j)P_t^-(I_t = j)}{\sum_{j=1}^J P(\boldsymbol{y}_t|I_t = j)P_t^-(I_t = j)} \end{aligned} \tag{5.15n}$$

Posterior distribution of $I_t = 1, \ldots, I_t = J$:

$$P_t^+(\theta) = (P_t^+(\theta_1), \ldots, P_t^+(\theta_J))^\top \tag{5.15o}$$

New time stamp

$$t = t + 1 \tag{5.15p}$$

Go to prediction step until process stops

$A(\theta_j)$, $B(\theta_j)$, $C(\theta_j)$, $W(\theta_j)$ and $V(\theta_j)$):

$$P(I_t = j|\boldsymbol{y}_t) = \frac{P(\boldsymbol{y}_t|I_t = j)P(I_t = j)}{\sum_{j=1}^{J} P(\boldsymbol{y}_t|I_t = j)P(I_t = j)} \tag{5.16}$$

where $P(I_t = j)$ is the prior probability of a system running with the j-th sub-model before $\boldsymbol{y}_t$ is measured, and $P(\boldsymbol{y}_t|I_t = j)$ is the likelihood probability of measuring $\boldsymbol{y}_t$ when a system is running with the j-th sub-model. The sub-model index at time t is determined by:

$$\arg\max_j P(I_t = j|\boldsymbol{y}_t). \tag{5.17}$$

The likelihood probability in Eq. 5.16 can be calculated by

$$\begin{aligned} P(\boldsymbol{y}_t|I_t = j) &= P(\boldsymbol{y}_t|y_{t-1}, y_{t-2}, \ldots, y_1, A(\theta_j), B(\theta_j), C(\theta_j), W(\theta_j), V(\theta_j)) \\ &= P(\boldsymbol{y}_t|\mathbb{E}(\boldsymbol{y}_t) = \boldsymbol{y}_t^-(\theta_j), \mathrm{cov}(\boldsymbol{y}_t) = \boldsymbol{M}_t^-(\theta_j)) \\ &= \det(2\pi \boldsymbol{M}_t^-(\theta_j))^{-\frac{1}{2}} \\ &\quad \times \exp^{-\frac{1}{2}(\boldsymbol{y}_t - \boldsymbol{y}_t^-(\theta_j))^\top (\boldsymbol{M}_t^-(\theta_j))^{-1}(\boldsymbol{y}_t - \boldsymbol{y}_t^-(\theta_j))} \end{aligned} \tag{5.18}$$

where det denotes the determinant calculation.

The common way of assigning $P(I_t = j)$ is recursive. Before having any measurements, $\forall j$, the prior probability $P(I_t = j)$ is set equal to $\frac{1}{J}$. Once measurements arrive, $\forall j$, the prior probability $P(I_t = j)$ is updated with the posterior probability $P(I_{t-1} = j|\boldsymbol{y}_{t-1})$. However, the drawback of this simple posterior-to-prior updating is that if any prior is updated with its posterior of value zero (or numerically indistinguishable from it), both prior and posterior values will be locked to zero for further computation. This can be problematic in cases of mode switching.

In similar approaches, prior probabilities are artificially lower bounded by a small value (Maybeck, 1999, Alkahe et al., 2002; Meskin et al., 2013), which is a design parameter. Another solution is the usage of a transition matrix describing the random change between sub-models. The choice of such a transition matrix is not unique, for example, a symmetric matrix (Isaksson et al., 1999). In the case of FKF to overcome this issue, a doubly stochastic matrix $F(\alpha)$ is specified (Baranowski et al., 2017):

$$F(\alpha) = \frac{1}{2}\begin{pmatrix} 1+\alpha & 1-\alpha & & & & & \\ 1-\alpha & 2\alpha & 1-\alpha & & & & \\ & 1-\alpha & 2\alpha & 1-\alpha & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & 1-\alpha & 2\alpha & 1-\alpha & \\ & & & & 1-\alpha & 1+\alpha \end{pmatrix} \in \mathbb{R}^{J\times J} \tag{5.19}$$

where the unfilled entries are zeros and $\alpha \in (0, 1)$ is a forgetting factor. The reason of choosing a doubly stochastic matrix is that the stationary distribution for $F(\alpha)$ is uniform:

$$\lim_{t\to\infty} F(\alpha)^t P_0 = (\frac{1}{J}, \frac{1}{J}, \ldots, \frac{1}{J})^\top \tag{5.20}$$

where $P_0 \in \mathbb{R}^J$ is a random vector where all the entries sum to one. Also, the use of $F(\alpha)$ can guarantees that the sum of priors is 1. Then the prior values are predicted by Eq. (5.15h) in Algorithm 3 where

$$P_t^-(\theta) = \begin{pmatrix} P(I_t = 1) \\ P(I_t = 2) \\ \vdots \\ P(I_t = J) \end{pmatrix} \tag{5.21}$$

$$P_{t-1}^+(\theta) = \begin{pmatrix} P(I_{t-1} = 1|\boldsymbol{y}_{t-1}) \\ P(I_{t-1} = 2|\boldsymbol{y}_{t-1}) \\ \vdots \\ P(I_{t-1} = J|\boldsymbol{y}_{t-1}) \end{pmatrix}. \tag{5.22}$$

The value selection of forgetting factor α determines the weights of posterior probabilities at time $t-1$ contributing to the current priors. When α is set close to its lower bound value zero, $P(I_t = J)$ is predicted with weighted $P(I_t = j), j = 1, 2 \ldots, J-1$. When α is set close to one, it means that $P(I_t = J)$ is only associated with its posterior probability at time $t-1$, $P(I_{t-1} = J|\boldsymbol{y}_{t-1})$. In this thesis, $\alpha = 0.99$. Users can adjust the value of α according to their experience or specific applications.

5.4.2. Application of the FKF for anomaly detection

This thesis also considers the case when $j \notin \{1, \ldots, J\}$, related to the appearance of anomalies. Bayes' theorem has limitations when measurements have large variance (Baranowski et al., 2017), or when the measurements are from a sub-model $j \notin \{1, \ldots, J\}$ (e.g. new operating modes or faults)(Song et al., 2007; Ge and Song, 2010a). The likelihoods of the measurements given $j \notin \{1, \ldots, J\}$ will be close to zero, leading to a numerical problem in calculating the posterior probability because the denominator in Eq. (5.16) will be close to zero. Bayes' theorem can still be applied to anomaly detection by utilising a small value as the monitoring threshold.

A monitoring indicator proposed in this thesis is L_t

$$L_t = \sum_{j=1}^{J} P(\boldsymbol{y}_t|I_t = j). \tag{5.23}$$

An anomaly is detected when the following condition holds for L_t:

$$L_t < L_{\text{LML}}. \tag{5.24}$$

where L_{LML} is the lower monitoring limit estimated from the validation data $Y_{\text{Va}} = \{Y_{\text{Va}}^{(1)}, \ldots, Y_{\text{Va}}^{(J)}\}$. L_{LML} is set with the 5th percentile of the values of L_t. These monitoring indicators are obtained by feeding Y_{Va} to the Algorithm 1 in which the calculation of the predicted and corrected prior probabilities in Eq. (5.15h), (5.15n) and (5.15o) are skipped and L_t is calculated by Eq. (5.23). The usage of validation data here is to simulate the on-line data of normal operation. The L_{LML} is a cutoff point in the validation data for flagging anomalies.

Algorithm 4: Anomaly detection and mode identification

The FKF model as (5.7) is trained according to the Section 5.3.2.
The lower monitoring limit L_{LML} is derived according to the Section 5.4.2.
while *Process continues* **do**
  $L_t = \sum_{j=1}^{J} P(\boldsymbol{y}_t | I_t = j)$
  if $L_t < L_{LML}$ **then**
    $\boldsymbol{y}_t$ is an anomaly, indicating $I_t \notin \{1, \ldots, J\}$.
    Instead of (5.15h) and (5.15n), the predicted and corrected prior probabilities are calculated using the following equations:

$$P_t^+(\theta_j) = P(I_t = j | \boldsymbol{y}_t) = 0 \quad \forall j \tag{5.25a}$$

$$P_{t+1}^-(\theta_j) = P(I_{t+1} = j) = \frac{1}{J} \quad \forall j. \tag{5.25b}$$

  else
    $\boldsymbol{y}_t$ is recognised as normal operation.
    $\forall j = 1, \ldots, J$, the predicted and corrected prior probabilities are calculated using (5.15h), (5.15n) and (5.15o).
    The mode identify I_t is determined by (5.17).
    $t = t + 1$
  end
end

Algorithm 4 presents the workflow of anomaly detection and mode identification based on the FKF. Given the FKF model and L_{LML}, the monitoring indicator at time t for the measurement $\boldsymbol{y}_t$ is calculated using Eq. (5.23). If $L_t < L_{\text{LML}}$, $I_t \notin \{1, \ldots, J\}$ denotes that $\boldsymbol{y}_t$ is an anomaly. There is no need to conduct mode identification. Hence, let $P_t^+(\theta_j) = 0 \; \forall j$ as Eq.(5.25a) be indicative of the appearance of anomaly. The predicted prior distribution of known sub-models for the next time instant are set to uniform distribution as Eq.(5.25b). If $\boldsymbol{y}_t$ is recognised as normal operation, the sub-model index is determined by Eq. (5.17).

5.5. Workflow for anomaly detection and mode identification

Fig. 5.1 summarises the workflow of how to implement the FKF for monitoring multimode processes. The dashed-box of off-line training shows the procedure of building a FKF monitoring model, including historical data labelling corresponding to the operating mode, the identification of the discrete FKF model and the determination of a monitoring threshold L_{LML}. Identifying the discrete FKF model entails MARSS learning using training data and noise estimation using validation data, resulting in J sub-models. In the on-line monitoring step, the discrete FKF model is applied to the on-line data $\boldsymbol{y}_t$ to obtain the L_t. If L_t exceeds the monitoring threshold L_{LML}, $\boldsymbol{y}_t$ is considered as an anomaly and the process is considered in anomalous operation. If $L_t < L_{\text{LML}}$ holds, $\boldsymbol{y}_t$ is recognised as normal operation. Sequentially, it will be classified to one of the known operating modes using Eq. (5.17). The prior distribution of known operating modes is updated according to Algorithm 3 so that it may be used for anomaly detection and mode identification of the next data sample.

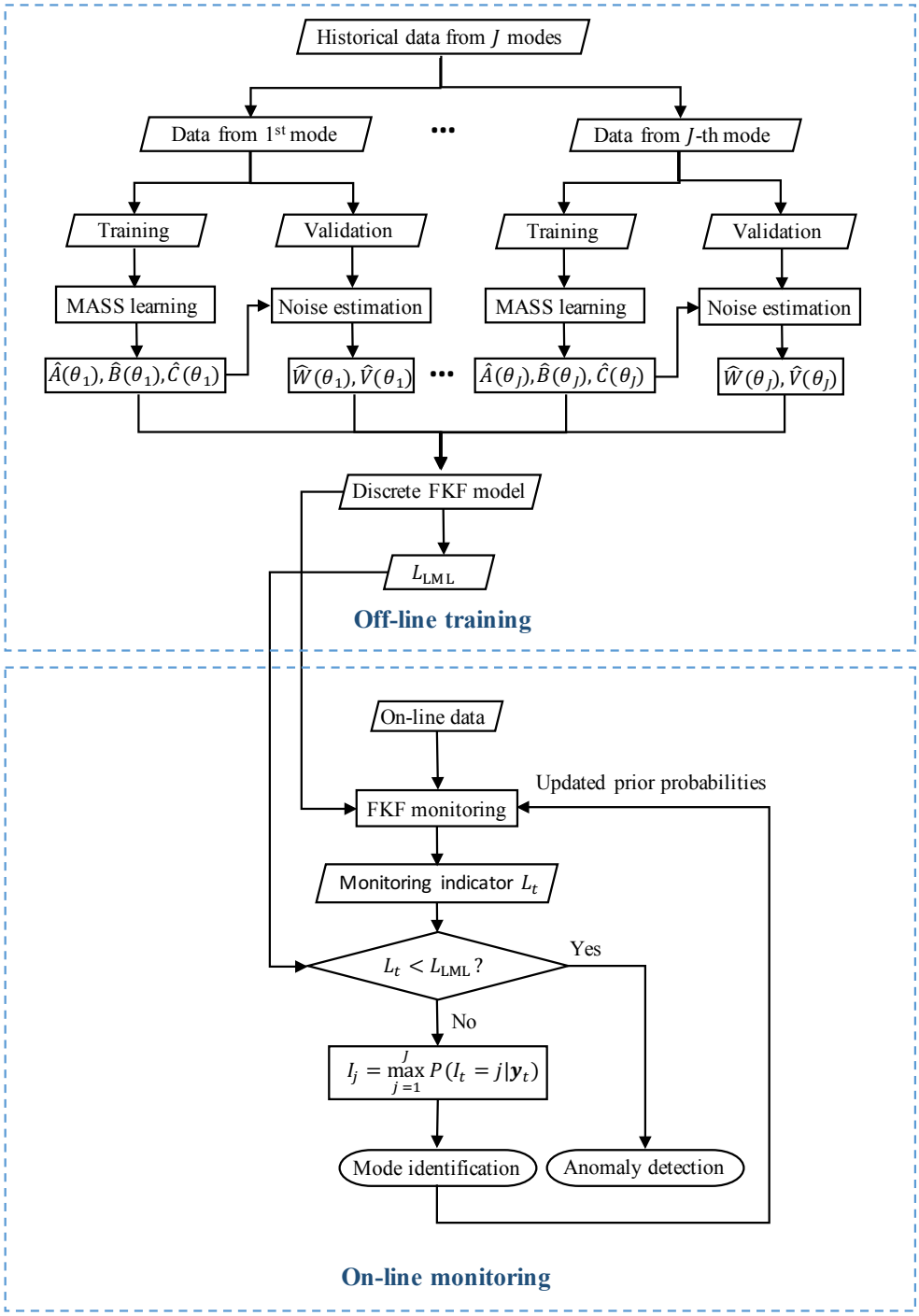


Figure 5.1: Flowchart of off-line training the FKF monitoring model and on-line monitoring: the off-line part comprises training the discrete FKF model based on MARSS models and the determination of a monitoring indicator; the on-line monitoring performs the FKF-based anomaly detection and mode identification (the flowchart credit to Ruomu Tan).

Table 5.1: Matrix specifications for simulated multimode processes

Simulated multimode model	Model number	Matrix A	Matrix B	Matrix C	Steady-states
Behaviour 1[a]	Model 1.1	$\begin{bmatrix} 0.9267 & -0.2183 \\ 0.3882 & 0.9558 \end{bmatrix}$	$\begin{bmatrix} 0.2917 \\ -0.3440 \end{bmatrix}$	$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$	$\begin{bmatrix} 1 \\ 1 \end{bmatrix}$
	Model 1.2	$\begin{bmatrix} 0.8791 & -0.4239 \\ 0.1884 & 0.9566 \end{bmatrix}$	$\begin{bmatrix} 0.5521 \\ -0.1451 \end{bmatrix}$	$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$	
	Model 1.3	$\begin{bmatrix} 0.0418 & -0.0703 \\ 0.1250 & 0.9796 \end{bmatrix}$	$\begin{bmatrix} 1.0285 \\ -0.1047 \end{bmatrix}$	$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$	
Behaviour 2[b]	Model 2.1	$\begin{bmatrix} 0.9267 & -0.2183 \\ 0.3882 & 0.9558 \end{bmatrix}$	$\begin{bmatrix} 0.6550 \\ 0.1327 \end{bmatrix}$	$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$	$\begin{bmatrix} 0 \\ 3 \end{bmatrix}$
	Model 2.2	$\begin{bmatrix} 0.9267 & -0.2183 \\ 0.3882 & 0.9558 \end{bmatrix}$	$\begin{bmatrix} 0.5834 \\ -0.6879 \end{bmatrix}$	$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$	$\begin{bmatrix} 2 \\ 2 \end{bmatrix}$
	Model 2.3	$\begin{bmatrix} 0.9267 & -0.2183 \\ 0.3882 & 0.9558 \end{bmatrix}$	$\begin{bmatrix} -0.1450 \\ -0.4324 \end{bmatrix}$	$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$	$\begin{bmatrix} 1 \\ -1 \end{bmatrix}$
Behaviour 3[c]	Model 3.1	$\begin{bmatrix} 0.9267 & -0.2183 \\ 0.3882 & 0.9558 \end{bmatrix}$	$\begin{bmatrix} 0.4375 \\ -0.5159 \end{bmatrix}$	$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$	$\begin{bmatrix} 1.5 \\ 1.5 \end{bmatrix}$
	Model 3.2	$\begin{bmatrix} 0.8791 & -0.4239 \\ 0.1884 & 0.9566 \end{bmatrix}$	$\begin{bmatrix} 1.5280 \\ -0.2468 \end{bmatrix}$	$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$	$\begin{bmatrix} 2 \\ 3 \end{bmatrix}$
	Model 3.3	$\begin{bmatrix} 0.0418 & -0.0703 \\ 0.1250 & 0.9796 \end{bmatrix}$	$\begin{bmatrix} 2.5712 \\ -0.2616 \end{bmatrix}$	$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$	$\begin{bmatrix} 2.5 \\ 2.5 \end{bmatrix}$

[a] Models with same steady-states and different dynamics;
[b] Models with different steady-states and same dynamics;
[c] Models with different steady- states and different dynamics.

