공학박사 학위논문

Data-driven Approaches to Fault Detection and Diagnosis under Multiple Faults

다중 원인 이상 감지 및 진단을 위한 데이터 기반 방법

2018년 2월

서울대학교 대학원
화학생물공학부
김 대 식

Abstract

Data-driven Approaches to Fault Detection and Diagnosis under Multiple Faults

Dae Shik Kim
School of Chemical and Biological Engineering
The Graduate School
Seoul National University

Fault detection and diagnosis (FDD) has been an important issue in chemical industry for optimal operation and process safety. FDD has three different approaches which are model-based approach, knowledge-based approach and data-driven approach. Recent advances in data acquisition and storage techniques have enabled high-frequency sampling and processing of sensor signals. Therefore, the data-driven methods can handle the limitations of the traditional FDD method.

To improve the FDD performance, three advanced FDD schemes were proposed. The first proposed method was the combination of model-based and data-driven approaches. If the unknown parameters of the process model is inaccurate, the result of FDD with model-based approach can be poor. In addition, since some processes, such as pharmaceutical process, are hard to collect measurement data, the robust parameter estimation with limited data is necessary. In this

reason, Bayesian inference was introduced to estimate the unknown parameters of physiologically based pharmacokinetic (PBPK) model with a small number of data. With the proposed estimation scheme, the estimation result was more robust than the least squares method. In addition, the model mismatch was reduced by introducing the drug dissolution model (DDM) into the PBPK model. With these results, FDD performance of model-based approach can be improved.

When the abundant data collection is possible, faulty state data can be classified by the differences between the normal data sets and fault data sets. To describe the data differences, Support vector machine (SVM) which is one of the machine learning technique was introduce to help the transient analysis of water pipe network to diagnose the partial blockage. The time domain transient data were convert to the frequency domain data to find the differences between the normal pipe and blocked pipe. With test experiences with various sizes of the blockage, normal, small blockage, medium blockage and harsh blockage transient data were collected. SVM structures of four cases of blockages were constructed with converted transient data. Finally, SVM structures can classify the blocked pipe and its blockage size automatically with the transient analysis data.

The data-based model is accurate when the learning data describes the characteristic of the process perfectly. Usually, it is impossible to collect perfect learning data from the operating process. Therefore, knowledge-based model can help to reduce model mismatch of the data-based model with prior information of the process and intuition of the expert engineer. Bayesian belief network (BBN)

is data-based model which describes the causality between the measurements of the process. To construct BBN structure with imperfect data, weight matrix from the signed digraph (SDG), which is one of the knowledge-based model, was proposed and applied to the structure learning algorithm. In addition, the root cause of the pre-defined fault scenario also introduced into the BBN with prior information of the process. Three case studies was conducted to verify the FDD performance of BBN-based fault diagnosis method with single fault scenarios and multiple fault scenarios. The BBN-based method was effective for all case studies compared with the traditional PCA-based method. Moreover, the fault diagnosis rate of the BBN-based method was better than the PCA-based method for not only single fault cases but multiple faults cases. Consequently, the BBN-based fault diagnosis method, which is the combination of knowledge-based and data-driven approaches, can improve the FDD performance compared with the traditional data-based approaches.

With the three proposed ways to improve the traditional FDD approaches, accurate and real-time process monitoring is possible. Therefore, the proposed methods can help to maintain the process when the failures occur and remain the process with optimal operation condition.

Keywords: Process monitoring, Data-driven approach, Bayesian network, Multivariate analysis, Fault diagnosis, Machine learning
Student Number: 2014-30259

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Fault Detection and Diagnosis

Fault detection and diagnosis (FDD) has been an important issue in chemical industry [1]. Because of various kinds of disturbance, equipment failure or mistake of operator, process faults can be presented during the operation. Some process faults can be easily identified because they changed some sensor values from normal condition to the out of the upper or lower limit. Even if those values remain in the normal range, huge differences in a short time can be monitored and the faults can be fixed by operator. However, lots of faults are hard to identified because sensor value change can be hidden under the measurement noise [2]. Moreover, investigation of the root cause is difficult since measurements of the process are highly correlated and a large number of feedback loops are located in the process. If the faults were not identified and the process was operating under the faulty condition, product quality should be worse and equipment will be severely damaged. Therefore, to maintain a proper product quality and reduce operation and maintenance cost, direct fault detection and accurate root cause search are necessary.