5.6. Simulated case studies

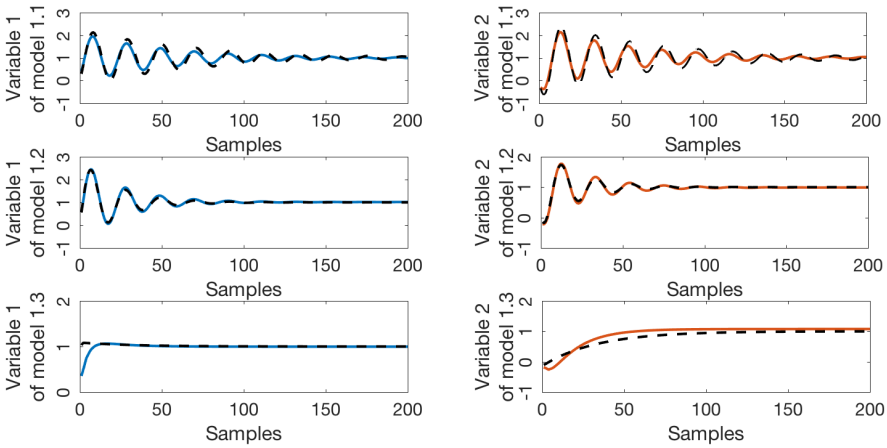
5.6.1. Performance of MARSS models

Dynamic and steady-state models

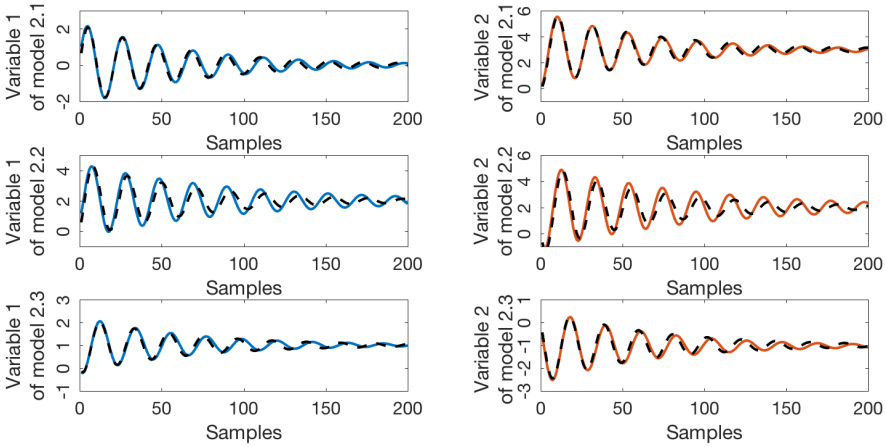
To evaluate the identification performance of MARSS models, a simulated multimode process is pre-defined. Simulated models were in state-space form, defined to give a range of under-damped and over-damped transient dynamics. Table 5.1 presents the matrix specifications of these models. Output data from the process have the following behaviours: 1) same steady-state, with three different dynamics, 2) three different steady-states with the same dynamics and 3) three different steady-states and three different dynamics.

Performance evaluation

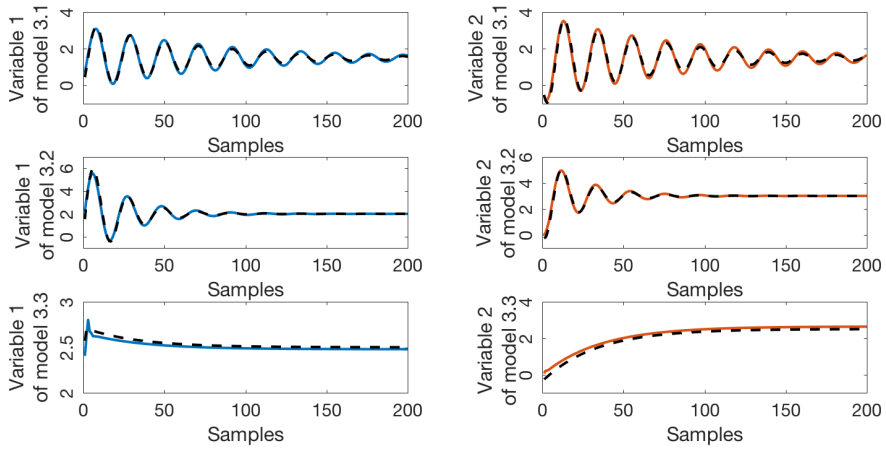
In this thesis, the number of autoregressive terms is estimated using the Partial Autocorrelation Function (PACF) against a confidence bound of 5% Adhikari and Agrawal (2013). For multivariate data, the calculation of PACF can be based on the summed squares of all measurements. With zero initial conditions and noise-free, the acquired MARSS models and predefined state-space models are excited by step response. The results of step responses are plotted in Fig. 5.2. It can be observed that the step responses of the models estimated using the MARSS learning agree well with the equivalent responses obtained for the original state-space models.



(a) Same steady-state, with three different dynamics



(b) Three different steady-states with the same dynamics

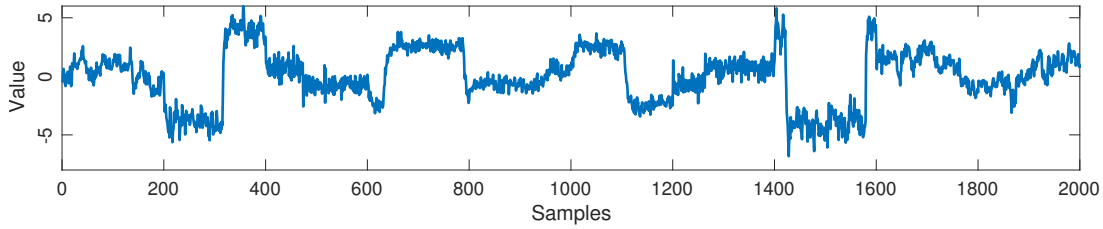


(c) Three different steady-states and three different dynamics

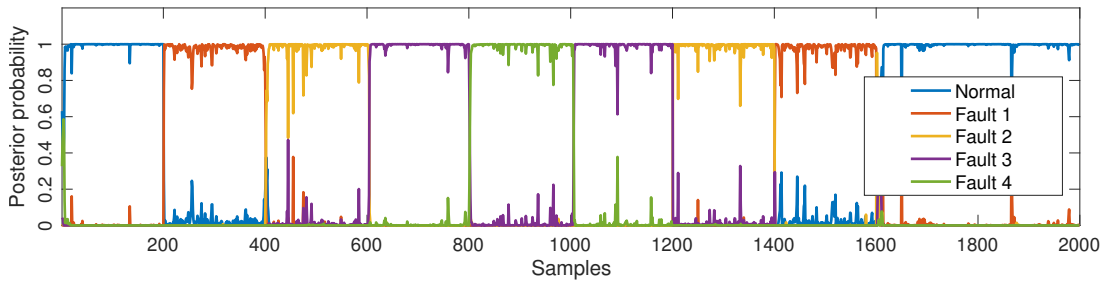
Figure 5.2: Step response comparison between the MARSS models (solid lines) and simulated models (dashed lines).

Table 5.2: System parameters dependent on operation conditions

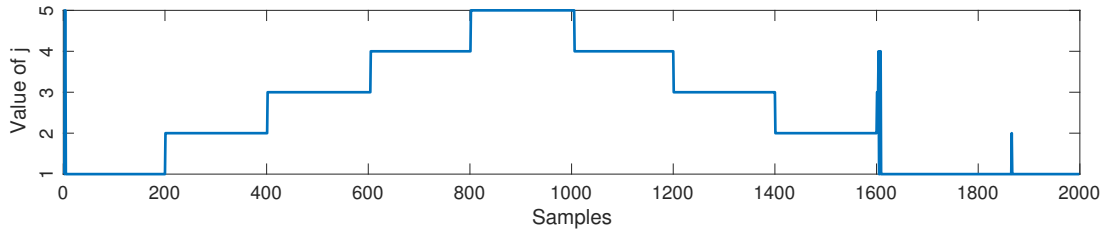
	j=1 (Normal)	j=2 (Fault 1)	j=3 (Fault 2)	j=4 (Fault 3)	j=5 (Fault 4)
$A(\theta_j)$	0.9	0.5	-0.5	0.8	0.7
$B(\theta_j)$	0.1	1	0.9	1	0.3
$C(\theta_j)$	1	2	1.2	0.5	0.6



(a) Trend plot of output measurements



(b) Bayesian statistical decisions: posterior probabilities are calculated using Eq. (5.16).



(c) Monitoring results

Figure 5.3: Apply the FKF monitoring approach to a scalar system

5.6.2. Apply the FKF to classification

Univariate system: fault detection and isolation

A univariate system with changes in parameters is considered for testing the differentiability of the FKF. Various system parameters are set for simulating one normal operation and four types of faulty operation. The specific values are shown in Table 5.2.

Relative to the normal operation, Fault 1 corresponds to the increases in system dynamics, input gain and output gain; Fault 2 introduces system oscillation while the increases in input gain and output gain are also taken into consideration; Fault 3 has its system dynamics and output gain decreased and input gain increased. Comparing with Fault 3, Fault 4 has larger system dynamics, smaller input gain and

output gain.

The initial conditions of the FKF were set with $\boldsymbol{x}_0^+(\theta_1) = \cdots = \boldsymbol{x}_0^+(\theta_5) = 0$ and $\boldsymbol{S}_0^+(\theta_1) = \cdots = \boldsymbol{S}_0^+(\theta_5) = 0.01$. The initial probabilities of these five cases are $P_0^+(\theta_1) = 0.95$ and $P_0^+(\theta_2) = \cdots = P_0^+(\theta_5) = 0.0125$. The operation scheme of the changes in parameters follows Normal, Fault 1, Fault 2, Fault 3, Fault 4, Fault 3, Fault 2, Fault 1, Normal and Normal. The duration of each operation scenario takes 200 samples. The system was excited by a sqaure input signal with zero mean, a period of 158, and the amplitude as 1.

Fig. 5.3(a) presents the plot of output measurements generated by the univariate system with various parameters. Applying the FKF algorithm to the univariate system, the Bayesian decisions plotted in Fig. 5.3(b) are calculated using Eq. (5.16). The monitoring results are highlighted in 5.3(c). It can be seen that the monitoring results are consistent with the planned operation scheme. Normal operation is distinguished from faulty operation and faulty conditions are successfully classified.

Multivariate system: mode identification

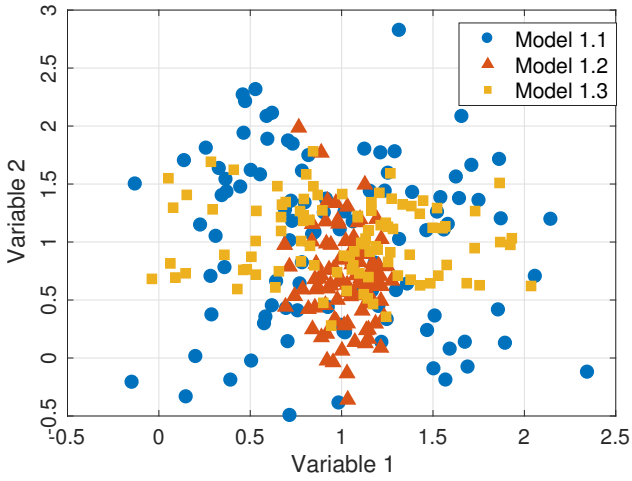
The dynamic and steady-state models in Section 5.6.1 are also used for investigating the performance of the FKF algorithm applied to mode identification. Applied with unit step inputs and uncorrelated, white Gaussian noise $\mathcal{N}(0, 0.1)$, starting at zero initial conditions, all of the state-space models were run separately to obtain training and validation data. Data of both transient response and fluctuations around steady-states were included in training and validation datasets. The test data were generated by running one of the simulation models with a unit step, noise $\mathcal{N}(0, 0.1)$ and a zero initial condition, then sequentially running the other two models. Since the jump from one simulated mode to another is instantaneous, the test data fluctuate around steady-state values.

Fig. 5.4 gives scatter plots of the test data described above. As the data points in each case overlap, it is difficult to visually distinguish each mode from one another, particularly in the first case shown in Fig. 5.4(a). Fig. 5.5(a) shows the trend plot of test data obtained by sequentially changing the operating modes, with duration of 600 samples for each mode. The on-line monitoring indicator result is presented in Fig. 5.5(b). The posterior probability of each model conditional on the measurements is calculated by Eq. (5.16) and shown in Fig. 5.5(c). The results of using models 2.1, 2.2 and 2.3 in Fig. 5.6 and models 3.1, 3.2 and 3.3 in Fig. 5.7 are obtained following the same experiment operation.

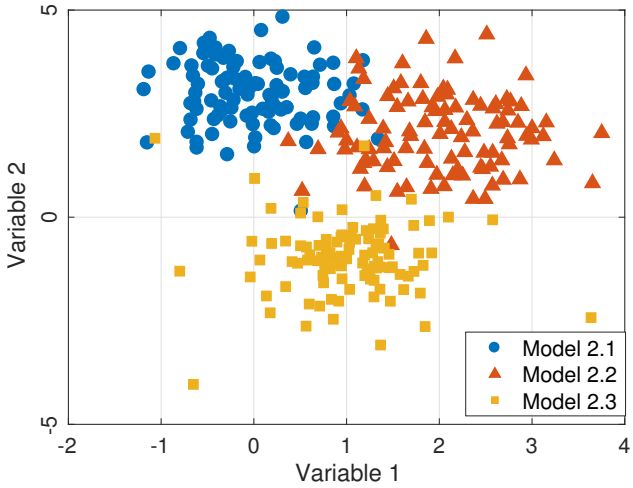
Table. 5.3 presents the performance of mode identification for the described models. The Mode Identification Accuracy (MIA) and False Alarm Rate (FAR) metrics are:

$$\begin{aligned} \text{MIA} &= \frac{n_{j,\text{classified}}}{n_j - n_{j,\text{FP}}} \\ \text{FAR} &= \frac{n_{j,\text{FP}}}{n_j} \end{aligned} \tag{5.26}$$

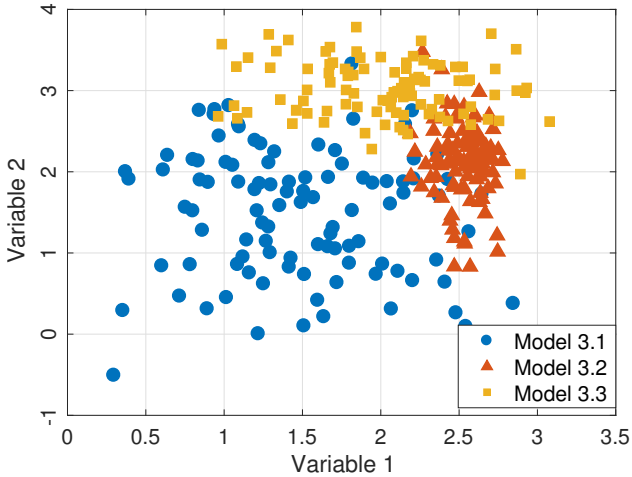
where j denotes the model number (e.g. 1.1, 2.1), $n_{j,\text{FP}}$, $n_{j,\text{classified}}$ and n_j are respectively the number of false alarms, the number of correctly classified samples and the number of samples from the model j. The results in Table 5.3 show that the FKF is capable of distinguishing data points derived from various dynamic and steady-state models.



(a) Behaviour 1: same steady-state, with three different dynamics

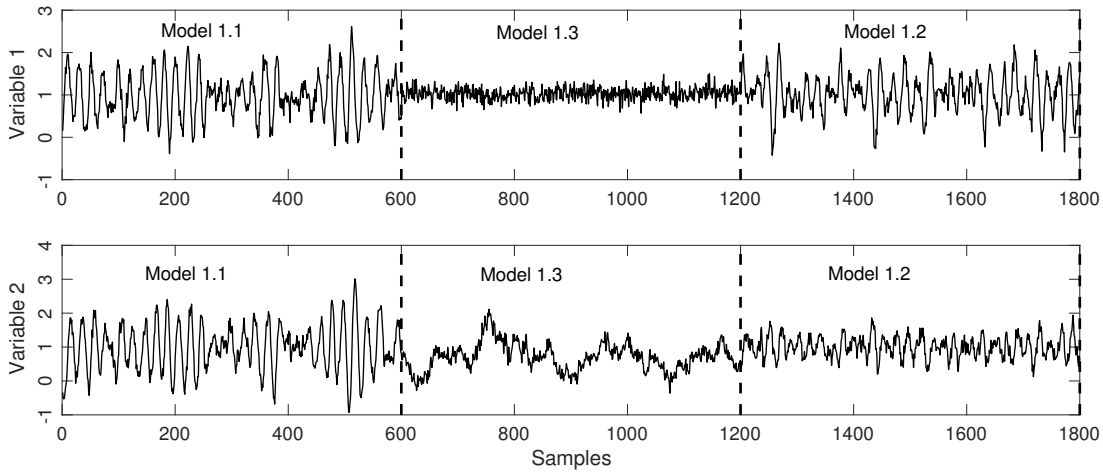


(b) Behaviour 2: three different steady-states with the same dynamics

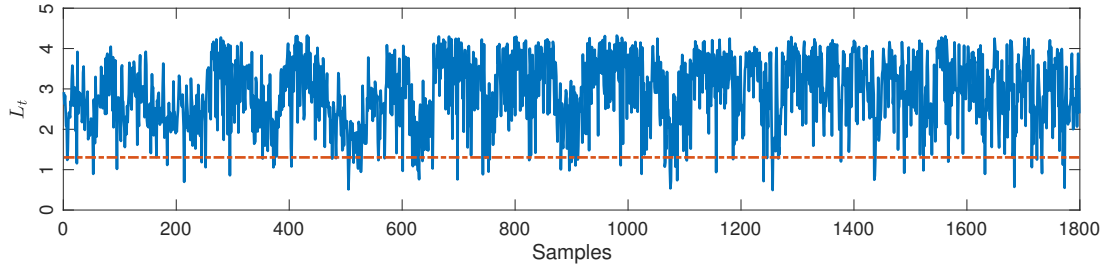


(c) Behaviour 3: three different steady-states and three different dynamics

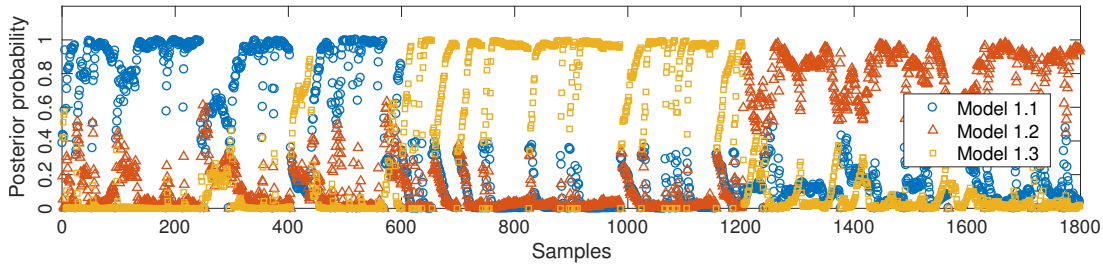
Figure 5.4: Scatter plot for test data from models defined in Table 5.1. The samples in each case are partially or completely overlapped. As models 1.1, 1.2 and 1.3 have the same steady-state, it is challenging to differentiate the samples of behaviour 1 without considering their dynamics.