For chemical plant, various faults, such as leakage, overheating, overpressure and overflow, can be exist. Because most of raw materials of chemical process is hazardous, flammable and toxic, chemical plant must be maintain safely. Moreover, chemical plant usually is operated in high pressure and high temperature condition [3]. Therefore, safety of chemical process is the most important issue in industrial field. There are two kinds of processes, widely used in chemical industry and examples of FDD for chemical industry. The one is continuous process and the other is batch process. Continuous stirred-tank reactor (CSTR) and Wet gas compressor (WGC) are typical continuous processes. Because most of industrial chemical reaction is exothermic reaction and operate under high pressure/temperature, monitoring chemical reactor is essential to maintain the process safely [4]. In addition, WGC consists of many mechanical equipment such as valves, suction drums, heat exchangers and compressor. Within the operation, equipment is exposed to external damage from repetitive vibration, corrosive materials and pressure. Since various faults can be occurred in WGC frequently compared with other process unit, FDD for WGC to immediate maintenance helps to improve the product quality and optimal operation [5]. However, accurate FDD for continuous processes is difficult problem due to the limitation of the number of sensors, disturbances and uncertainties.

Another example of continuous process is water pipe network. Especially, effective management of aged water pipe infrastructures is an important issue of water supply. Partial blockage owing to the precipitation of salts from water is one of the serious faults that can occur in aged water pipes. Because the partial blockage shortens the

lifespan of pipeline and increases the water supply pressure, it can cause leak, burst and energy waste during the water supply [6]. Moreover, fragments of the blockages mixed with drinking or home water can cause the hygiene problem [7, 8]. Since the blockages in water pipeline can be easily removed by chemical treatment [9] [10], immediate scale removal or blocked pipe replacement given early diagnosis can help the maintenance significantly. However, fault diagnosis for water pipe networks is a difficult task because of the lack of direct measurement of the internal state of underground pipeline. Monitoring the interior of pipe using endoscope is the most accurate method to detect partial blockage for early diagnosis. Despite the accuracy of the endoscope monitoring, this method can observe limited area of the water pipe network where entering site for endoscope exists. Therefore, most of the water pipes are maintained by preset schedules, which prescribe to inspect or replace pipes periodically [11, 12, 13]. The maintenance period is determined by regulation, past experience or pipe lifespan depending on the pipe materials. Since the frequency of manual maintenance is pre-determined without observations, it is neither economically optimal nor able to handle serious faults in a timely fashion [14].

Although improvement of FDD performance for continuous process, nonlinearity and correlation between the process variables can be acceptable. However, effective monitoring for batch process is more challenging problem since the representative characteristic of batch process is highly nonlinear and complex correlation and those make hard to diagnose the multiple root causes of faults [15, 16]. Moreover, typical batch process operate with more high temperature

and pressure compared with continuous process. Therefore, the probability of fault occurrence of batch process is higher and fault of batch process can cause severe damage to equipment and operators [17]. In addition, because the batch process is used for bioproduct and pharmaceutics which are high-value products, fault-free operation with FDD can improve the product quality and process productivity. Therefore, efficient process monitoring approach to overcome a nonlinearity of the batch process monitoring is essential.

There are three ways to detect and diagnose process failures: model-based methods, knowledge-based methods and data-driven methods [18]. The model-based method uses analytic model to monitor the process statement. The knowledge-based methods use qualitative causal models (digraphs, fault-tree) or abstraction hierarchy [19, 20]. Recent advances in data acquisition and storage techniques have enabled high-frequency sampling and processing of sensor signals. Therefore, since it is available to find differences between normal state and abnormal state from the process sensor signals, the data-driven methods such as principal component analysis (PCA) or partial least squared (PLS) can be used to FDD [21, 22]. In followings, details of three ways for FDD will be introduced.

1.1.1 Model-based approaches

Basic idea of model-based FDD approach is observation of the differences between process outputs and predicted outputs from analytic model. If the analytic model of process is accurate, predicted

outputs should have neighboring values of process outputs. However, if the process outputs have different values compared with prediction, failures should be occurred in target process. Therefore, the most important issue of model-based approach is accuracy of analytic model. To construct accurate analytic model, effective parameter estimation is essential. Least squares method is widely used parameter estimation tool because of its simplicity [23, 19]. Kalman filter can be also used for state estimation of process model [24, 25]. In addition, some effective estimation methods have been developed which are relying on parity equation for residual generation [26]. However, the chemical process is complex and the variables are highly correlated. Moreover, since there exist lots of uncertainties and nonlinearities which make hard to estimate model parameters, accuracy of analytic model is hard to be guaranteed. In addition, because characteristics of target process will be changed within the process operation, the parameters should be updated properly to maintain the FDD performance.