(a) Trend plot of samples

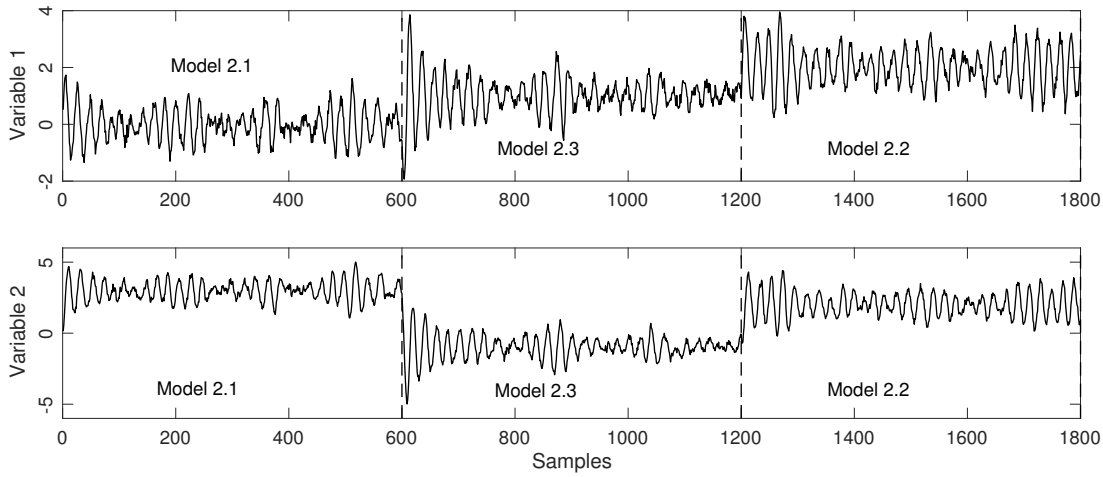


(b) On-line monitoring indicator. Blue-solid line is L_t; red-dashed line is L_{LML}.

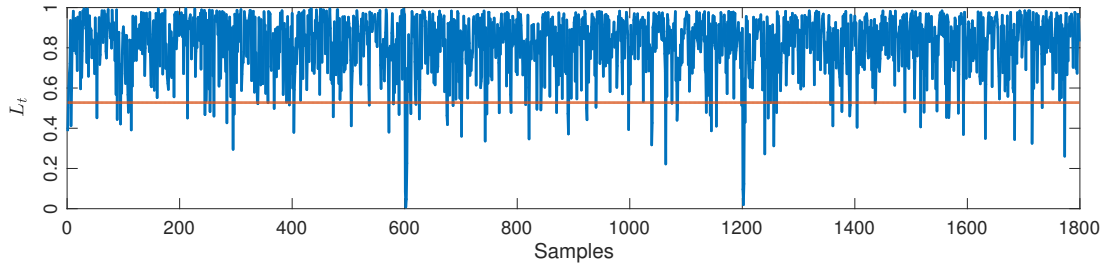


(c) On-line mode identification

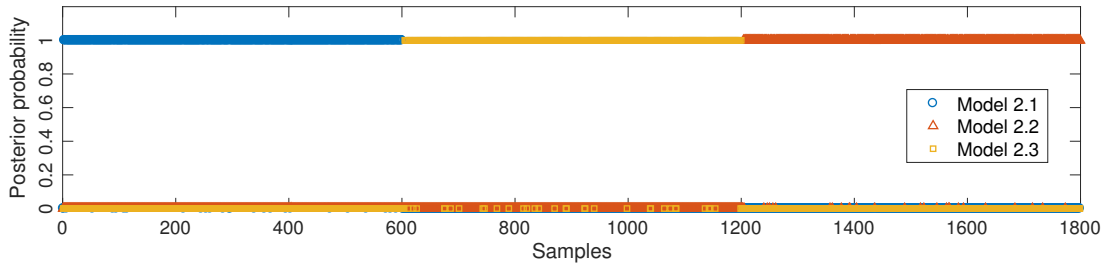
Figure 5.5: Monitoring results: the true operation scheme is model 1.1, model 1.3 and model 1.2.The training of monitoring model uses the data from model 1.1, 1.2 and 1.3 while the test data consists of data from model 1.1, 1.2 and 1.3. Sub-figure (a) is the trend plot for two variables. Sub-figure (b) plots the monitoring indicator L_t against samples. The red dashed line is the monitoring threshold. L_t below the monitoring threshold indicates anomalies. Since the indicators only occasionally dip below the monitoring threshold instead of remaining below the threshold for a prolonged period of time, thus there is no anomaly detected. Sub-figure (c) presents the mode identification results. The current mode is the one with the maximal posterior value.



(a) Trend plot of samples.

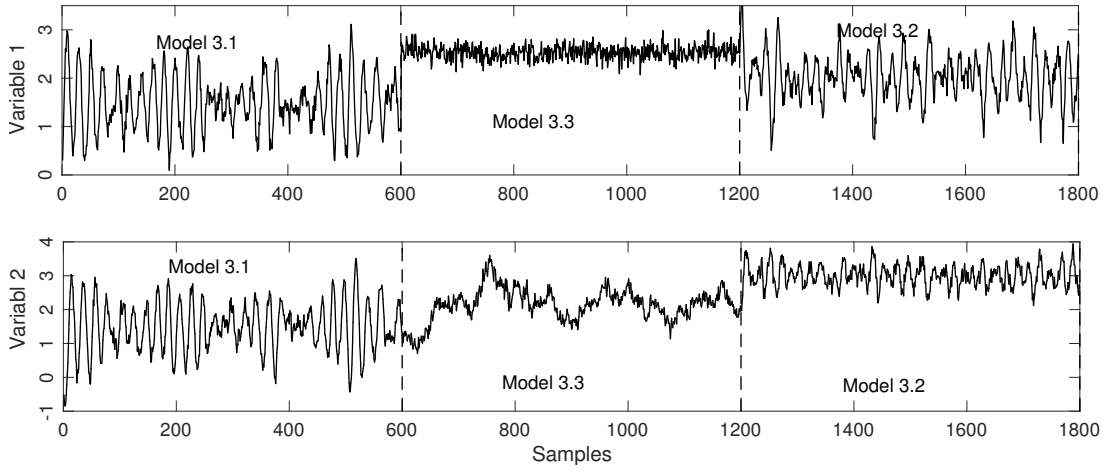


(b) On-line monitoring indicator. Blue-solid line is L_t; red-dashed line is L_{LML}.

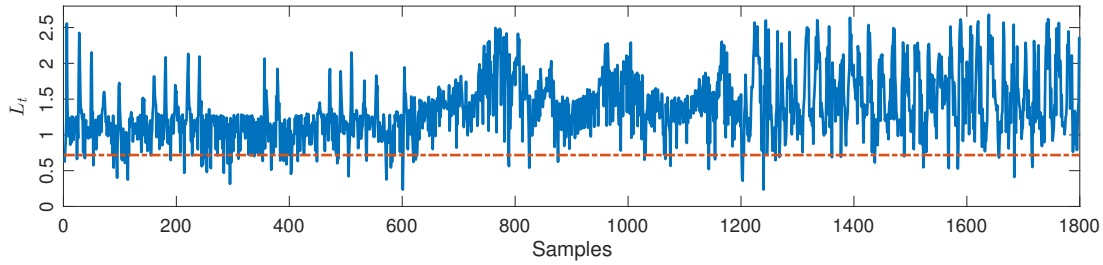


(c) On-line mode identification

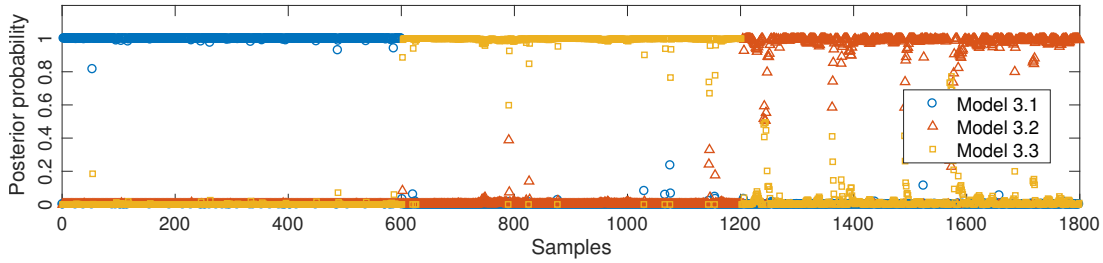
Figure 5.6: Monitoring results: the true operation scheme is model 2.1, model 2.3 and model 2.2. The training of monitoring model uses the data from model 2.1, 2.2 and 2.3 while the test data consists of data from model 2.1, 2.2 and 2.3. Sub-figure (a) is the trend plot for two variables. Sub-figure (b) plots the monitoring indicator L_t against samples. The red dashed line is the monitoring threshold. L_t below the monitoring threshold indicates anomalies. Since the indicators only occasionally dip below the monitoring threshold instead of remaining below the threshold for a prolonged period of time, thus there is no anomaly detected. Sub-figure (c) presents the mode identification results. The current mode is the one with the maximal posterior value.



(a) Trend plot of samples.



(b) On-line monitoring indicator. Blue-solid line is L_t; red-dashed line is L_{LML}.



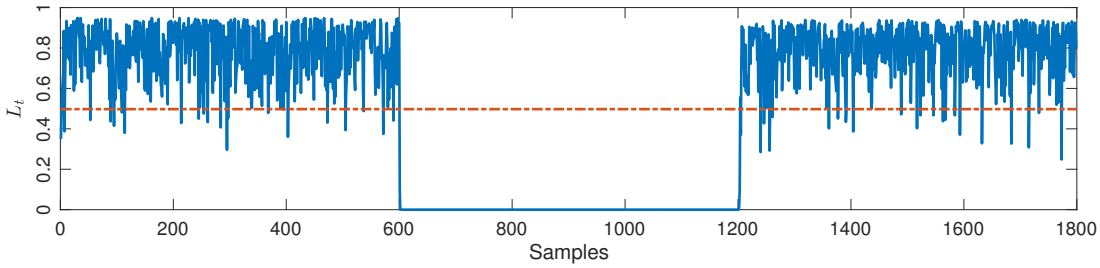
(c) On-line mode identification

Figure 5.7: Monitoring results: the true operation scheme is model 3.1, model 3.3 and model 3.2. The training of monitoring model uses the data from model 3.1, 3.2 and 3.3 while the test data consists of data from model 3.1, 3.2 and 3.3. Sub-figure (a) is the trend plot for two variables. Sub-figure (b) plots the monitoring indicator L_t against samples. The red dashed line is the monitoring threshold. L_t below the monitoring threshold indicates anomalies. Since the indicators only occasionally dip below the monitoring threshold instead of remaining below the threshold for a prolonged period of time, thus there is no anomaly detected. Sub-figure (c) presents the mode identification results. The current mode is the one with the maximal posterior value.

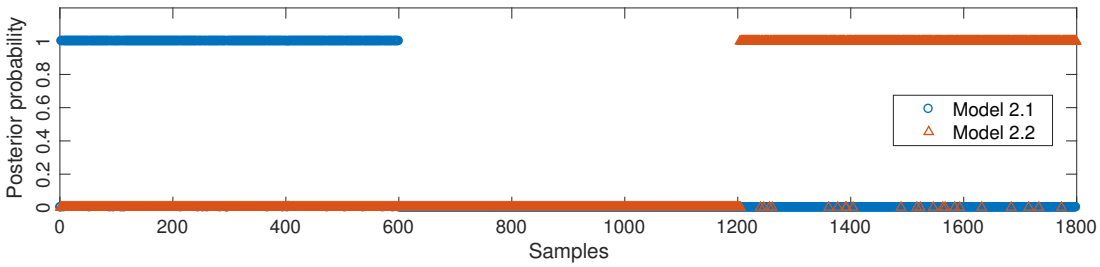
Table 5.3: Performance of mode identification on three multimode simulation models

Dynamic and steady-state models	Model number	$n_{j,\text{FP}}$[a]	FAR [b]	n_j[c] – $n_{j,\text{FP}}$	$n_{j,\text{classified}}$[d]	MIA[e]
Same steady-state, with three different dynamics	Model 1.1	23	3.83%	577	499	86.48%
	Model 1.2	18	3.01%	581	562	96.90%
	Model 1.3	35	5.83%	565	557	98.58%
three different steady-states with the same dynamics	Model 2.1	26	4.33%	574	574	100%
	Model 2.2	30	5.01%	569	569	100%
	Model 2.3	37	6.17%	563	563	100%
three different steady-states and three different dynamics	Model 3.1	49	8.17%	551	551	100%
	Model 3.2	18	3.01%	573	573	98.79%
	Model 3.3	12	2.00%	588	588	100%

[a] The number of false alarms w.r.t. model j;
[b] False Alarm Rate;
[c] The number of samples from the model j;
[d] The number of correctly classified samples from model j;
[e] Mode Identification Accuracy;



(a) On-line anomaly detection. Blue-solid line is L_t; red-dashed line is L_{LML}.



(b) On-line mode identification

Figure 5.8: Anomaly detection and mode identification when a new operating mode appears: the training of monitoring model uses the data from model 2.1 and 2.2 while the test data consists of data from model 2.1, 2.2 and 2.3. Sub-figure (a) plots the monitoring indicator L_t against samples. The red dashed line is the monitoring threshold. L_t below the monitoring threshold indicates anomalies. Samples between 600 and 1200 are from mode 2.3, and are identified as anomalies. Sub-figure (b) shows the mode identification results where the current mode is the one with the maximal posterior probability. Mode 2.1 and 2.2 are correctly recognised. As mode 2.3 is abnormal operation, its posterior values are zeros.

5.6.3. Apply the FKF to anomaly detection

This subsection demonstrates the monitoring performance for a process containing known operating modes and one new operating mode. Training and validation data were generated from state-space mod-

els 2.1 and 2.2 while the data to be classified were generated from models 2.1, 2.2 and 2.3. The FKF model only contains the dynamic and steady-state characteristics from models 2.1 and 2.2. Model 2.3 was treated as a new mode.

Fig. 5.8 shows the monitoring results of test data including known operating modes and an additional mode. The on-line operating scheme is model 2.1, 2.3 and 2.2. Initially, the posterior probability and L_t indicate that the process was operating in the mode described by known model 2.1. It can be seen that L_t quickly responds to the occurrence of model 2.3, dropping down below the red-dashed line of L_{LML}, and posterior probabilities for model 2.1, 2.3 and 2.2 are zero. When model 2.2 occurs in the process, L_t returns to a level above L_{LML} and the maximal posterior belongs to model 2.2. The results show that the proposed monitoring indicator L_t is able to detect the unknown operating modes.

5.7. Summary

In this chapter, the Field Kalman Filter (FKF) has been extended for the application of monitoring multimode processes. The discretised FKF has been proposed for classification, for example, fault detection and isolation. Further, an extension of the discretised FKF incorporating a monitoring indicator has been developed for anomaly detection and mode identification. Considering the implementation of the FKF in practice, the Multivariate Autoregression State-Space (MARSS) models have been employed for modelling the normal operation. To systematically apply the MARSS models and the FKF monitoring approach, a framework has been designed. Through the simulated case studies, the proposed framework and monitoring approach have been validated.

6. Experiment case studies

This chapter investigates the monitoring performance of the Binary Classifier for Fault Detection (BaFFle) algorithm, introduced in Chapter 3, and the Field Kalman Filter (FKF) algorithm described in Chapter 5, respectively. Validation experiments are conducted with the PRONTO dataset which is collected from an industrial-scale two-phase flow facility. With this dataset and a newly designed experiment, compared with the experiment in publications (Cong and Baranowski, 2018a,b), the fault detection performance is evaluated. The results obtained using the FKF algorithm have been previously reported in Cong et al. (2020).

6.1. Introduction to the PRONTO benchmark case study

The PRONTO benchmark dataset (Stief et al., 2018a,b, 2019) was collected from a multiphase flow facility. This multiphase flow case study simulated multiple normal operating modes and various types of faults by utilising different set-ups of the facility. Fig. 6.1 highlights the configuration of flow controls and valves in the facility. The derived dataset from this case study can be used for developing and testing algorithms dedicated to mode identification, fault detection, fault diagnosis and other process monitoring tasks. The usage of measured process variables are given in Table 6.1. The variables which are used in the validation of the BaFFle and FKF algorithms are marked with ✓.

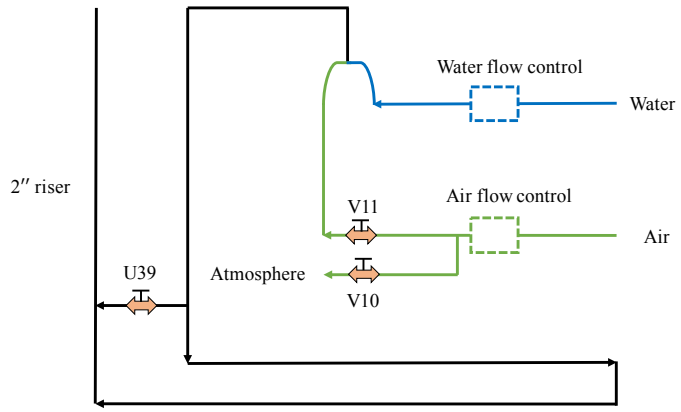


Figure 6.1: The configuration of flow controls and valves V10, V11 and U39

Table 6.1: List of process variables recorded as part of the PRONTO benchmark dataset. Variables which are used in the validation of the BaFFle and FKF methods are also highlighted.

Number	Sensor tag	Variable description	unit	BaFFle	FKF
1	FT305/302	Input air flow rate	$m^3\ h^{-1}$	✓	✓
2	PT312	Air delivery pressure	bar(g)	✓	✓
3	FT102/104	Input water flow rate	$kg\ s^{-1}$	✓	✓
4	PT417	Pressure in the mixing zone	bar(g)	✓	✓
5	PT408	Pressure at the riser top	bar(g)	✓	✓
6	FT406	2-phase separator output water flow rate	bar(g)	✓	✓
7	PT501	3-phase separator pressure	bar(g)	✓	✓
8	LI502	3-phase separator water level	%	✓	✓
9	LI503	Water coalescer level	%	✓	
10	LVC502-SR	Water coalescer outlet level	%	✓	
11	LI101	Water tank level	m	✓	

Table 6.2: Details of normal operating modes

Mode identity	Mode 1	Mode 2	Mode 3
Water flow rate ($kg\ s^{-1}$)	0.1	0.5	1
Air flow rate ($m^3\ h^{-1}$)	120	150	200

6.1.1. Normal operating modes

The simulated normal operation was performed by controlling the air and water flow rates. The specific flow rates for each mode are presented in Table 6.2. The healthy data were used for training monitoring models.

6.1.2. Faults

Three types of faults were simulated. These faults are developing fault, seeded in the system by manually adjusting valves. The faulty operation are as follows:

– Air blockage: valve V11 controls the air volume into the pipelines. A progressively worsening blockage fault was simulated by closing V11 from 90° to 10° in a stepwise fashion with a step size of 10°. As a result, the portion of air in the flow mixture was gradually reduced. The same valve closing operation was conduced both in Mode 1 and Mode 2.

– Air leakage: valve V10 controls the air volume exhausted into the atmosphere. The simulation of air leakage was performed by adjusting the valve opening to 5°, 10° and 15° in Mode 1, and to 5°, 10°, 15°, 20°, 25°, 30°,40° and 90° in Mode 2.

– Diverted flow: valve U39 controls the flowing path of air and water mix. When U39 is open, the mixed flow can be partially directed to the 2″ riser. The adjustment of U39 valve ranges from 5° to 60°.

6.2. BaFFle for fault detection

6.2.1. Density estimation considerations

Two density estimation approaches, namely Gaussian distribution and Kernel Density Estimation (KDE), are compared in this section. Since the control limits in the BaFFle algorithm are determined based on the density estimation of Principal Components (PCs), the plots of distribution of PCs can give insights to the precision of control limits. Fig. 6.2 and 6.3 demonstrate the distribution plots of PCs, for Mode 1 and Mode 2, respectively. The number of PCs are determined with the cumulative explained variance as introduced in Section 3.3.3 of Chapter 3. According to Jolliffe and Cadima (2016), a commonly used cumulative explained variance is 70%. In this analysis, 70% process variations results in 5 and 4 PCs, respectively in Mode 1 and Mode 2.

As shown in Fig. 6.2 and 6.3, there are three representations of the probability density of each PC, which are histogram, Gaussian distribution, and KDE estimation. Taking the histogram plots as the true probability distribution, it can be seen that the KDE curves are more fitting to the ground truth than the Gaussian curves, especially when the probability distributions are not symmetric and bell-shaped (e.g. Fig. 6.3 (a)). In addition, Fig. 6.2 shows that the Lower Control Limit (LCL) and Upper Control Limit (UCL) obtained from the KDE cover a narrower range defining normal operation, compared with the ones obtained from the Gaussian distribution. Nevertheless, the LCLs based on the KDE in Fig. 6.3 (c) and (d) are more relaxed than the LCLs based on the Gaussian distribution. This might be because there are bars detached from the others. For example, the bar located at -0.87 in Fig. 6.3 (c) is isolated from the other bars. The influences of the LCLs and UCLs, obtained with different density estimation approaches, on monitoring performance will be further elaborated in Section 6.2.2.

6.2.2. Experiment results

Performance metrics

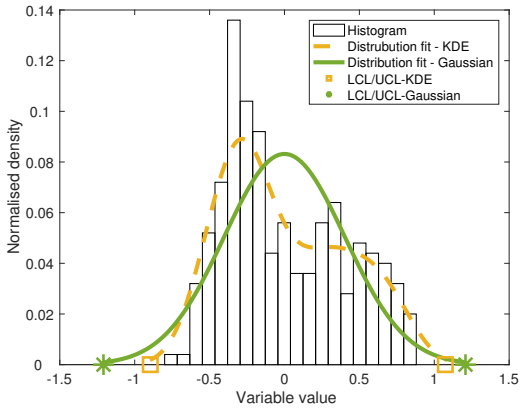
To measure the time lag between fault start and fault detection, Detection Time (DT) metric is used, and defined as below:

$$\mathrm{DT} = t_{\text{Fault detection}} - t_{\text{Fault start}}. \tag{6.1}$$

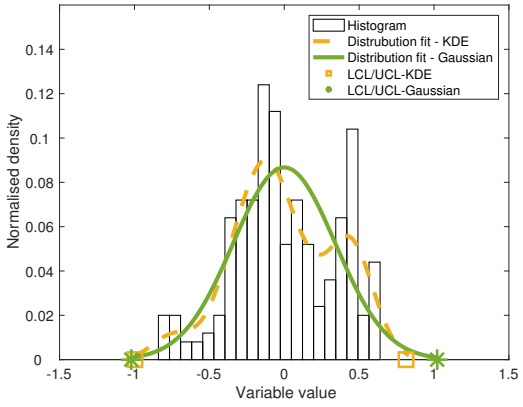
where $t_{\text{Fault detection}}$ indicates the time stamp from where a consecutive sequence of 30 data points are recognised as faulty. The objective of this definition is to avoid false alarms caused by random fluctuations. Other used performance metrics include Sensitivity, Specification and Accuracy, the calculation of which are given in Section 3.1 of Chapter 3.

Monitoring demonstration under blockage fault

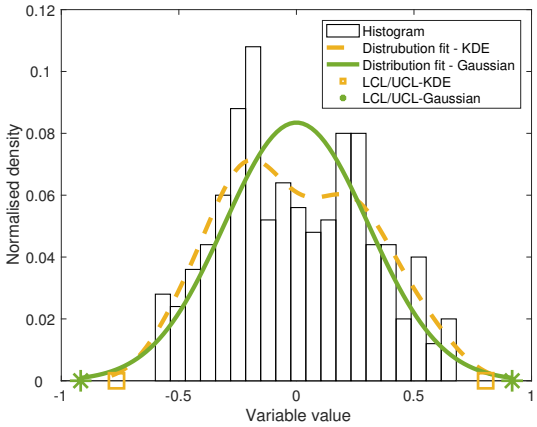
Fig. 6.4 illustrates the detection performance of the blockage fault in operating Mode 1. Before the fault was seeded in the facility, there were few PCs voting faulty operation. The fault started at sample 251 (A_1 in Fig. 6.4). In the course of blockage becoming severe (valve opening becoming smaller), the number of PCs supporting the decision of fault occurring in the system increased gradually. From sample 936 (A_2 in Fig. 6.4), there were 30 consecutive samples categorised as faulty, thus faulty operation was confirmed in the system. When the valve opening was decreased to 20° and 10°, all 5 PCs gave the decisions that the facility run at faulty operation.



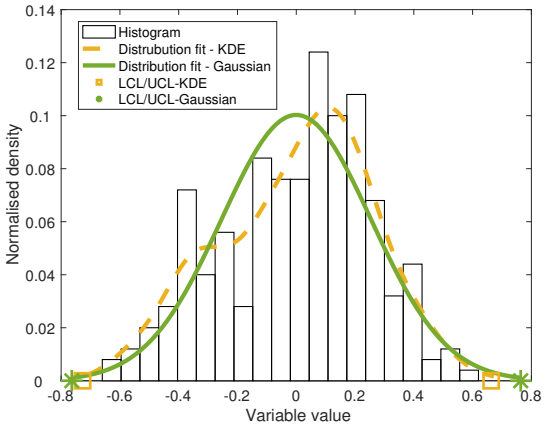
(a) Distribution plot of the 1st PC



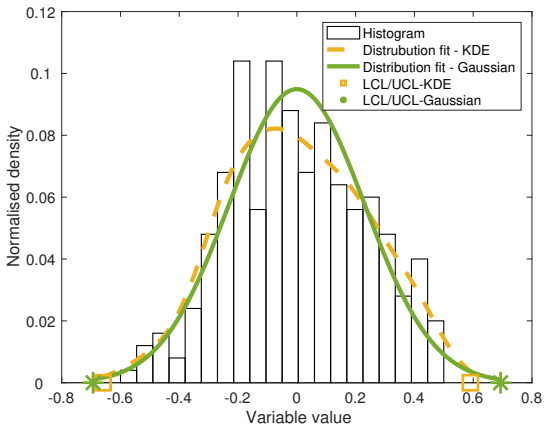
(b) Distribution plot of the 2nd PC



(c) Distribution plot of the 3rd PC

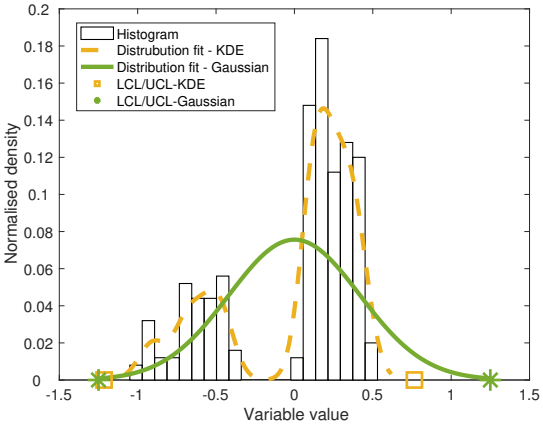


(d) Distribution plot of the 4th PC

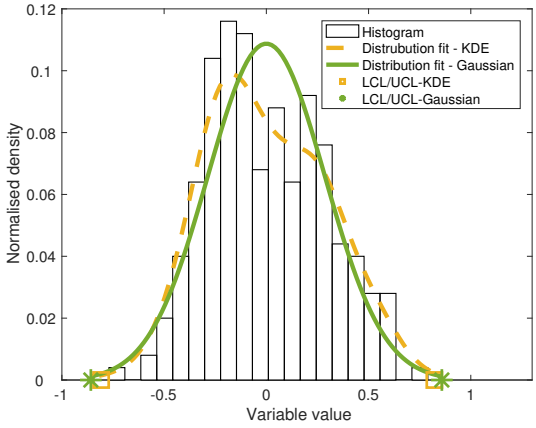


(e) Distribution plot of the 5th PC

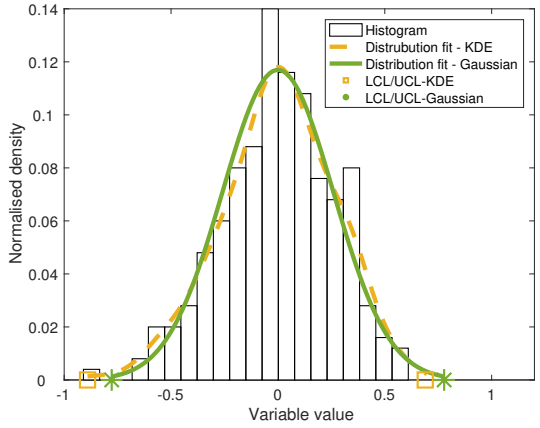
Figure 6.2: Distribution plots:5 PCs are extracted for explaining 70% variance in normal data of Mode 1.



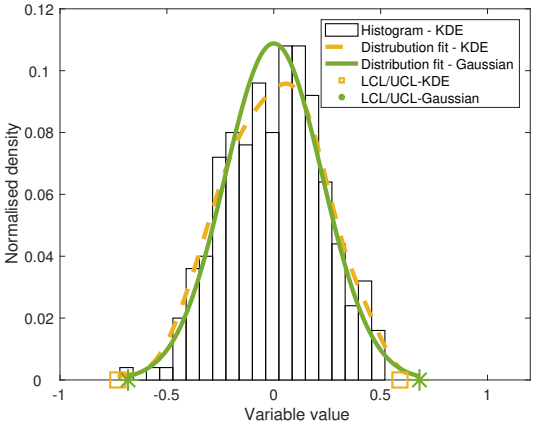
(a) Distribution plot of the 1st PC



(b) Distribution plot of the 2nd PC



(c) Distribution plot of the 3rd PC



(d) Distribution plot of the 4th PC

Figure 6.3: Distribution plots:4 PCs are extracted for explaining 70% variance in normal data of Mode 2.

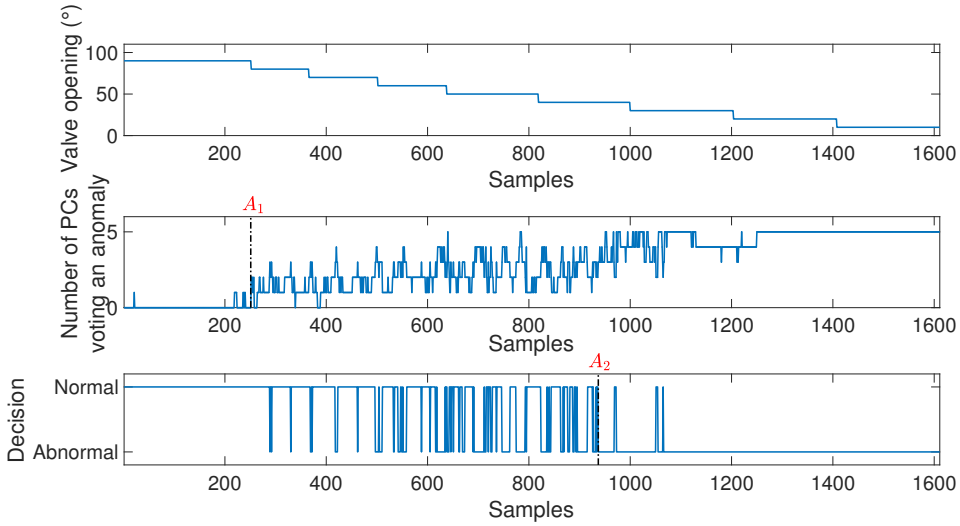


Figure 6.4: Blockage fault detection in operating Mode 1. A_1: fault start. A_2: fault detection.

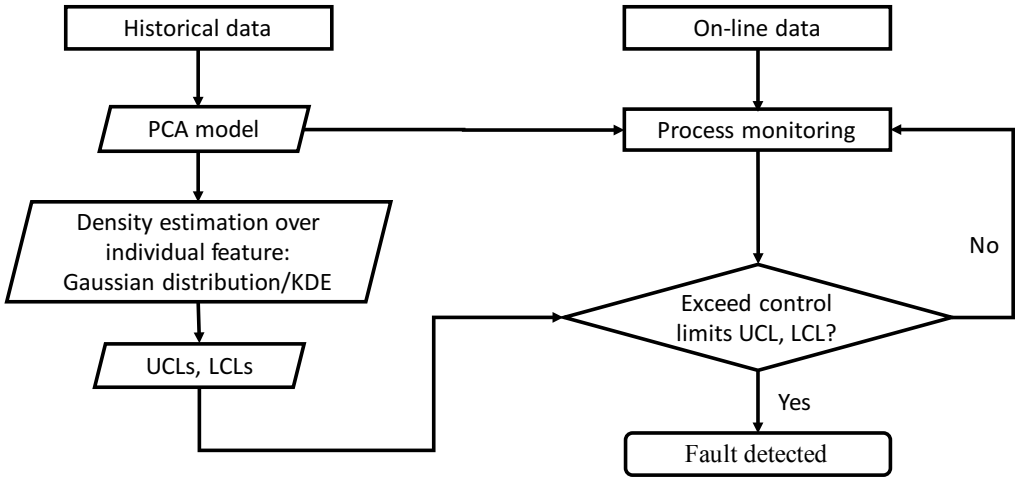


Figure 6.5: Gaussian/KDE-based fault detection with constant control limits: UCL and LCL are derived from the historical data. The values of UCL and LCL are not updated over time.

Comparison study

A comparison experiment is designed to prove the advantages of the proposed BaFFle algorithm. There are four algorithms involved in this experiment, Gaussian-based fault detection with constant monitoring limits, Gaussian-based BaFFle, KDE-based fault detection with constant monitoring limits and KDE-based BaFFle, respectively. Each algorithm was implemented for each type of fault under a specific operating mode. The workflow of Gaussian/KDE-based fault detection with constant monitoring limits is demonstrated in Fig. 6.5. The monitoring limits are fixed values.

Table 6.3 presents the fault detection comparison results using the aforementioned four monitoring algorithms. For reference, Algorithm 1 is Gaussian-based fault detection with constant monitoring limits; Algorithm 2 is Gaussian-based BaFFle; Algorithm 3 is KDE-based fault detection with constant monitoring limits; Algorithm 4 is KDE-based BaFFle. Generally, Table 6.3 consists of two parts. The first part presents the comparison results using blockage, leakage and diverted data of Mode 1 while the test data for the second part are from Mode 2. There are six metrics, namely DT, FP, FN, Sensitivity, Specification and Accuracy, to evaluate the performance of fault detection. The comparison is conducted from following two aspects:

- **Comparison between dynamic and constant monitoring limits**: the comparison is Algorithm 1 vs. 2 and Algorithm 3 vs. 4. DT metric shows that, except for leakage fault of Mode 2, the four algorithms all require a long time to detect the appearance of faults, no matter of the type of faults and operating mode. The delay of fault detection also can be reflected by the high missed alarms (FN). The delay likely results from a non-linear relationship between the fault severity and the associated valve adjustment. In the early stage, small valve adjustments cause minor variation in flow regimes, thus it is difficult to distinguish the incipient faults from normal operation. Nonetheless,

Table 6.3: Fault detection comparison results. The comparison experiment is conducted among four algorithms which are Gaussian-based fault detection with constant monitoring limits, Gaussian-based BaFFle, KDE-based fault detection with constant monitoring limits and KDE-based BaFFle. The test data include two operating modes. Under each mode, there are three types of faulty operation. The metrics for evaluating the monitoring performance are namely, DT, FP, FN, Sensitivity, Specification and Accuracy.

			Algorithm 1[a]	Algorithm 2[b]	Algorithm 3[c]	Algorithm 4[d]
Mode 1	Bloackage	DT	999	738	999	686
		FP	0	0	0	0
		FN	970	638	919	495
		Sensitivity	28.68	53.09	32.43	63.60
		Specification	100	100	100	100
		Accuracy	39.97	60.40	42.95	69.72
	Leakage	DT	615	452	478	210
		FP	0	0	0	0
		FN	562	394	518	327
		Sensitivity	38.91	57.17	43.70	64.46
		Specification	100	100	100	100
		Accuracy	52.01	66.35	55.76	72.08
	Diverted	DT	Fail	723	Fail	103
		FP	Fail	0	Fail	0
		FN	Fail	535	Fail	119
		Sensitivity	Fail	61.65	Fail	91.47
		Specification	Fail	100	Fail	100
		Accuracy	Fail	67.50	Fail	92.77
Mode 2	Blockage	DT	935	597	935	908
		FP	6	1	4	7
		FN	922	170	920	418
		Sensitivity	26.83	84.52	26.98	55.72
		Specification	98.54	99.74	99.03	98.12
		Accuracy	44.46	88.52	44.70	67.71
	Leakage	DT	8	14	8	14
		FP	6	1	4	7
		FN	537	27	525	34
		Sensitivity	68.80	98.38	69.49	97.98
		Specification	98.54	99.74	99.03	98.12
		Accuracy	74.53	98.64	75.19	98.00
	Diverted	DT	519	264	519	509
		FP	6	1	4	7
		FN	776	62	855	136
		Sensitivity	29.45	93.79	22.27	85.77
		Specification	98.54	99.74	99.03	98.12
		Accuracy	48.25	95.47	43.15	89.23

[a] Gaussian-based fault detection with constant monitoring limits.
[b] Gaussian-based BaFFle.
[c] KDE-based fault detection with constant monitoring limits.
[d] KDE-based BaFFle.