Analytic model-based approach can be applied to pharmaceutical industry because the drug discovery process takes an enormous amount of time, money, and effort to produce and monitor the process. Nevertheless, to prevent side-effects of a drug and find an optimal dosage, a large number of tests need to be conducted on various subjects. However, experiments of a new drug on human carry a great risk, because the toxicity and side-effects of a new drug are often unknown. Mathematical models describing drug delivery mechanisms in terms of drug concentrations in each organ over a time course can be of significant help in reducing the cost of development and the risk of failure. Therefore, time-course data are collected to construct

physiologically based pharmacokinetics (PBPK) models during animal and human trials (Phases I-III) [27]. Pharmacokinetics (PK) is the study of absorption, distribution, metabolism, and excretion of the chemicals in a living body, and plays an important role in the development of drugs [28, 29]. The PBPK model includes mechanistic and physiological basis describing the pharmacokinetics of drugs within a biological entity [30]. If a PBPK model is constructed, it can be used not only for the prediction of PK profiles of drugs, but also for dose regulation as in feedback control strategies [27]. Therefore, the PBPK model can also allow for determination of the optimal dosage and administration time [31].

Because PBPK models are only concerned with the dynamics inside a body, they commonly include a number of organs and blood vessels. However, drug dissolution dynamics is usually not considered. In the past, most medicines with serious side-effects or dosage sensitivity were usually in the form of a liquid, because they need to be absorbed quickly [32]. Since this type of drugs is absorbed at a high rate, drug dissolution dynamics could be ignored. However, with the recent development of various drug dosage forms, some drugs are produced in the form of a tablet or capsule to control the dissolution rate [33, 34]. Since the dissolution dynamics of medicine in a non-liquid form is an important part of the drug's PK profile, it is necessary to describe the dissolution dynamics when constructing a PBPK model. In addition, the dissolution rate of drugs for individuals will be very different with a large variance because of their physical properties and individual genotypic variations [35]. It is also impossible to collect dissolution data for every patient to estimate a personal disso-

lution rate. In this manner, the dissolution parameter would be better described as a probability distribution to reflect personal differences. Nevertheless, PBPK and drug dissolution models have been developed and their parameters have been estimated independently. Therefore, development of a model that combines PBPK and drug dissolution dynamics in a compatible manner can help to provide reasonable estimates of the dissolution parameters in a lumped form, which can also reflect personal differences of each patient.

PBPK models involve both physiological and kinetic parameters. Physiological parameters, include organ volume, blood flow rate, and blood volume, etc. Kinetic parameters include absorption rate and Michaelis-Menten constants for enzyme reactions. Whereas physiological parameters can be measured or specified easily, kinetic parameters are generally difficult to specify [36]. Therefore, unknown parameters should be estimated with experimental data. However, experiments to collect in vivo data are expensive, and often have poor repeatability [37]. Estimating the parameters of a PBPK model with such poor data sets is further complicated by the concentration profiles, which show a pattern of declining exponential functions, with amplitudes and decay times [38]. In addition, since each individual may have different parameter values depending on their own physiological properties, the drug concentration profiles can vary between test subjects. With these uncertainties, parameters can be treated as random variables and described by probability distributions. The least squares method is the most widely used estimation method, and finds the point estimate of the parameters by minimizing the sum of squared errors between actual observations and pre-

dictions. Since there is no probabilistic structure in the least squares method, it is sensitive to the presence of unusual data points; one or two outliers can sometimes seriously skew the results, thus raising the requirement for a large number of data points. However, the number of available in vivo drug concentration data points is often limited, and involves a high degree of uncertainty. This requires a robust estimation method which would be suitable for PBPK models.

Statistical inference based on Bayes' rule can be used for parameter estimation of the PBPK model. In particular, the maximum a posteriori (MAP) principle is one of the Bayesian inference methods considering the prior information on the parameter and differences between model outputs and experimental observations [39]. Because the prior knowledge is incorporated into the estimation, MAP methods can be more robust than the least squares or maximum likelihood estimation (MLE) methods [40, 41], and can provide more accurate estimates when the data are contaminated with noise, or when the number of data points is small [39]. For simple PK models with one or two model equations, the MAP method is easy to implement because the model parameters are nearly uncorrelated. In this case, the parameter distribution obtained from in vitro experiments can be used as the prior information [42]. However, model parameters are often correlated since the organs are interconnected by blood vessels. While the prior information cannot be incorporated into a MAP method without information on the parameter correlations, i.e., covariance matrix, a large number of experimental data sets are necessary for estimating the correlations between each pair of parameters [43].