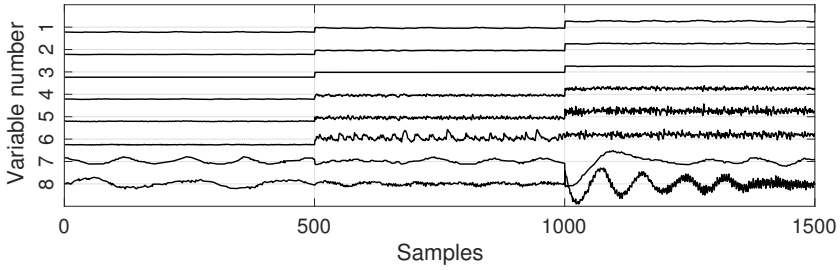
it can be seen that the algorithms with dynamic monitoring thresholds (Algorithm 2/4) achieve fewer missed alarms than the algorithms with fixed monitoring thresholds (Algorithm 1/3). Taking the blockage fault of Mode 1 as an example, compared with Algorithm 1, Algorithm 2 improved DT by 261 samples while Algorithm 4 shortened DT from 999 (Algorithm 3) to 686 samples. Similarly, FN dropped from 970 to 638 when using Algorithm 2 instead of Algorithm 1, and from 919 (Algorithm 3) to 495 (Algorithm 4). In addition, it should be noted that unsuitable constant monitoring thresholds would result in the failure to detect faults, for example, the diverted case in Mode 1. While, due to the adaptability of the monitoring limits, faults can be distinguished from normal operation.

The improvement in detection response of dynamic monitoring algorithms is owing to the use of the moving window (Section 3.5.2 of Chapter 3). The moving window can continuously learn the process variation by incorporating new measurements while increasing the accuracy of monitoring limits by discarding old measurements. Moreover, reduced DT and FN as well as low FP contribute to improving Sensitivity, Specification and Accuracy. As Table 6.3 shown, Sensitivity, Specification and Accuracy of dynamic algorithms (Algorithm 2 and 4) have been improved. For instance, Algorithm 1 only achieved 29.45% Sensitivity and 48.25% Accuracy in monitoring diverted flow fault, yet these two metrics were significantly improved by a 64.35% and 47.22% rate, respectively, using Algorithm 2. Similar improvements can also be observed in the comparison between Algorithm 3 and 4.

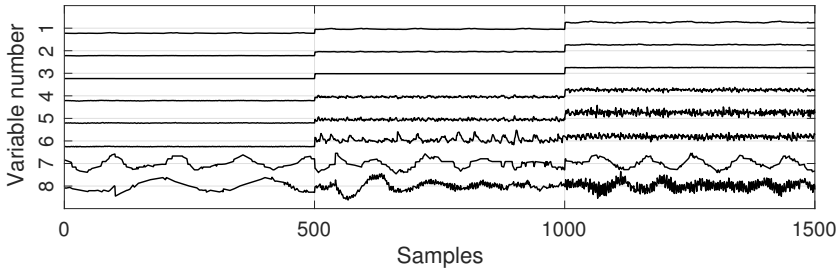
- **Comparison between Gaussian- and KDE-based BaFFle**: the comparison is Algorithm 2 vs. 4. It can be found that for Mode 1, the detection performance with Algorithm 4 outperforms the use of Algorithm 2; however, the monitoring results of Mode 2 show that Algorithm 2 is more suitable than Algorithm 4. These observations can be explained with the distribution plots in Fig. 6.2 and 6.3. In these figures, the LCL and UCL, given a confidence level of 99.7%, are marked. For all the remaining PCs of Mode 1, the normal operation intervals defined by the KDE-based LCL and UCL are narrower than the ones defined by Gaussian LCL and UCL. It means that KDE-based BaFFle is more suitable to be applied to Mode 1. In Fig. 6.3, it can be observed that the data points located at the edge of the dataset are detached from the majority. For example, in Fig. 6.3 (c), apart from the main cluster of bars in the range of -0.64 and 0.57, there is a bar occurring at -0.87. In the cases where there are separated bars, given a specific confidence level, the LCL and UCL derived from KDE might draw a wider range defining normal operation. This is because the KDE is prone to retain the information from all the individual points while under Gaussian distribution assumption, the data samples with low occurrence frequency might be ignored. As a result, in Mode 2, the control limits calculated according to KDE probability density will be less sensitive to the occurrence of abnormal operation, compared with the ones according to Gaussian probability density.

6.3. FKF for anomaly detection and mode identification

This thesis validates the monitoring performance of FKF using data from three normal operating mode (see Table 6.2) and one blockage fault under Mode 1 (introduced in Section 6.1.2). Eight variables are selected for monitoring, presented in Table 6.1. The workflow of the application of the FKF follows Section 5.5 of Chapter 5.



(a) Training data



(b) Validation data

Figure 6.6: Trend plots of training and validation data

6.3.1. Data preparation

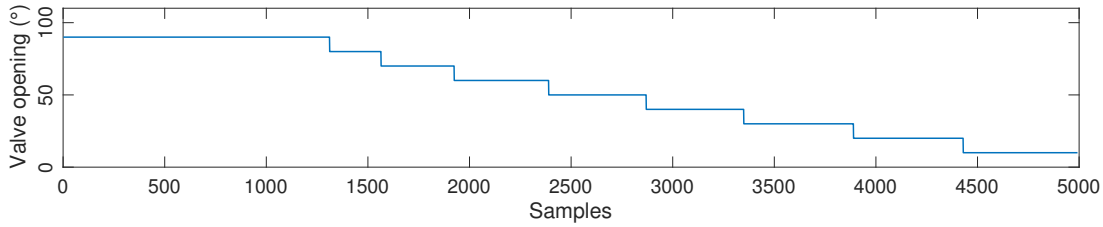
In the application of FKF to monitoring a process with multiple operating modes, explicit state-space models for individual modes are required. To obtain these models, a data-based method, Multivariate Autoregression State-space (MARSS), which was introduced in Chapter 5, is used. Since the data of normal operation consists of multiple operating modes, it is necessary to partition the healthy data with respect to operating modes. In this paper, the Dirichlet Process-Gaussian Mixture Models (DP-GMMs) (see Chapter 4) is employed to automate the data partition.

Given data labels, normal data are separated into training set, validation set and test set. The training set contains the first 500 samples $(1, 2, \ldots, 500)$ from Mode 1, Mode 2 and Mode 3, and the validation set contains the second 500 samples $(501, 502, \ldots, 1000)$ from Mode 1, Mode 2 and Mode 3. Additional samples from three normal modes, as well as all the samples from the air blockage fault, compose the test set. Fig. 6.6 plots the values for process variables against the time index to present the process trends in the training and validation sets.

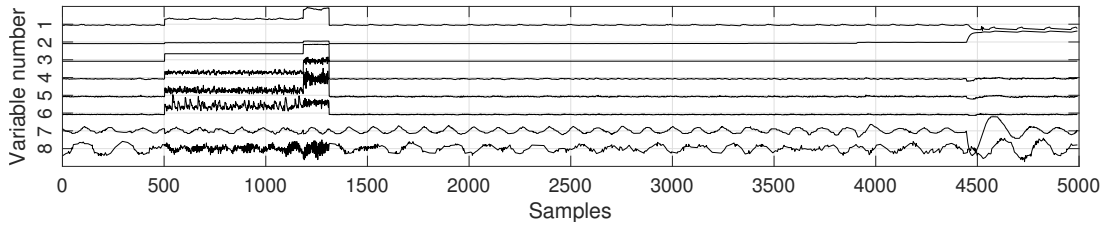
The vertical axis of Fig. 6.7(a) shows the degree of valve openings. Fig 6.7(b) highlights the trends of test data. The process begins with three different normal operating modes where the valve opening is kept maximum. Then after Mode 3, an air blockage fault is seeded by gradually closing the valve.

6.3.2. Results

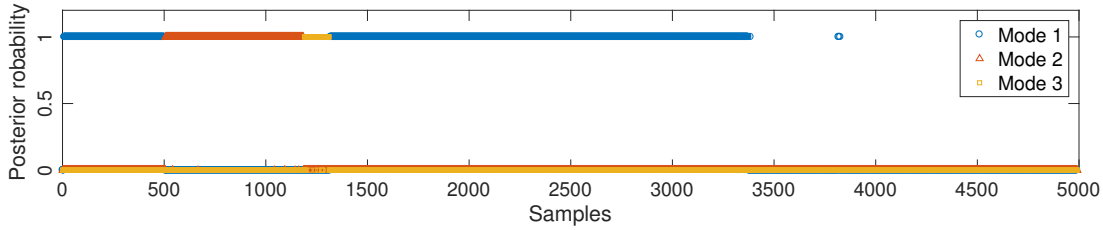
To differentiate anomalies from normal operation, the FKF introduces a unified monitoring indicator L_t (Section 5.4.2 of Chapter 5). Fig. 6.7(c) and (d) plots the posterior probability and monitoring indicator L_t against the time index, respectively. The process is considered to be running at faulty operating mode, when L_t falls below the monitoring threshold (the dashed line in Fig. 6.7(d)) for an extended period of time. When L_t remains at a normal level (above the monitoring threshold), the mode selection function



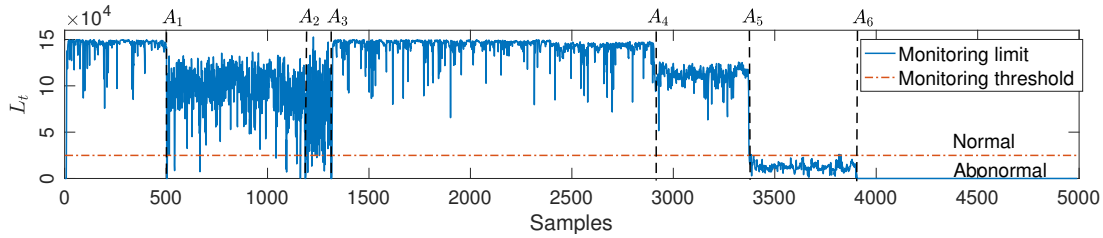
(a) The sequences of valve opening



(b) Trend plot of test data



(c) On-line mode identification: the current mode is the one with the maximal posterior probability.



(d) On-line anomaly detection: the monitoring threshold L_{LML} is in the red dashed line. The solid line is the monitoring indicator L_t. A_1 and A_2: the time stamp of mode switch. A_3: the time stamp of fault occurrence. A_4: the time stamp of L_t downward shifting, but not yet triggering the alarm of fault occurrence. A_5: the time stamp of fault detection. A_6: the time stamp of L_t approaching zero.

Figure 6.7: FKF for anomaly detection and mode identification on PRONTO benchmark dataset

works. The selection results are in posterior probabilistic form.

In this experiment, the initial conditions (at time 0) of the FKF algorithm are artificially set to zero, which are distinct from the set points of the normal operating modes. The zero initialisation causes a fast transient response in the L_t indicator. The transient responses also appear when mode switches take place, such as at sample 503 (A_1 in Fig. 6.7(d)) and 1185 (A_2 in Fig. 6.7(d)). This is because the FKF needs time to adapt to the current mode. During the adaption, L_t might travel across the monitoring threshold. After the adaption, L_t returns back to normal, and the current operating mode is the one with the maximal posterior probability. As seen in Fig. 6.7, both the monitoring indicators and the posterior probability correctly indicate the periods where the process was operating in Mode 1, Mode 2 and Mode 3. The monitoring indicators L_t varies over a wider range at Mode 2 and Mode 3, compared with at

Mode 1. This is because the process variables have larger variances (see Fig. 6.6(b)).

At sample 1312 (A_3 in Fig. 6.7(d)), the valve closure-caused fault is induced. However, till time stamp 3375 (A_5 in Fig. 6.7(d)), L_t continues to indicate the process running at normal mode, and the maximal posterior probability points to Mode 1. The inconsistency between the true process operation and the FKF results is likely due to the nonlinear relationship between the valve adjustment and the associated flow variations. The changes in flow regimes are gradually developing. Small adjustments in valve openings might cause minor and undetectable flow variations. Therefore, the incipient fault behaves similar to Mode 1.

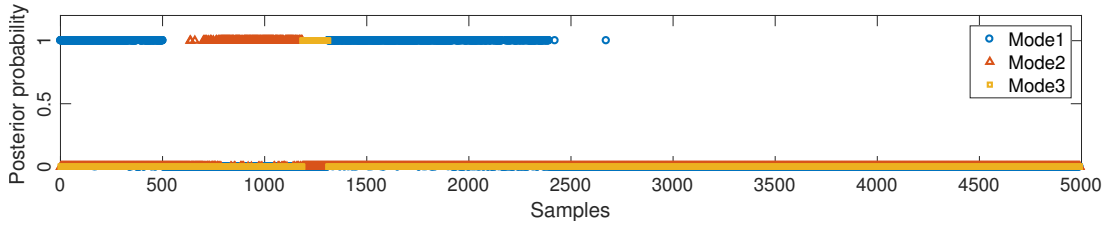
As the fault progresses, there is a downward shift in the monitoring indicator L_t between A_4 (sample 2908) and A_5 (sample 3375) in Fig. 6.7(d), but not yet triggering the alarm of fault occurrences. The valve opening at degree of 30 further worsen the fault severity, causing a sharp fall in L_t below L_{LML}. At this stage, the fault is detected, and all posterior probabilities of value zero indicates that the operating mode is unknown. From sample 3900 (A_6 in Fig. 6.7(d)), the monitoring indicator approaches to zero corresponding to a severe fault.

6.3.3. Comparison study

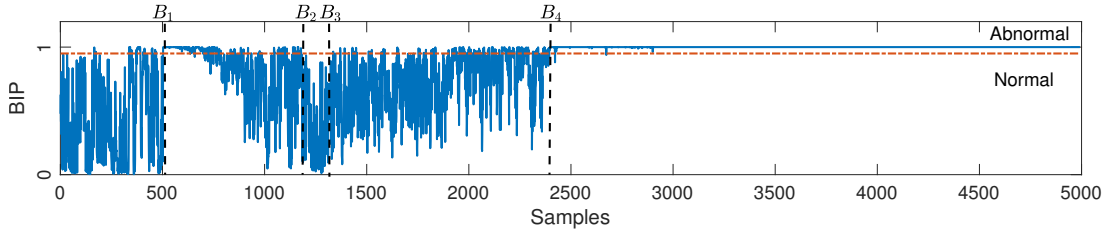
This section presents a comparison study among process monitoring methods, namely FGMM-BIP index and PCA-based T^2 and SPE and the FKF method. In FGMM-BIP index (Yu and Qin, 2008), multiple operating modes are described by Finite Gaussian Mixture Models (FGMM). BIP index is a unified monitoring indicator which is defined by integrating Bayesian statistical decisions and distance-based probability. The implementation of FGMM-BIP index followed the description by Yu and Qin (2008). The PCA-based T^2 and SPE are commonly used fault detection algorithm in process industries (Pilario et al., 2020). The computation of T^2 used the Matlab code from (Pilario, 2020) while SPE was implemented according to Ruiz-Cárcel et al. (2015). In this comparison experiment, the training and test data sets are the same as the ones in Section 6.3.1.

Fig. 6.8 and 6.9 give the monitoring results using FGMM-BIP index and PCA-based T^2 and SPE, respectively. Fig. 6.8(a) shows the mode identification results. The running mode is determined by the maximal posterior probability. The dashed line in Fig.6.8(b) is the BIP control limit set as 95% according to Yu and Qin (2008). When monitoring statistics exceed the BIP index, it means that there is faulty operation in the system. B_1 (sample 503) and B_2 (sample 1185) in Fig. 6.8(b) represent the action of mode switch. FGMM-BIP index mistakenly interprets the mode switch at B_1 (sample 503) as faulty operation, and persists this recognition until sample 658. This wrong recognition is also reflected in Fig. 6.8(a) as no mode is identified between sample 503 and 658. Starting from sample 1312, a blockage fault is seeded in the system by gradually closing the valve. The fault is detected by FGMM-BIP at sample 2394 (B_4 in Fig. 6.8(b)). This detection is earlier 981 samples than by the FKF method. The samples between B_3 and B_4 are identified as Mode 1 (see Fig. 6.8(a)). Since PCA-based T^2 and SPE can not perform mode identification, Fig. 6.9 only plots the fault detection results. The fault is detected at sample 2907 (E_1 in Fig 6.9) and 4457 (E_2 in Fig.6.9) by T^2 and SPE, respectively. It can be seen that T^2 outperform SPE indicator in this fault detection case study.

Table 6.4 gives the quantitative comparisons among the FKF, FGMM-BIP index and PCA-based T^2 and SPE. As the FKF and FGMM-BIP are designed on the basis of individual operating modes, both methods require the classification of training data. Furthermore, the monitoring models of the FKF and



(a) On-line mode identification: the current mode is the one with the maximal posterior probability.



(b) On-line anomaly detection: the monitoring indicator BIP (the solid line) above the monitoring threshold (the dashed line) indicates that the process is at abnormal operation whereas below the monitoring threshold means normal operation. B_1 and B_2: the time stamp of mode switch. B_3: the time stamp of fault occurrence. B_4: the time stamp of fault detection.

Figure 6.8: FGMM-BIP for anomaly detection and mode identification on PRONTO benchmark dataset

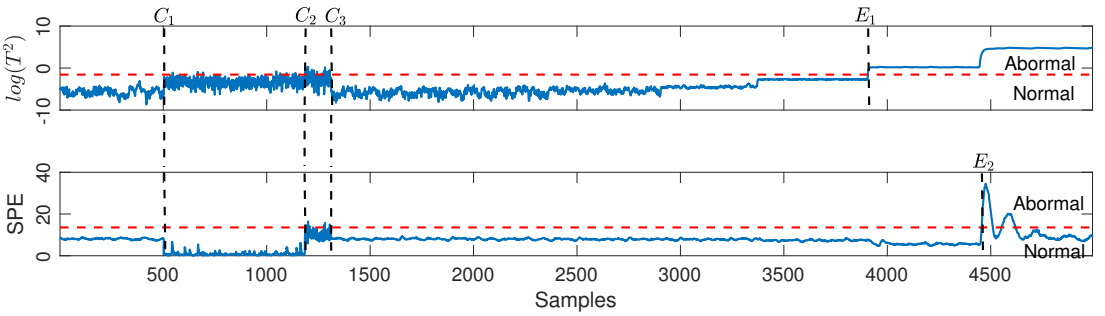


Figure 6.9: PCA-based T^2 and SPE for anomaly detection on PRONTO benchmark dataset: the monitoring indicators T^2 and SPE (the solid lines) above the monitoring thresholds (the dashed line) indicates that the process is at abnormal operation whereas below the monitoring threshold means normal operation. Mode switches occur at C_1 and C_2 while fault occurs at C_3. E_1 is when T^2 detects the fault. E_2 is when SPE detects the fault. It can been that in this case study, T^2 can identify the occurrence of fault earlier that SPE.

FGMM-BIP index are able to perform mode identification while the unified monitoring indicators can give a comprehensive understanding of the health state of process plants. In terms of the ability of mode identification, FGMM-BIP index shortens the delay of anomaly detection, compared with the other two algorithms. However, the FKF has less false alarm rate than the FGMM-BIP index and T^2 statistics, and outperforms FGMM-BIP index at task of identifying Mode 2 and 3.

6.4. Summary

This chapter has validated two monitoring algorithms, the BaFFle and FKF, using industrial-scale data. The results of the BaFFle in Section 6.2 has shown that for monitoring single mode processes,

Table 6.4: Quantitative comparison among the FKF, FGMM-BIP index and PCA-based T^2 and SPE. In this metric, the FKF outperforms the FGMM-BIP index and T^2. DT metric shows that the FGMM-BIP can detect the occurrence of faulty operation earlier than the FKF, T^2 and SPE. However, it should be noted that the FGMM-BIP index has a significant high false alarm rate. The operating mode is switched to Mode 2 and 3 at time stamp 503 and 1185, respectively. The FKF detects the mode switches a few time stamps later, whereas the delay of identifying the switch to Mode 2 by the FGMM-BIP index is 155 time instances.

	FKF	FGMM-BIP index	PCA
Classified training data?	✓	✓	✗
Mode identification	✓	✓	✗
Anomaly detection?	✓	✓	✓
Process models necessary?	✗[a]	✗	✗
Unified monitoring indicator?	✓	✓	✗
False alarm rate	2.67%	26.01%	T^2: 3.05% SPE: 0.53%
DT[b]	3375	2394	T^2: 3907 SPE: 4457
Time stamp of identifying Mode 2	507	658	N/A[c]
Time stamp of identifying Mode 3	1190	1185	N/A

[a] Incorporating an MARSS learning step in the proposed FKF removes the necessity of acquiring process model with first-principles;
[b] DT stands for Detection Time, the calculation of which is based on Eq. 6.1
[c] N/A means that the capability of mode identification is not applicable to the PCA-based T^2 and SPE.

- Gaussian- and KDE-based BaFFle approaches, featuring dynamic monitoring limits, are capable of reducing both false and missed alarms;

- The use of moving window not only benefits the adaptability by incorporating new measurements and removing old measurements, but also eliminates the need of large-scale data preparation and analysis. Users can employ this algorithm to monitoring single-mode processes without large-scale data as a-prior. The adaptability of the BaFFle algorithm enables control limits to be updated while the process is running;

- KDE is able to estimate the probability density distributions of non-Gaussian data.

- The fusion of multiple univariate control charts by a majority voting strategy can be successfully applied to monitoring multivariate processes. In addition, the fusion results are interpretable, and might indicate the severity of fault.

However, it should be noted that although KDE method can handle the probability density estimation of non-Gaussian data, the control limits based on KDE might be relaxed due to data samples that are detached from the majority. Such relaxed control limits will cause missed alarms. Thus, there still is space to improve the sensitivity of KDE-based monitoring limits.

Section 6.3 has proved the proposed FKF workflow (Section 5.5 of Chapter 5) in practical settings. Also through the experiments in Section 6.3, the strengths of the FKF for monitoring multimode processes can be concluded as follows.

- The proposed systematic framework can be applied to process condition monitoring when explicit mathematical process models are required. Process models in state-space form can be obtained using measurement data of individual operating modes.
- The FKF possesses the capabilities of both mode identification and anomaly detection.
- The FKF is capable of having a low false alarm rate while performing quick response to mode identification.
- The unified monitoring indicator eases fault detection by inspecting one index instead of a series of indices.

To conclude, the experiments presented in this chapter have confirmed the effectiveness of the BaFFle algorithm in fault detection, of the workflow of applying the FKF in practice, and of the FKF in mode identification and anomaly detection.

7. Conclusion

7.1. Summary of thesis

This thesis focused on systematically designing novel monitoring methods and algorithms for detecting abnormal behaviours from measurement data. This includes considering processes with a variety of complexity such that, with the newly designed approaches, monitoring systems would respond to the occurrence of faults and anomalies with more reliability and efficiency.

Chapter 2 firstly reviewed the key points of Process Condition Monitoring (PCM), such as the tasks of PCM, the categories of PCM and characteristics of PCM, to have a conceptual framework. Secondly, Chapter 2 investigated the development of techniques that are frequently involved in PCM, for example, process modelling. In addition, the processes with multiple operating modes were well studied with respect to their mathematical definition, data characteristics, data labelling, monitoring models as well as monitoring indices. Moreover, the decision-making methods were reviewed. Through the understanding of the concepts and techniques of PCM, Chapter 2 confirmed that for the sake of reliability and efficiency, the required monitoring algorithms should have acceptable levels of missed alarms and false alarms. Chapter 3 developed a single mode monitoring algorithm, named Binary Classified for Fault Detection (BaFFle). An adaptability design is involved in the BaFFle to enable it to work on any single mode system with only a small amount of historical data. For data management, Chapter 4 introduced the Dirichlet Process-Gaussian Mixture Models (DP-GMMs) for automating the data clustering without specifying the number of clusters in advance. The DP-GMMs-based clustering results were also discussed. In addition, a monitoring framework which incorporates cluster-based Multivariate Statistic Process Monitoring (MSPM) was proposed. Chapter 5 reviewed a model-based algorithm, the Field Kalman Filter (FKF). Moreover, the practical issues of the FKF, when applied to monitoring a process with multiple operating modes, were addressed. The training of the FKF monitoring model have been achieved using clustered historical data. Bayesian statistics have been used for differentiating various operation behaviours while a unified monitoring indicator has been designed to extend the FKF for anomaly detection. Chapter 6 gave the validation of the BaFFle and FKF algorithms with the PRONTO dataset.

7.2. Contributions and future work

As her main contributions in the thesis, the author considers:

- Development of the BaFFle algorithm with adaptability for fault detection. The adaptability of the BaFFle is apparent when there are only a small amount of measurements, particularly if these

measurements are insufficiently representative. In such cases, the BaFFle can adjust its monitoring thresholds;

- Development of a method for applying univariate control charts in monitoring multivariate processes. The Principal Component Analysis (PCA) technique is used for extracting uncorrelated features from multivariate data so as to have multiple unbiased univariate control charts. To fuse the decision across individual control charts, a majority voting strategy is adopted;

- Investigation of the approaches of probability density estimation. Considering the non-Gaussian distributions, the Kernel Density Estimation (KDE), a nonparametric method, is employed;

- Creation of a mechanism of warning and detection, where the warning indicator is used for adjusting control limits while the detection indicator is used for determining whether a process is healthy or not;

- Application of the BaFFle algorithm for fault detection. The fault detection results have presented the effectiveness and interpretability of the BaFFle algorithm. The effectiveness of the dynamic control limits in reducing false and missed alarms have been demonstrated in the comparison with constant control limits;

- Investigation of the DP-GMMs in the use of data clustering. Since the number of operating modes might be unknown, the DP-GMMs algorithm is selected for clustering recorded measurements corresponding to the operating modes. The quality of clustering is dependent on the initial values of the parameters of DP-GMMs. A discussion on how to properly initialise the parameter is presented;

- Development of a monitoring framework. In the collaboration work (Tan et al., 2019, 2020), a monitoring framework was proposed, in which the monitoring model is trained off-line by the DP-GMMs clustering and a kernel-based MSPM algorithm. Another function of the DP-GMMs in this framework is to identify if the incoming data are from new modes. The identified data of new modes will be incorporated in the training of monitoring model. In such a form, missed alarm caused by new modes might be significantly reduced;

- Development of the FKF algorithm for fault isolation. The FKF is an example of the integration of process models and Bayes' theorem, thus has the advantage of differentiating various faults or operating modes deterministically or stochastically. Additionally, as Bayesian decision statistics are traceable and interpretable, it would be convenient for process inspectors to spot anomalies or identify the fault type;

- Investigation of process modelling in the absence of prior knowledge of process industries. With large-scale historical data, data-based process modelling methods can be used. As the FKF monitoring model for a multimode process is a set of state-space models, the MARSS approach is employed in this thesis;

- Development of the FKF algorithm for anomaly detection. The anomaly detection is an extension of the FKF, from inferring within known process conditions to recognition of new process operation. This is achieved by designing a novel unified monitoring indicator;

- Development of a workflow for systematically using the FKF in the industrial applications of anomaly detection and mode identification. Practical issues, such as the difficulties in data labelling, first-principle modelling and interpretability of the monitoring results, have been taken into consideration;

- Application of the proposed workflow for monitoring a multimode process. The effectiveness of proposed workflow has been proved using the PRONTO benchmark data. The experiment results have shown that the use of Bayesian statistic decisions can differentiate various operating modes, and that anomalies can be detected with the unified monitoring indicator. In addition, the comparison experiment has highlighted that the FKF monitoring model can achieve a relative low false alarm rate, and outperforms the two selected approaches at mode identification.

Based on the research in this thesis, a few directions might be worthy of further exploration.

- The dynamics in data has not been considered in the BaFFle algorithm. The extraction of dynamic features might give more insights into the variations in the data. Therefore, rather than the PCA used in this thesis, dynamic feature extraction methods, such as dynamic PCA and Canonical Variate Analysis, might be of interest in future work.

- Development of the FKF for monitoring nonlinear operating modes. In the proposed FKF, data from normal operation are described with linear models. Nevertheless, nonlinearity might also appear in the data, thereby linear models would be insufficient to monitoring nonlinear processes. Thus, it is worth to work on the extension of the FKF to monitor nonlinear processes.

- Development of the FKF for predicting the health state of the monitored system. The design of the FKF takes the dynamics of processes into account. The knowledge of dynamics might be further utilised for prognosis of potential harmful behaviour in the process so as to take actions in advance. Furthermore, the prognosis result can be incorporated into the FKF-based fault detection in order to increase the accuracy of detection.

A. Appendix

A.1. Definitions of terms in Process Condition Monitoring (PCM)

- Normal operation refers to the system performance under a desired function (Tidriri et al., 2016). In PCM, “normality”, “normal condition/operating mode”, “healthy state”, “standard condition/-operation”, and “fault-free” all mean that a process plant is running at normal operation.

- Failure refers to, under specified operating conditions, the permanent inability of a system to implement a required function (Isermann, 2006). Venkatasubramanian et al. (2003) categorised failures into three classes according to the sources of a failure: parameter-caused failure arises when a disturbance from a single or multiple external variables exceeds acceptable normal range; structure-caused failure arises in the structure itself; sensor- and actuator- caused failure arises due to errors, such as a constant bias and an out-of-range malfunction, occurring with sensors and actuators.

- Malfunction is defined by Isermann (2006) as an intermittent irregularity in the fulfilment of desired function of a system.

- Fault is a non-permitted deviation of at least one characteristic property or feature of the system from the normal operation (Isermann, 2005, 2006). Moreover, a fault can be classified as hard faults and soft faults depending on whether it is predicable or not (Martin, 1994). The occurrence of a hard fault is instantaneous and unpredictable, depicted in Fig. A.1. A soft fault is progressive over time, shown in Fig. A.1 as a gradual time-continuous trajectory. A fault, if left unresolved, will develop to a failure or a malfunction.

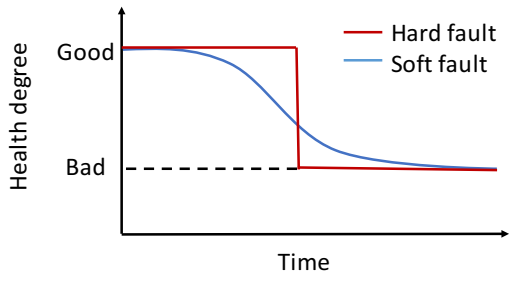


Figure A.1: Hard and soft faults (Martin, 1994)

- Degradation, which is considered as a normal, but detrimental, physical change (Gonzalez et al., 1997), is a representative soft fault. Remaining undetected, a soft fault may lead to severe defects

occurring in mechanical components (e.g. engines and motors) or in physical structures (e.g. cracks in pipelines and leakage in valves), eventually leading to critical failures.

- Anomaly refers to the cases where the system performance is not in conformity with normal operation. An anomaly might be indicative of, for example, a fault. Some literature, for example, (Cateni et al., 2008), also uses "outliers" instead of "anomaly". Without additional domain knowledge, it is typically not possible to distinguish between a fault and a change. Domain knowledge can be obtained from experienced operators or domain experts.

A.2. List of publications

1. J. Baranowski, P. Bania, I. Prasad, and T. Cong. Bayesian fault detection and isolation using field Kalman filter. *EURASIP Journal on Advances in Signal Processing*, 2017 (1):79, 2017.

2. T. Cong and J. Baranowski. Binary classifier for fault detection based on gaussian model and PCA. *IFAC-PapaerOnLine*, 51(24):1317-1323, 2018a.

3. T. Cong and J. Baranowski. Binary classifier for fault detection based on KDE and PCA. In *2018 23rd International Conference on Methods & Models in Automation & Robotics (MMAR)*, pages 821-825. IEEE, 2018b.

4. R. Tan, T. Cong, N. F. Thornhill, J. R. Ottewill, and J. Baranowski. Statistical monitoring of processes with multiple operating mode. *IFAC-PapersOnLine*, 52(1):635-642, 2019.

5. R. Tan, T. Cong, J. R. Ottewill, J. Baranowski, and N. Thornhill. An on-line framework for monitoring nonlinear processes with multiple operating modes. *Journal of Process Control*, 89:119-130, 2020.

6. T. Cong, R. Tan, J. R. Ottewill, N. Thornhill, and J. Baranowski. Anomaly detection and mode identification in multimode process using the field Kalman filter. *IEEE Transactions on Control Systems Technology*, pages 1-14, 2020. doi: 10.1109/TCST.2020.3027809.

Bibliography

R. Adhikari and R. K. Agrawal. *An introductory study on time series modeling and forecasting*. 2013.

A. Adhitya, S. F. Cheng, Z. Lee, and R. Srinivasan. Quantifying the effectiveness of an alarm management system through human factors studies. *Computers & chemical engineering*, 67:1–12, 2014.

S. Agatonovic-Kustrin and R. Beresford. Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research. *Journal of pharmaceutical and biomedical analysis*, 22(5):717–727, 2000.

H. Akaike. Fitting autoregressive models for prediction. *Annals of the institute of Statistical Mathematics*, 21(1):243–247, 1969.

B. M. Åkesson, J. B. Jørgensen, N. K. Poulsen, and S. B. Jørgensen. A generalized autocovariance least-squares method for Kalman filter tuning. *Journal of Process control*, 18(7-8):769–779, 2008.

S. Akhlaghi, N. Zhou, and Z. Huang. Adaptive adjustment of noise covariance in Kalman filter for dynamic state estimation. In *2017 IEEE power & energy society general meeting*, pages 1–5. IEEE, 2017.

B. Al-Najjar and I. Alsyouf. Enhancing a company's profitability and competitiveness using integrated vibration-based maintenance: A case study. *European journal of operational research*, 157(3):643–657, 2004.

D. J. Aldous. Exchangeability and related topics. In *École d'Été de Probabilités de Saint-Flour XIII-1983*, pages 1–198. Springer, 1985.

J. Alkahe, Y. Oshman, and O. Rand. Adaptive estimation methodology for helicopter blade structural damage detection. *Journal of guidance, control, and dynamics*, 25(6):1049–1057, 2002.

A. Alkaya and İ. Eker. Variance sensitive adaptive threshold-based PCA method for fault detection with experimental application. *ISA transactions*, 50(2):287–302, 2011.

F. B. Alt and N. D. Smith. *Multivariate process control*, volume 7. Elsevier, 1988.

I. Alvarez, J. Niemi, and M. Simpson. Bayesian inference for a covariance matrix. *arXiv preprint arXiv:1408.4050*, 2014.

M. Amini and S. Chang. A review of machine learning approaches for high dimensional process monitoring. In *Proceedings of the 2018 Industrial and Systems Engineering Research Conference, Orlando, FL*, 2018.

B. D. Anderson and J. B. Moore. *Optimal filtering*. Prentice Hall, 1979.

T. Anderson. An introduction to multivariate statistical analysis (wiley series in probability and statistics). *July 11*, 2003.

C. E. Antoniak. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The annals of statistics*, pages 1152–1174, 1974.

K. Baker. Singular value decomposition tutorial. *The Ohio State University*, 24, 2005.

P. Bania and J. Baranowski. Field Kalman filter and its approximation. In *2016 IEEE 55th Conference on Decision and Control (CDC)*, pages 2875–2880, Dec 2016. doi: 10.1109/CDC.2016.7798697.

Y. Bar-Shalom. Optimal simultaneous state estimation and parameter identification in linear discrete-time systems. *IEEE Transactions on Automatic Control*, 17(3):308–319, June 1972. doi: 10.1109/TAC.1972.1100005.

J. Baranowski, P. Bania, I. Prasad, and T. Cong. Bayesian fault detection and isolation using field Kalman filter. *EURASIP Journal on Advances in Signal Processing*, 2017(1):79, 2017.

A. Baratloo, M. Hosseini, A. Negida, and G. El Ashal. Part 1: simple definition and calculation of accuracy, sensitivity and specificity. 2015.

J. Barnard, R. McCulloch, and X.-L. Meng. Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Statistica Sinica*, pages 1281–1311, 2000.

A. Bedoui. *Comparison of Bayesian nonparametric density estimation methods*. PhD thesis, University of Texas at El Paso, 2013.

J. Bernoulli. *Ars conjectandi*. Impensis Thurnisiorum, fratrum, 1713.

S. Bersimis, J. Panaretos, and S. Psarakis. Multivariate statistical process control charts and the problem of interpretation: a short overview and some applications in industry. In *Proceedings of the 7th Hellenic European Conference on Computer Mathematics and its Applications, Athens Greece*, 2005.

S. Bersimis, S. Psarakis, and J. Panaretos. Multivariate statistical process control charts: an overview. *Quality and Reliability engineering international*, 23(5):517–543, 2007.

N. Bhutani, G. Rangaiah, and A. Ray. First-principles, data-based, and hybrid modeling and optimization of an industrial hydrocracking unit. *Industrial & engineering chemistry research*, 45(23):7807–7816, 2006.

G. Bialic, M. Zmarzły, and R. Stanisławski. Design of the scales for the power boiler fuel feeding system based on the process identification. *IFAC Proceedings Volumes*, 42(13):292–295, 2009.

C. M. Bishop. *Pattern recognition and machine learning*. springer, 2006.

G. Bishop, G. Welch, et al. An introduction to the Kalman filter. *Proc of SIGGRAPH, Course 8*, (27599-23175):41, 2001.

M. Blanke, M. Kinnaert, J. Lunze, M. Staroswiecki, and J. Schröder. *Diagnosis and fault-tolerant control*, volume 2. Springer, 2006.