Model-based FDD for water pipe network is another issue of infrastructure maintenance. One dimensional momentum and continuity equations are used to detect and diagnose the partial blockage with transient analysis [44, 45, 46, 47]. Such fundamental models can describe the relationships between the hydraulic head of transient flow and the location and size of partial blockage. The models can also compute difference between the hydraulic heads of normal and faulty states. However, even if a fundamental model is available, hydraulic head signals from a real pipe network are characterized by a very low signal-to-noise ratio. Therefore, the monitoring method based on a fundamental model is difficult to apply to real pipeline networks, and another monitoring methods are necessary to solve the problem.

1.1.2 Knowledge-based approaches

Prior knowledge of the process can help to detect and diagnose the process failures. Since the experienced process operators and process engineers can detect the process failures and figure out the root causes of the faults empirically. Knowledge-based approach captures the prior knowledge to detect and diagnose the process faults in a formal methodology [48]. While the model-based approaches build process model with analytic equations from physical and chemical behaviors of the process, knowledge-based approaches only consider the prior knowledge from the experience and intuition without mathematical equations. Therefore, knowledge-based models are described as graphical models which show causal interaction between the process variables.

A signed digraph (SDG) is widely used knowledge-based model [49, 20]. In SDG, each variable is called 'node', and arrow described casual interaction between two nodes called 'edge'. The origin node of edge arrow is called 'parent' and the end node of edge is called 'child'. If two variables have casual interaction each other, the parent variable affect the child variable. Therefore, the direction of edge is determined by figuring out which variable is effected and which variable affect other. The nodes of SDG are usually output variables of the process. However, the edges of SDG and the direction of edges are determined by the prior knowledge of target process. Therefore, the knowledge of the process can describe the process behaviors perfectly, SDG model can be used for FDD of the process. In contrast, if the prior knowledge was wrong or ignored some interactions between the process variables, FDD performance with SDG model should be poor [50]. Fault-tree is another knowledge-based model which is a logic diagram [51]. Fault-tree is built with process failures and consequent failures which are connected by cause and effect relationship. The process failures which are identified from prior knowledge can be arranged in a fault-tree structure and the consequent failures, which is brought by the process failures, are evaluated in the structure. There are some standard logic and event symbols in Figure 1.1 [52]. Similar to SDG, if the logic and event tree from the prior knowledge can describe the causes and effects between the variables, FDD performance of fault-tree analysis is guaranteed. However, if the undesired event or different characteristic process faults, which were not described in fault-tree, was generated, fault-tree analysis cannot deal with the unknown event.

Nowadays, knowledge-based model is not used to FDD solely. Although traditional knowledge-based approaches only depend on the prior knowledge of system and personal experience, advanced knowledge-based model is modified with process history data. Knowledge-based approaches was studied in early 90s when sensor and computer technology is not enough to collect, store and handle a large number of data. Recently, since the computational technology and data acquisition are great improved, shortcomings of knowledge-based approaches can be adjusted by process history data and unknown process failures and unexpected failures' consequent can be detected. Therefore, data-driven approaches for FDD becomes more important.