G. Bonciolini, E. Boujo, and N. Noiray. Output-only parameter identification of a colored-noise-driven van-der-pol oscillator: Thermoacoustic instabilities as an example. *Physical Review E*, 95(6):062217, 2017.

M. Borowski, S. Siebig, C. Wrede, and M. Imhoff. Reducing false alarms of intensive care online-monitoring systems: an evaluation of two signal extraction algorithms. *Computational and mathematical methods in medicine*, 2011, 2011.

A. W. Bowman and A. Azzalini. *Applied smoothing techniques for data analysis: the kernel approach with S-Plus illustrations*, volume 18. OUP Oxford, 1997.

J. Calado, J. Korbicz, K. Patan, R. J. Patton, and J. S. Da Costa. Soft computing approaches to fault diagnosis for dynamic systems. *European Journal of Control*, 7(2-3):248–286, 2001.

I. T. Cameron and G. Ingram. A survey of industrial process modelling across the product and process lifecycle. *Computers & Chemical Engineering*, 32(3):420–438, 2008.

G. A. Carpenter and S. Grossberg. A massively parallel architecture for a self-organizing neural pattern recognition machine. *Computer vision, graphics, and image processing*, 37(1):54–115, 1987.

G. Casella and E. I. George. Explaining the Gibbs sampler. *The American Statistician*, 46(3):167–174, 1992.

S. Cateni, V. Colla, M. Vannucci, J. Aramburo, and A. R. Trevino. Outlier detection methods for industrial applications. *Advances in Robotics, Automation and Control*, pages 265–282, 2008.

M. Chaabane, A. Ben Hamida, M. Mansouri, H. Nounou, and M. Nounou. Improved Shewhart chart for damage detection of structural health monitoring systems. In *2018 4th International Conference on Advanced Technologies for Signal and Image Processing (ATSIP)*, pages 1–5, March 2018. doi: 10.1109/ATSIP.2018.8364484.

Y. Chang, W. Yang, and D. Zhao. Energy efficiency and emission testing for connected and automated vehicles using real-world driving data. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 2058–2063, Nov 2018. doi: 10.1109/ITSC.2018.8569806.

Y.-H. Chang, W. Wang, E. A. Patterson, J.-Y. Chang, and J. E. Mottershead. Output-only full-field modal testing. *Procedia engineering*, 199:423–428, 2017.

M. Chegini, J. Bernard, P. Berger, A. Sourin, K. Andrews, and T. Schreck. Interactive labelling of a multivariate dataset for supervised machine learning using linked visualisations, clustering, and active learning. *Visual Informatics*, 3(1):9–17, 2019.

J. Chen and R. J. Patton. *Robust model-based fault diagnosis for dynamic systems*, volume 3. Springer Science & Business Media, 2012.

Y.-C. Chen. A tutorial on kernel density estimation and recent advances. *Biostatistics & Epidemiology*, 1(1):161–187, 2017.

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

T. Cong and J. Baranowski. Binary classifier for fault detection based on gaussian model and PCA. *IFAC-PapersOnLine*, 51(24):1317–1323, 2018a.

T. Cong and J. Baranowski. Binary classifier for fault detection based on KDE and PCA. In *2018 23rd International Conference on Methods & Models in Automation & Robotics (MMAR)*, pages 821–825. IEEE, 2018b.

T. Cong, R. Tan, J. R. Ottewill, N. F. Thornhill, and J. Baranowski. Anomaly detection and mode identification in multimode processes using the field Kalman filter. *IEEE Transactions on Control Systems Technology*, pages 1–14, 2020. doi: 10.1109/TCST.2020.3027809.

P. Czop, G. Kost, D. Sławik, and G. Wszołek. Formulation and identification of first-principle data-driven models. *Journal of Achievements in materials and manufacturing Engineering*, 44(2):179–186, 2011.

O. Dahunsi, J. Pedro, and O. Nyandoro. System identification and neural network based PID control of servo-hydraulic vehicle suspension system. *SAIEE Africa Research Journal*, 101(3):93–105, 2010.

S. Dasani, S. L. Shah, T. Chen, J. Funnell, and R. W. Pollard. Monitoring safety of process operations using industrial workflows. *IFAC-PapersOnLine*, 48(8):451–456, 2015.

P. De Marco and C. C. Nóbrega. Evaluating collinearity effects on species distribution models: An approach based on virtual species simulation. *PLoS One*, 13(9):e0202403, 2018.

C. de Prada, C. C. Pantelides, and J. L. Pitarch. Special issue on "process modelling and simulation", 2019.

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)*, 39(1):1–22, 1977.

X. Deng, N. Zhong, and L. Wang. Nonlinear multimode industrial process fault detection using modified kernel principal component analysis. *IEEE Access*, 5:23121–23132, 2017.

F. Ding, P. X. Liu, and G. Liu. Multiinnovation least-squares identification for system modeling. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 40(3):767–778, June 2010. doi: 10.1109/TSMCB.2009.2028871.

S. Ding. Data-driven design of monitoring and diagnosis systems for dynamic processes: A review of subspace technique based schemes and some recent results. *Journal of Process Control*, 24(2):431–449, 2014.

S. Ding, P. Zhang, A. Naik, E. Ding, and B. Huang. Subspace method aided data-driven design of fault detection and isolation systems. *Journal of process control*, 19(9):1496–1510, 2009.

S. X. Ding. *Model-based fault diagnosis techniques: design schemes, algorithms, and tools*. Springer Science & Business Media, 2008.

J. Dos Reis and C. O. Costa. Review report on fault detection in sensor networks, 2013. URL http://repositorio.lnec.pt:8080/bitstream/123456789/1005804/2/Rel%20418_13%20dspace.pdf. Accessed: 2020-1-21.

N. B. Erichson, S. Voronin, S. L. Brunton, and J. N. Kutz. Randomized matrix decompositions using R. *arXiv preprint arXiv:1608.02148*, 2016.

M. Escobar. *Estimating the means of several normal popilations by nonparametric estimation of the distribution of the means*. PhD thesis, Department of Statistics, Yale University, 1988.

M. D. Escobar. Estimating normal means with a Dirichlet process prior. *Journal of the American Statistical Association*, 89(425):268–277, 1994.

M. D. Escobar and M. West. Bayesian density estimation and inference using mixtures. *Journal of the american statistical association*, 90(430):577–588, 1995.

W. Favoreel, B. De Moor, and P. Van Overschee. Subspace state space system identification for industrial processes. *Journal of process control*, 10(2-3):149–155, 2000.

J. Figwer. Continuous-time input-output linear dynamic system identification using sampled data. In *2015 20th International Conference on Methods and Models in Automation and Robotics (MMAR)*, pages 712–717. IEEE, 2015.

D. Fink. A compendium of conjugate priors. Technical report, Montana State Univeristy, 1997.

C. Forbes, M. Evans, N. Hastings, and B. Peacock. *Statistical distributions*. John Wiley & Sons, 2011.

S. Formentin and S. Bittanti. An insight into noise covariance estimation for Kalman filter design. *IFAC Proceedings Volumes*, 47(3):2358–2363, 2014.

C. Fraley and A. E. Raftery. Bayesian regularization for normal mixture estimation and model-based clustering. *Journal of classification*, 24(2):155–181, 2007.

P. M. Frank and X. Ding. Survey of robust residual generation and evaluation methods in observer-based fault detection systems. *Journal of process control*, 7(6):403–424, 1997.

B. Frigyik, A. Kapila, and M. Gupta. Introduction to the Dirichlet distribution and related processes. Master's thesis, University of Washington, Seattle, Washington, the United States, 2010.

D. Gamerman and H. F. Lopes. *Markov chain Monte Carlo: stochastic simulation for Bayesian inference*. Chapman and Hall/CRC, 2006.

X. Gao, D. You, and S. Katayama. Seam tracking monitoring based on adaptive Kalman filter embedded elman neural network during high-power fiber laser welding. *IEEE Transactions on Industrial Electronics*, 59(11):4315–4325, Nov 2012. doi: 10.1109/TIE.2012.2193854.

D. Garcia-Alvarez, M. Fuente, and G. Sainz. Fault detection and isolation in transient states using principal component analysis. *Journal of Process Control*, 22(3):551–563, 2012.

Z. Ge. Process data analytics via probabilistic latent variable models: A tutorial review. *Industrial & Engineering Chemistry Research*, 57(38):12646–12661, 2018.

Z. Ge and Z. Song. Mixture bayesian regularization method of PPCA for multimode process monitoring. *AIChE journal*, 56(11):2838–2849, 2010a.

Z. Ge and Z. Song. Maximum-likelihood mixture factor analysis model and its application for process monitoring. *Chemometrics and Intelligent Laboratory Systems*, 102(1):53–61, 2010b.

Z. Ge and Z. Song. *Multivariate statistical process control: Process monitoring methods and applications*. Springer Science & Business Media, 2012.

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.

Z. Ge, Z. Song, S. X. Ding, and B. Huang. Data mining and analytics in the process industry: The role of machine learning. *IEEE Access*, 5:20590–20616, 2017. doi: 10.1109/ACCESS.2017.2756872.

A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. *Bayesian data analysis*. Chapman and Hall/CRC, 2013.

S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6(6):721–741, Nov 1984. doi: 10.1109/TPAMI.1984.4767596.

J. Gertler. *Fault detection and diagnosis in engineering systems*. Routledge, 1998.

K. Ghosh, Y. S. Ng, and R. Srinivasan. Evaluation of decision fusion strategies for effective collaboration among heterogeneous fault diagnostic methods. *Computers & chemical engineering*, 35(2):342–355, 2011.

W. R. Gilks, S. Richardson, and D. Spiegelhalter. *Markov chain Monte Carlo in practice*. Chapman and Hall/CRC, 1995.

O. Gonzalez, H. Shrikumar, J. A. Stankovic, and K. Ramamritham. Adaptive fault tolerance and graceful degradation under dynamic hard real-time scheduling. In *Proceedings Real-Time Systems Symposium*, pages 79–89, Dec 1997. doi: 10.1109/REAL.1997.641271.

D. Görür and C. E. Rasmussen. Dirichlet process gaussian mixture models: Choice of the base distribution. *Journal of Computer Science and Technology*, 25(4):653–664, 2010.

D. M. Green, J. A. Swets, et al. *Signal detection theory and psychophysics*. Wiley New York, 1966.

C. Grinstead and J. Snell. *Introduction to probability*. American Mathematical Society, 1998.

J. A. Gubner. *Probability and random processes for electrical and computer engineers*. Cambridge University Press, 2006.

S. Guo, P. Rösch, J. Popp, and T. Bocklitz. Modified PCA and PLS: Towards a better classification in Raman spectroscopy–based biological applications. *Journal of Chemometrics*, 34(4):e3202, 2020.

D. M. Hawkins. The detection of errors in multivariate data using principal components. *Journal of the American Statistical Association*, 69(346):340–344, 1974.

E. E. Holmes, E. J. Ward, and K. Wills. MARSS: Multivariate autoregressive state-space models for analyzing time-series data. *R journal*, 4(1):11–19, 2012.

X. Hong, R. J. Mitchell, S. Chen, C. J. Harris, K. Li, and G. W. Irwin. Model selection approaches for non-linear system identification: a review. *International journal of systems science*, 39(10):925–946, 2008.

O. Hryniewicz and K. Kaczmarek-Majer. Monitoring of non-stationary health-recovery processes with control charts. *Int. J. Adv. Life Sci*, 2018.

S. Huang, K. K. Tan, and T. H. Lee. Fault diagnosis and fault-tolerant control in linear drives using the Kalman filter. *IEEE Transactions on Industrial Electronics*, 59(11):4285–4292, Nov 2012. doi: 10.1109/TIE.2012.2185011.

D.-H. Hwang and C. Han. Real-time monitoring for a process with multiple operating modes. *Control Engineering Practice*, 7(7):891–902, 1999.

A. Isaksson, F. Gustafsson, and N. Bergman. Pruning versus merging in Kalman filter banks for manoeuvre tracking. *IEEE Transactions on Aerospace and Electronic Systems, AES-33*, 1999.

R. Isermann. Model-based fault-detection and diagnosis–status and applications. *Annual Reviews in control*, 29(1):71–85, 2005.

R. Isermann. *Fault-diagnosis systems: an introduction from fault detection to fault tolerance*. Springer Science & Business Media, 2006.

A. K. Jardine, D. Lin, and D. Banjevic. A review on machinery diagnostics and prognostics implementing condition-based maintenance. *Mechanical systems and signal processing*, 20(7):1483–1510, 2006.

B. Jiang, D. Huang, X. Zhu, F. Yang, and R. D. Braatz. Canonical variate analysis-based contributions for fault identification. *Journal of Process Control*, 26:17–25, 2015.

Q. Jiang and X. Yan. Nonlinear plant-wide process monitoring using MI-spectral clustering and bayesian inference-based multiblock kpca. *Journal of Process Control*, 32:38–50, 2015.

J. Jin and J. Shi. State space modeling of sheet metal assembly for dimensional control. *Journal of manufacturing science and engineering*, 121(4):756–762, 1999.

S. Joe Qin. Statistical process monitoring: basics and beyond. *Journal of Chemometrics: A Journal of the Chemometrics Society*, 17(8-9):480–502, 2003.

I. Jolliffe. *Principal component analysis*. Springer, 2011.

I. T. Jolliffe and J. Cadima. Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374 (2065):20150202, 2016.

M. Jordan. The exponential family: conjugate priors, 2010. URL https://people.eecs.berkeley.edu/~jordan/courses/260-spring10/other-readings/chapter9.pdf. Accessed: 2020-10-21.

T. Jung. The basic distribution probability tutorial for deep learning researchers., 2019. URL https://pythonawesome.com/the-basic-distribution-probability-tutorial-for-deep-learning-researchers/. Accessed: 2019-11-13.

R. E. Kalman. A new approach to linear filtering and prediction problems. *Journal of basic Engineering*, 82(1):35–45, 1960.

P. A. Kline. *Fault detection and isolation for integrated navigation systems using the Global Positioning System*. PhD thesis, Ohio University, 1991.

T. Kobayashi and D. L. Simon. Evaluation of an enhanced bank of Kalman filters for in-flight aircraft engine sensor fault diagnostics. *Journal of Engineering for Gas Turbines and Power*, 127(3):497–504, 2005.

J. R. Kolodziej and D. J. Mook. Model determination for nonlinear state-based system identification. *Nonlinear Dynamics*, 63(4):735–753, 2011.

K. Kosanovich and M. Piovoso. Process data analysis using multivariate statistical methods. In *1991 American Control Conference*, pages 721–724. IEEE, 1991.

R. Kothamasu, S. H. Huang, and W. H. VerDuin. System health monitoring and prognostics—a review of current paradigms and practices. *The International Journal of Advanced Manufacturing Technology*, 28(9-10):1012–1024, 2006.

U. Kruger and L. Xie. *Advances in Statistical Monitoring of Complex Multivariate Processes: With Applications in Industrial Process Control*. Wiley, 2012.

W. E. Larimore. Canonical variate analysis in identification, filtering, and adaptive control. In *29th IEEE Conference on Decision and Control*, pages 596–604 vol.2, Dec 1990. doi: 10.1109/CDC.1990.203665.

D. K. Lee, J. In, and S. Lee. Standard deviation and standard error of the mean. *Korean journal of anesthesiology*, 68(3):220, 2015.

W. Li and S. J. Qin. Consistent dynamic PCA based on errors-in-variables subspace identification. *Journal of Process Control*, 11(6):661–678, 2001.

W. Li, H. H. Yue, S. Valle-Cervantes, and S. J. Qin. Recursive PCA for adaptive process monitoring. *Journal of process control*, 10(5):471–486, 2000.

J. Lin. On the Dirichlet distribution. Master's thesis, Department of Mathematics and Statistics, Queens University, Kingston, Ontario, Canada, 2016.

R. Y. Liu. Control charts for multivariate processes. *Journal of the American Statistical Association*, 90 (432):1380–1387, 1995.

X. Liu, U. Kruger, T. Littler, L. Xie, and S. Wang. Moving window kernel PCA for adaptive monitoring of nonlinear processes. *Chemometrics and intelligent laboratory systems*, 96(2):132–143, 2009.

Y. Liu and A. M. Bazzi. A review and comparison of fault detection and diagnosis methods for squirrel-cage induction motors: State of the art. *ISA transactions*, 70:400–409, 2017.

S. Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129–137, March 1982. doi: 10.1109/TIT.1982.1056489.

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.

J. Lu, M. Li, and D. Dunson. Reducing over-clustering via the powered Chinese restaurant process. *arXiv preprint arXiv:1802.05392*, 2018.

Y. Ma, H. Shi, and M. Wang. Adaptive local outlier probability for dynamic process monitoring. *Chinese Journal of Chemical Engineering*, 22(7):820–827, 2014.

J. Márquez-Flores. TP, FN, FP, FN and derived measures for a test, 2010. URL `http://www.academicos.ccadet.unam.mx/jorge.marquez/cursos/Instrumentacion/FalsePositive_TrueNegative_etc.pdf`. Accessed: 2020-10-20.

K. Martin. A review by discussion of condition monitoring and fault diagnosis in machine tools. *International Journal of Machine Tools and Manufacture*, 34(4):527–551, 1994.

P. Matisko and V. Havlena. Noise covariance estimation for Kalman filter tuning using bayesian approach and Monte Carlo. *International Journal of Adaptive Control and Signal Processing*, 27(11):957–973, 2013.

P. S. Maybeck. Multiple model adaptive algorithms for detecting and compensating sensor and actuator/surface failures in aircraft flight control systems. *International Journal of Robust and Nonlinear Control*, 9(14):1051–1070, 1999.

S. D. J. McArthur, S. M. Strachan, and G. Jahn. The design of a multi-agent transformer condition monitoring system. *IEEE Transactions on Power Systems*, 19(4):1845–1852, Nov 2004. doi: 10.1109/TPWRS.2004.835667.

R. Mehra. On the identification of variances and adaptive Kalman filtering. *IEEE Transactions on Automatic Control*, 15(2):175–184, April 1970. doi: 10.1109/TAC.1970.1099422.

N. Meskin, E. Naderi, and K. Khorasani. A multiple model-based approach for fault diagnosis of jet engines. *IEEE Transactions on Control Systems Technology*, 21(1):254–262, Jan 2013. doi: 10.1109/TCST.2011.2177981.

D. Moshou, D. Kateris, N. Sawalhi, S. Loutridis, and I. Gravalos. Fault severity estimation in rotating mechanical systems using feature based fusion and self-organizing maps. In *International Conference on Artificial Neural Networks*, pages 410–413. Springer, 2010.

K. Murphy. Conjugate bayesian analysis of the gaussian distribution. Technical report, The University of British Columbia, Canada, 2007.