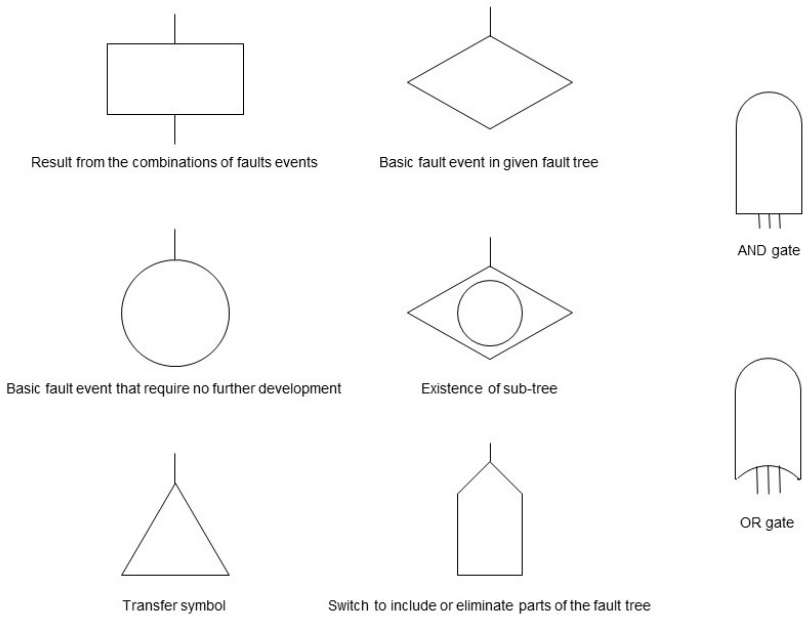

Figure 1.1: Examples of standard fault tree symbolism

1.1.3 Data-driven approaches

Recent advances in data acquisition and storage techniques have enabled high-frequency sampling and processing of sensor signals. Therefore, the data-driven methods can handle the problems of the previous research. PCA and PLS are widely used multivariate analysis methods for FDD. PCA/PLS projected the process data to the latent space and classified the differences between normal and faulty state using statistic methods such as Mahalanobis and Euclidian distances [53]. However, for nonlinear system, lineal PCA/PLS has limitation to describe its nonlinearity, various nonlinear PCA/PLS methods was developed [54, 55]. Nowadays, since the collecting huge number of process data sets is available, Artificial neural network (ANN), also a part of qualitative approach, can be also used for process monitoring [56]. Because ANN model is more efficient to describe nonlinearity compared with PCA/PLS, it can help to find root cause of faults from the nonlinear system [57].

Support vector machine (SVM) is a statistical learning method for classification or regression [58]. SVM constructs $n - 1$ dimensional hyperplanes which can classify n dimensional data with two classes [59, 60, 61]. Even if the hyperplanes for classification are nonlinear and difficult to define, SVM can identify the hyperplanes with a proper kernel function. If a characteristic of faulty state can be categorized, a model of faulty state data can be constructed using SVM the input of which is an original signal and output is a classified signal. The differences between normal and faulty states can be analyzed using the scores of SVM model.

With high frequency hydraulic head signal, identification of momentary differences in hydraulic head signals generated by transient flow in normal and faulty conduits has become possible. Hydraulic transient is the short-lived pressure wave that is changed and returns to the original after a certain periods of time. In order to create a transient flow, hydraulic head of pipe should be temporally changed by external manipulations such as water hammering effect or rapid opening and closing of valve. If there exist a fault inside the pipe, such as leak, burst or partial blockage, the trend of transient signal is different from that of the normal pipe. Fault diagnosis using the temporal differences in signals is called the transient analysis [62, 63, 64]. Transient analysis cannot only detect the partial blockage but also estimate the size of blockage [65, 66, 67, 68, 47].

When the transient flow is generated in a partially blocked pipe, pressure signal will be disturbed by reflected pressure waves from the blockage [69]. As a result, the trend of signal in a partially blocked pipe can be different from that of the normal pipe. Differences in the pressure signals of normal and faulty states are quite irregular and difficult to classify due to the measurement noise and the effect of pipe materials [70]. To identify a faulty state, the differences arising from faulty states should be well modelled with various data sets from repetitive experiments reflecting normal and faulty states under different conditions. Manual classification may yield inaccurate results because of its ambiguous criteria. Hence, a data-based learning method, SVM method, to provide the quantitative and objective criteria for classification is required.

However, conventional data-driven FDD approaches were limited to diagnose accurate root causes [71]. Especially when the faults were generated more than one from multiple causes, root cause diagnosis using multivariate analysis was hard to figure out accurately because influence of faults was overlapped [72]. Moreover, for batch process, correlation between the process variables and nonlinearities of process make hard to construct analytic and knowledge-based model and hard to diagnose the multiple root causes of faults [15]. Although the ANN also deals with nonlinearity of batch process, ANN needs a huge number of data sets to construct an accurate model for FDD. In addition, it is difficult to ensure the reliability of FDD result to field operator because ANN is basically a black box model that can not show real causality between the process variables.

Bayesian belief network (BBN) is one of the machine learning technique to estimate causality between the variables stochastically. BBN is acyclic graph that consists of nodes which represents the process variables and edges which represents the dependency of linked variables [73]. From the data, the edges of the BBN is built called 'structure learning' and the quantity of the dependency is estimated called 'parameter learning'. Using constructed BBN structure and edges' parameters, probability of root cause can be calculated by Bayes' rule. Therefore, BBN can be applied to FDD using the causality between the variables, widely used in the mechanical systems [74, 75]. Since BBN can be used to estimate unknown dependency between the variables from the data, it was also applied to bioinformatics research for handling time-series gene expression

data [76, 77]. Moreover, BBN can use prior knowledge about system and its faults, BBN help to improve the diagnosis performance for multiple faults from various causes [78]. If BBN model of the batch process constructed by the operation data can describe the correlation and nonlinearity, accurate FDD for the batch process is possible. In addition, in contrast with ANN, BBN model can describe the causality between the process variables intuitionally. Therefore, BBN can help to understand the result of FDD for field operator within the maintenance.

1.2 Objective & Outlook

Main objective of the thesis is proposal advanced FDD schemes to overcome present limitations of traditional FDD approaches. There are two kinds of effort to adjust the problems. The first approach is improvement of analytic model for model-based methods with Bayesian parameter estimation. Because the most important factor of model-based approaches is accuracy of analytic models. To assure the model accuracy, a proper parameter estimation scheme is necessary. With Bayesian parameter estimation, which is one of the data science field, the combination of model-based and data-driven can be proposed in this thesis. The other approach focuses on data-driven methods. To classify normal data and faulty data, machine learning technique was applied. Moreover, prior knowledge of the process was combined to data-driven method to improve FDD performance. The proposed combined data-driven method was applied to the batch process which was known as highly nonlinear system with complex correlations.

The summary of three proposed approaches are below:

- Improvement of model-based approaches I: Bayesian parameter estimation scheme to improve accuracy of analytic model with a small number of data.
- Proposed data-driven approach I: SVM-based FDD scheme for partial blocked water pipe system.
- Proposed data-driven approach II: BBN-based FDD scheme for multiple faults detection and diagnosis and application for the batch process monitoring.

The first work is advanced parameter estimation scheme to improve accuracy of analytic model. The model-based approaches can be improved with proposed parameter estimation scheme because the FDD performance of model-based approaches mostly depends on the accuracy of analytic model. The first study presents a PBPK model augmented with dissolution dynamics and a robust parameter estimation scheme, given a limited number of data sets shown in Figure 1.2. The maximum a posterior (MAP) method is used to estimate the parameters of the PBPK model. The covariance matrix in prior distribution is calculated by simulation. With the prior distribution and likelihood, the objective function of the MAP method is minimized for optimal parameter estimation. The proposed model and estimation scheme are illustrated with a PBPK model for a rat with Tegafur administration, and are also compared with a conventional least squares method and a MAP estimation method ignoring parameter correlations.

The second study is machine learning based data-driven approach for water pipe system monitoring. This work proposes a diagnosis scheme for partial blockage in water pipeline shown in Figure 1.3. The pressure signal data were collected by generating transient flows with valve opening test, which is easy to implement for artificial leakage in real water pipe network such as hydrant opening test. Using a novel peak search algorithm in the frequency domain, a frequency range of 'fault-characteristic' peak (FC-peak) located in a particular frequency range for each blockage state is identified. With the FC-peak, SVM structure for each blockage case are constructed to detect partial blockage and diagnose its size. Lastly, the performance of proposed scheme is verified with the test data sets from different blockage sizes in the test bed.

The last topic is another data-driven approach using BBN. Although SVM-based approach only used data-driven model, BBN is combination of knowledge-based and data-driven approaches. Therefore, the proposed BBN-based approach can not only use the characteristic from the process outputs but prior knowledge of target system from engineer's intuition and experience. The structure learning algorithm using prior knowledge is proposed to reduce the computation time with SDG which contains prior knowledge of target process. With process data and prior knowledge, edges and likelihood values of BBN are learned. The Root causes of multiple faults probability (RCP) was calculated with BBN and real-time process data. If the process had undesired failures, RCP value can show probability of the faults' root causes. The proposed FDD scheme was verified

with case studies and applied to the batch process monitoring shown in Figure 1.4. For batch process monitoring, because most of FDD research used PCA/PLS based methods, the result from BBN-based approach was compared with traditional PCA-based approach.

All methodologies used for the three proposed approaches were introduced for this research in Chapter 2. To verify the proposed approaches, case studies and simulation conditions were described in Chapter 3. In Chapter 4, simulation results were shown and they were discussed in Chapter 5. In addition, the summary of this research and suggested future works were in the last part of Chapter 5.