K. P. Murphy. *Machine learning: a probabilistic perspective*. Cambridge, MA, 2012.

F. Murtagh and P. Contreras. Methods of hierarchical clustering. *arXiv preprint arXiv:1105.0121*, 2011.

S. Natarajan and R. Srinivasan. Multi-model based process condition monitoring of offshore oil and gas production process. *Chemical Engineering Research and Design*, 88(5-6):572–591, 2010.

D. Navarro and A. Perfors. The Chinese restaurant process, 2014. URL `http://compcogsci-3016.djnavarro.net/technote_chineserestaurantprocesses.pdf`. Accessed: 2020-10-21.

R. M. Neal. Bayesian mixture modeling. In *Maximum Entropy and Bayesian Methods*, pages 197–211. Springer, 1992.

R. M. Neal. Markov chain sampling methods for Dirichlet process mixture models. *Journal of computational and graphical statistics*, 9(2):249–265, 2000.

G. Niu and H. Li. IETM centered intelligent maintenance system integrating fuzzy semantic inference and data fusion. *Microelectronics Reliability*, 75:197–204, 2017.

G. Niu, B.-S. Yang, and M. Pecht. Development of an optimized condition-based maintenance system by data fusion and reliability-centered maintenance. *Reliability Engineering & System Safety*, 95(7): 786–796, 2010.

S. W. Nydick. The Wishart and inverse Wishart distributions. *Electronic Journal of Statistics*, 6(1-19), 2012.

B. J. Odelson, M. R. Rajamani, and J. B. Rawlings. A new autocovariance least-squares method for estimating noise covariances. *Automatica*, 42(2):303–308, 2006.

P.-E. P. Odiowei and Y. Cao. Nonlinear dynamic process monitoring using canonical variate analysis and kernel density estimations. *IEEE Transactions on Industrial Informatics*, 6(1):36–45, 2009.

M. Orkisz. Practical aspects of machine diagnostics accuracy. In *2017 IEEE 11th International Symposium on Diagnostics for Electrical Machines, Power Electronics and Drives (SDEMPED)*, pages 260–266, Aug 2017. doi: 10.1109/DEMPED.2017.8062365.

J. Orloff and J. Bloom. Conjugate priors: beta and normal, 2018. URL `http://www-math.mit.edu/~dav/05.dir/class15-slides-all.pdf`. Accessed: 2019-10-18.

J. Paisley. A tutorial on the Dirichlet process for engineers. Master's thesis, Duke University, Durham, North Carolina, 2015.

C. C. Pantelides and J. Renfro. The online use of first-principles models in process operations: Review, current status and future needs. *Computers & Chemical Engineering*, 51:136–148, 2013.

K. Patan. *Artificial neural networks for the modelling and fault diagnosis of technical processes*. Springer, 2008.

R. J. Patton, P. M. Frank, and R. N. Clark. *Issues of fault diagnosis for dynamic systems*. Springer Science & Business Media, 2013.

K. Pearson. Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London. A: Mathematical, Physical and Engineerign Sciences*, 186:343–414, 1895.

D. Pebrianti, R. Samad, M. Mustafa, N. R. H. Abdullah, and L. Bayuaji. Bank of Kalman filters for fault detection in quadrotor MAV. *Journal of Engineering and Applied Sciences*, 11(10):6668–6674, 2016.

E. J. Perreault, R. F. Kirsch, and A. M. Acosta. Multiple-input, multiple-output system identification for characterization of limb stiffness dynamics. *Biological cybernetics*, 80(5):327–337, 1999.

K. E. Pilario. Pca-based fault detection for 2D multivariate process data, 05 2020. URL https://www.mathworks.com/matlabcentral/fileexchange/65983-pca-based-fault-detection-for-2d-multivariate-process-data,. Accessed: 2020-05-14.

K. E. Pilario, M. Shafiee, Y. Cao, L. Lao, and S.-H. Yang. A review of kernel methods for feature extraction in nonlinear process monitoring. *Processes*, 8(1):24, 2020.

M. Piovoso and A. Owens. Sensor data analysis using artificial neural networks. *Chemical Process Control CPC IV*, pages 101–118, 1991.

S. Qin and T. McAvoy. A data-based process modeling approach and its applications. In *Dynamics and Control of Chemical Reactors, Distillation Columns and Batch Processes*, pages 93–98. Elsevier, 1993.

S. J. Qin. Recursive PLS algorithms for adaptive data modeling. *Computers & Chemical Engineering*, 22(4-5):503–514, 1998.

S. J. Qin. Survey on data-driven industrial process monitoring and diagnosis. *Annual reviews in control*, 36(2):220–234, 2012.

M. Quiñones-Grueiro, A. Prieto-Moreno, C. Verde, and O. Llanes-Santiago. Data-driven monitoring of multimode continuous processes: A review. *Chemometrics and Intelligent Laboratory Systems*, 189: 56–71, 2019.

M. R. Rajamani and J. B. Rawlings. Estimation of the disturbance structure from data using semidefinite programming and optimal weighting. *Automatica*, 45(1):142–148, 2009.

B. Rao. *Handbook of condition monitoring*, chapter 1. Elsevier, 1996.

C. E. Rasmussen. The infinite gaussian mixture model. In *Advances in neural information processing systems*, pages 554–560, 2000.

A. Rastegari and M. Bengtsson. Implementation of condition based maintenance in manufacturing industry-a pilot case study. In *2014 International Conference on Prognostics and Health Management*, pages 1–8. IEEE, 2014.

R. S. Risuleo, G. Bottegal, and H. Hjalmarsson. Kernel-based system identification from noisy and incomplete input-output data. In *2016 IEEE 55th Conference on Decision and Control (CDC)*, pages 2061–2066, Dec 2016. doi: 10.1109/CDC.2016.7798567.

Rong Chen, Xiaodong Wang, and J. S. Liu. Adaptive joint detection and decoding in flat-fading channels via mixture Kalman filtering. *IEEE Transactions on Information Theory*, 46(6):2079–2094, Sep. 2000. doi: 10.1109/18.868479.

J. Roux. An introduction ot Kalman filtering: probabilistic and deterministic approaches. Technical report, University of Nice, 2003.

Rui Xu and D. Wunsch. Survey of clustering algorithms. *IEEE Transactions on Neural Networks*, 16(3): 645–678, May 2005. doi: 10.1109/TNN.2005.845141.

C. Ruiz-Cárcel, Y. Cao, D. Mba, L. Lao, and R. Samuel. Statistical process monitoring of a multiphase flow facility. *Control Engineering Practice*, 42:74–88, 2015.

E. Russell, L. Chiang, and R. 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.

E. L. Russell, L. H. Chiang, and R. D. Braatz. *Data-driven methods for fault detection and diagnosis in chemical processes*. Springer Science & Business Media, 2012.

R. Schlaifer and H. Raiffa. *Applied statistical decision theory*. Harvard University Press, 1961.

N. Schuurman, R. Grasman, and E. Hamaker. A comparison of inverse-Wishart prior specifications for covariance matrices in multilevel autoregressive models. *Multivariate Behavioral Research*, 51(2-3): 185–206, 2016.

J. Shang, M. Chen, and D. Zhou. Multimode process monitoring based on conditionally independent bayesian learning. In *2017 IEEE 56th Annual Conference on Decision and Control (CDC)*, pages 2705–2710, Dec 2017. doi: 10.1109/CDC.2017.8264052.

X. Shen and S. Agrawal. Kernel density estimation for an anomaly based intrusion detection system. In *Proceedings of the 2006 World Congress in Computer Science, Computer Engineering and Applied Computing*, pages 161–167, 2006.

W. A. Shewhart. Quality control charts. *The Bell System Technical Journal*, 5(4):593–603, 1926.

B. W. Silverman. *Density estimation for statistics and data analysis*. Monographs on Statistics and Applied Probability, London: Chapman and Hall, 1986.

S. Simani, C. Fantuzzi, and R. J. Patton. Model-based fault diagnosis techniques. In *Model-based Fault Diagnosis in Dynamic Systems Using Identification Techniques*, pages 19–60. Springer, 2003.

P. Smyth. Clustering sequences with hidden Markov models. In *Advances in neural information processing systems*, volume 9, pages 648–654, 1997.

X. Song, M. Wu, C. Jermaine, and S. Ranka. Conditional anomaly detection. *IEEE Transactions on Knowledge and Data Engineering*, 19(5):631–645, May 2007. doi: 10.1109/TKDE.2007.1009.

R. Srinivasan, C. Wang, W. Ho, and K. Lim. Dynamic principal component analysis based methodology for clustering process states in agile chemical plants. *Industrial & engineering chemistry research*, 43 (9):2123–2139, 2004.

R. Srinivasan, P. Viswanathan, H. Vedam, and A. Nochur. A framework for managing transitions in chemical plants. *Computers & chemical engineering*, 29(2):305–322, 2005.

A. Stief. *Combining data from disparate sources for condition monitoring purposes*. PhD thesis, AGH University of Science and Technology in Krakow, 2019.

A. Stief, R. Tan, Y. Cao, J. R. Ottewill, N. F. Thornhill, and J. Baranowski. Multiphase flow facility case study with heterogeneous data. 2018a. doi: http://dx.doi.org/10.5281/zenodo.1341583.

A. Stief, R. Tan, Y. Cao, J. R. Ottewill, N. F. Thornhill, and J. Baranowski. Multiphase flow facility case study technical report. 2018b. doi: http://dx.doi.org/10.5281/zenodo.1341583.

A. Stief, R. Tan, Y. Cao, J. R. Ottewill, N. F. Thornhill, and J. Baranowski. A heterogeneous benchmark dataset for data analytics: Multiphase flow facility case study. *Journal of Process Control*, 79:41–55, 2019.

V. M. Stijn. Change detection in system parameters of lithography machines. Master's thesis, Eindhoven University of Technology, 2018.

A. Stuart and J. K. Ord. *Kendall's Advanced Theory of Statistics: Volume I- Distribution Theory*. Edward Arnold, 1994.

R. Tan, T. Cong, N. F. Thornhill, J. R. Ottewill, and J. Baranowski. Statistical monitoring of processes with multiple operating modes. *IFAC-PapersOnLine*, 52(1):635–642, 2019.

R. Tan, J. R. Ottewill, and N. F. Thornhill. Nonstationary discrete convolution kernel for multimodal process monitoring. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–12, 2019.

R. Tan, T. Cong, J. Ottewill, J. Baranowski, and N. Thornhill. An on-line framework for monitoring nonlinear processes with multiple operating modes. *Journal of Process Control*, 89:119–130, 2020.

S. Tan, F. Wang, J. Peng, Y. Chang, and S. Wang. Multimode process monitoring based on mode identification. *Industrial & Engineering Chemistry Research*, 51(1):374–388, 2012.

Y. W. Teh. Dirichlet process. *Encyclopedia of machine learning*, pages 280–287, 2010.

N. Thornhill, M. Oettinger, and P. Fedenczuk. Refinery-wide control loop performance assessment. *Journal of Process Control*, 9(2):109–124, 1999.

N. F. Thornhill, H. Melbø, and J. Wiik. Multidimensional visualization and clustering of historical process data. *Industrial & engineering chemistry research*, 45(17):5971–5985, 2006.

K. Tidriri, N. Chatti, S. Verron, and T. Tiplica. Bridging data-driven and model-based approaches for process fault diagnosis and health monitoring: A review of researches and future challenges. *Annual Reviews in Control*, 42:63–81, 2016.

L. Tierney. Markov chains for exploring posterior distributions. *the Annals of Statistics*, 22(4):1701–1728, 1994.

L. Y. Tseng and S. B. Yang. A genetic approach to the automatic clustering problem. *Pattern recognition*, 34(2):415–424, 2001.

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.

F. Vicario, M. Q. Phan, R. Betti, and R. W. Longman. Output-only observer/Kalman filter identification (O3KID). *Structural Control and Health Monitoring*, 22(5):847–872, 2015.

H. Wang, X. Ye, and M. Yin. Study on predictive maintenance strategy. *International. Journal of Science and Technology*, 9(4):295–300, 2016.

X. Wang, U. Kruger, and G. W. Irwin. Process monitoring approach using fast moving window PCA. *Industrial & Engineering Chemistry Research*, 44(15):5691–5702, 2005.

X. Z. Wang, S. Medasani, F. Marhoon, and H. Albazzaz. Multidimensional visualization of principal component scores for process historical data analysis. *Industrial & engineering chemistry research*, 43(22):7036–7048, 2004.

S. Weisberg. *Applied linear regression*. John Wiley & Sons, 1980.

M. West. Hyperparameter estimation in Dirichlet process mixture models. Technical report, Institute of Statistics and Decision Sciences, Duke University, 1992.

D. T. Westwick and E. J. Perreault. Closed-loop identification: Application to the estimation of limb impedance in a compliant environment. *IEEE Transactions on Biomedical Engineering*, 58(3):521–530, March 2011. doi: 10.1109/TBME.2010.2096424.

S. Wold. Exponentially weighted moving principal components analysis and projections to latent structures. *Chemometrics and intelligent laboratory systems*, 23(1):149–161, 1994.

O. Wu, A. Bouaswaig, L. Imsland, S. M. Schneider, M. Roth, and F. M. Leira. Campaign-based modeling for degradation evolution in batch processes using a multiway partial least squares approach. *Computers & Chemical Engineering*, 2019.

X. Xie and H. Shi. Dynamic multimode process modeling and monitoring using adaptive gaussian mixture models. *Industrial & Engineering Chemistry Research*, 51(15):5497–5505, 2012.

E. Xing. Bayesian nonparametrics: Dirichlet processes, 2014. URL https://www.cs.cmu.edu/~epxing/Class/10708-14/scribe_notes/scribe_note_lecture19.pdf Accessed: 2019-10-18.

W. Xue, Y.-q. Guo, and X.-d. Zhang. A bank of Kalman filters and a robust Kalman filter applied in fault diagnosis of aircraft engine sensor/actuator. In *Second International Conference on Innovative Computing, Informatio and Control (ICICIC 2007)*, pages 10–10. IEEE, 2007.

F. Yang, D. Xiao, and S. L. Shah. Optimal sensor location design for reliable fault detection in presence of false alarms. *Sensors*, 9(11):8579–8592, 2009.

X. Yang, Y. Zhou, and M. Liu. Fusion recognition fingerprints and handwritten signature recognition fusion based on the bayesian algorithm. In *The International Conference on Artificial Intelligence and Software Engineering (ICAISE 2013)*. Atlantis Press, 2013.

Y. Yang, Y. Ma, B. Song, and H. Shi. An aligned mixture probabilistic principal component analysis for fault detection of multimode chemical processes. *Chinese Journal of Chemical Engineering*, 23(8): 1357–1363, 2015.

I. Yildirim. Bayesian inference: Gibbs sampling. Technical note, Department of Brain and Cognitive Sciences Univerisyt of Rochester, August 2012.

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

Z. Yu and J. Falnes. State-space modelling of a vertical cylinder in heave. *Applied Ocean Research*, 17 (5):265–275, 1995.

H. H. Yue and S. J. Qin. Reconstruction-based fault identification using a combined index. *Industrial & engineering chemistry research*, 40(20):4403–4414, 2001.

C. Zhang, X. Gao, T. Xu, and Y. Li. Nearest neighbor difference rule–based kernel principal component analysis for fault detection in semiconductor manufacturing processes. *Journal of Chemometrics*, 31 (6):e2888, 2017a.

K. Zhang, Y. A. Shardt, Z. Chen, and K. Peng. Using the expected detection delay to assess the performance of different multivariate statistical process monitoring methods for multiplicative and drift faults. *ISA transactions*, 67:56–66, 2017b.

K. Zhang, K. Peng, and J. Dong. A common and individual feature extraction-based multimode process monitoring method with application to the finishing mill process. *IEEE Transactions on Industrial Informatics*, 14(11):4841–4850, Nov 2018. doi: 10.1109/TII.2018.2799600.

Y. Zhang, H. Zhou, S. J. Qin, and T. Chai. Decentralized fault diagnosis of large-scale processes using multiblock kernel partial least squares. *IEEE Transactions on Industrial Informatics*, 6(1):3–10, 2009.

C. Zhao and F. Gao. Fault-relevant principal component analysis (FPCA) method for multivariate statistical modeling and process monitoring. *Chemometrics and Intelligent Laboratory Systems*, 133:1–16, 2014.

S. J. Zhao, J. Zhang, and Y. M. Xu. Monitoring of processes with multiple operating modes through multiple principal component analysis models. *Industrial & engineering chemistry research*, 43(22): 7025–7035, 2004.

S. J. Zhao, J. Zhang, and Y. M. Xu. Performance monitoring of processes with multiple operating modes through multiple PLS models. *Journal of process Control*, 16(7):763–772, 2006.

M. Zheng, S. Krishnan, and M. P. Tjoa. A fusion-based clinical decision support for disease diagnosis from endoscopic images. *Computers in Biology and Medicine*, 35(3):259–274, 2005.

L. Zhou, Z. Song, J. Chen, Z. Ge, and Z. Li. Process-quality monitoring using semi-supervised probability latent variable regression models. *IFAC Proceedings Volumes*, 47(3):8272–8277, 2014.

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.

N. Zhou, Z. Huang, Y. Li, and G. Welch. Local sequential ensemble Kalman filter for simultaneously tracking states and parameters. In *2012 North American Power Symposium (NAPS)*, pages 1–6. IEEE, 2012.

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.

J. Zhu, Z. Ge, and Z. Song. Multimode process data modeling: A Dirichlet process mixture model based bayesian robust factor analyzer approach. *Chemometrics and Intelligent Laboratory Systems*, 142: 231–244, 2015.

W. Zhu, N. Zeng, N. Wang, et al. Sensitivity, specificity, accuracy, associated confidence interval and ROC analysis with practical SAS implementations. *NESUG proceedings: health care and life sciences, Baltimore, Maryland*, 19:67, 2010.

AUTHORS DECLARATION

I DECLARE, WHILE AWARE OF CRIMINAL RESPONSIBILITY FOR STATING AN UNTRUTH, THAT I HAVE CARRIED OUT MY DOCTORAL DISSERTATION PERSONALLY AND INDEPENDENTLY, AND HAVE NOT USED ANY SOURCES OTHER THAN THOSE LISTED IN THE BIBLIOGRAPHY.

OŚWIADCZENIE AUTORKI PRACY

OŚWIADCZAM, ŚWIADOMA ODPOWIEDZIALNOŚCI KARNEJ ZA POŚWIADCZENIE NIEPRAWDY, ŻE NINIEJSZĄ PRACĘ DYPLOMOWĄ WYKONAŁAM OSOBIŚCIE I SAMODZIELNIE, I NIE KORZYSTAŁAM ZE ŹRÓDEŁ INNYCH NIŻ WYMIENIONE W PRACY.

.....

PODPIS