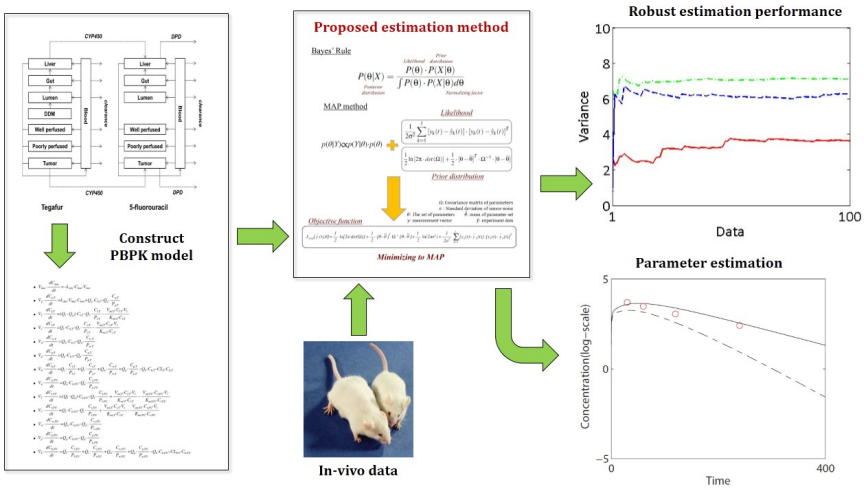

Figure 1.2: Proposed parameter estimation scheme of PBPK model. Based on the Baye's rule, parameters of PBPK model was estimated with in vivo data. The robustness of estimation result was verified with various in vivo data from the simulation.

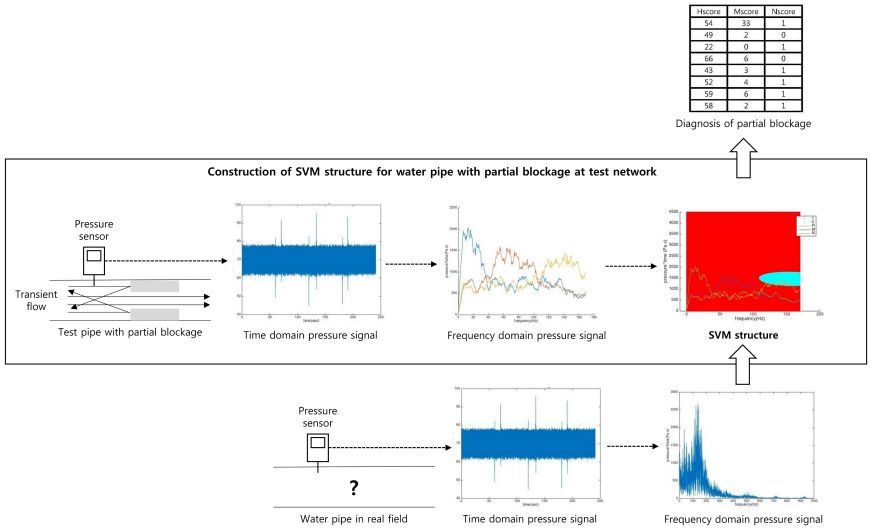

Figure 1.3: Proposed scheme to diagnose a partial blockage in water pipe. To construct SVM classifier for each blockage size, the pressure data from transient flow are obtained in test water pipe network with different sizes of blockage. To identify characteristic of each signal, time domain pressure signal is converted to frequency domain. In frequency domain, each signal has specific peaks for blockage size. With these peaks, SVM classifier for each blockage size is constructed.

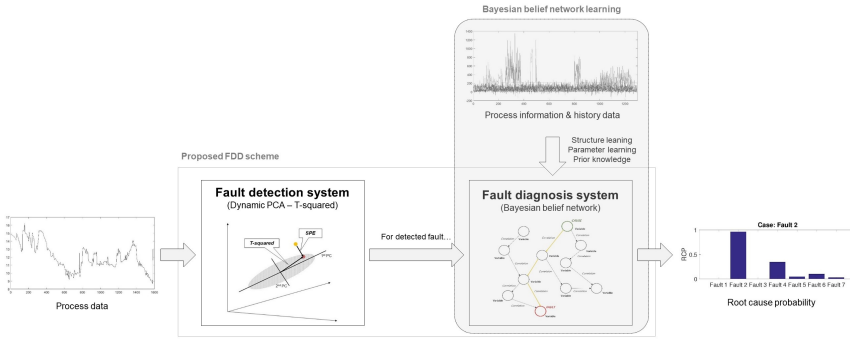

Figure 1.4: Proposed scheme to detect process multiple faults and diagnose their root causes. BBN was constructed by structure and parameter learning with process history data. The proposed scheme is divided two parts, detection system and diagnosis system. Traditional PCA-based T-squared method was used to fault detection with current process data. If the fault was detected by the detection system, fault diagnosis system was activated to find root cause of the detected fault with constructed BBN.

Chapter 2

Methodologies

2.1 Parameter estimation techniques

2.1.1 Least squares method

The least squares method is an estimation technique to identify the parameters that minimize the sum of the squared difference between observations and predictions [79]. The objective function of least squares method is defined as:

$$J_{lse}[\hat{y}_k(t), \theta] = \sum_{k=1}^{l} [y_k(t) - \hat{y}_k(t)]^T \cdot [y_k(t) - \hat{y}_k(t)] \qquad (2.1)$$

where $y_k(t) \in R^{q\times 1}$ and $\hat{y}_k(t) \in R^{q\times 1}$ are the measurements and the model predictions in the k^th organ, respectively, and θ is the parameter vector. When the experimental observations are given, the optimal parameter estimate, $\hat{\theta}$, is determined by minimizing the objective function $J_{lse}[\hat{y}_k(t), \theta]$.

$$\hat{\theta}_{lse} = arg \min_{\theta} J_{lse}[\hat{y}_k(t), \theta] \tag{2.2}$$

The advantage of the least squares method is its simplicity [79] . However, there are no claims about optimality for non-linear models or non-Gaussian residuals, and statistical assessment cannot be performed owing to its deterministic nature [80].

2.1.2 Parameter estimation via maximum a posteriori principle

Bayes' rule provides the conditional probability of event θ, given X [81]:

$$P(\theta \Theta X) = \frac{P(\theta) \cdot P(X|\theta))}{\int P(\theta) \cdot P(X|\theta) d\theta} \tag{2.3}$$

In this equation, $P(\theta)$ is 'the prior distribution,' and $P(X|\theta)$ is the 'likelihood function', describing the conditional probability of X, given θ. $P(\theta|X)$ is the 'posterior distribution', describing the conditional probability of θ, given X. The denominator term is constant, referred to as the 'normalizing factor', which adjusts the maximum value of the posterior probability to be unity [82]. In parameter estimation problems, θ is the vector of the parameters, and X is the given data. Hence, the posterior distribution describes the conditional probability distribution of the parameter vector, given certain observations. Therefore, the parameter vector estimate can be determined

if the posterior distribution is available. $P(\theta|X)$ can be approximated by the product of the likelihood function and prior distribution, since the normalizing factor is constant. The maximum a posteriori (MAP) principle calculates the point estimate of θ by maximizing the product of likelihood and prior distribution:

$$\hat{\theta}_{map} = arg\max_{\theta} P(\theta) \cdot P(X|\theta) \tag{2.4}$$

Because the unbiased measurement noise can be assumed to follow a normal distribution, the likelihood function can be described as follows [82, 83]:

$$p[y;\hat{\theta}] = \prod_{k=1}^{l} \frac{1}{\sqrt{2 \cdot det(\Sigma)}} \cdot \exp\left(\frac{-1}{2 \cdot [y_k(t) - \hat{y}_k(t)]^T \cdot \Sigma^{-1} \cdot [y_k(t) - \hat{y}_k(t)]}\right) \tag{2.5}$$

where Σ is the q-by-q covariance matrix of the estimation error, $y_k(t) - \hat{y}_k(t)$. Since the concentrations in each organ are measured independently, the measurement noises are uncorrelated. Therefore, the off-diagonal entries of the covariance matrix are zero. In addition, because the measurements are from the same type of sensor, the diagonal entries of the covariance matrix can be assumed to have the same value, σ^2. Therefore, the likelihood term can be simplified as:

$$p[y;\hat{\theta}] = \prod_{k=1}^{l} \frac{1}{\sqrt{2\pi \cdot \sigma^{2q}}} \cdot \exp\left(\frac{-1}{(2 \cdot \sigma^2) \cdot [y_k(t) - \hat{y}_k(t)]^T \cdot [y_k(t) - \hat{y}_k(t)]}\right) \quad (2.6)$$

For the priori information term, the probability parameter vector can be assumed to follow the normal distribution, $P(\theta) < N(\bar{\theta}, \Omega)$:

$$p(\theta) = \frac{1}{\sqrt{2\pi \cdot det(\Omega)}} \cdot \exp\left(\frac{-1}{2 \cdot [\theta - \bar{\theta}]^T \cdot \Omega^{-1} \cdot [\theta - \bar{\theta}]}\right) \quad (2.7)$$

where $\bar{\theta}$is the m-by-1 vector of the mean values of parameter estimates, and Ω is the m-by-m covariance matrix of θ. To make the optimization simpler, negative logarithm is taken by the product of (2.6) and (2.7).

$$\begin{aligned} J_{Cov-map}[\hat{y}(t), \theta] = {} & \frac{1}{2} \cdot \ln\left[2\pi \cdot det(\Omega)\right] + \frac{1}{2} \cdot [\theta - \bar{\theta}]^T \cdot \Omega^{-1} \cdot [\theta - \bar{\theta}] \\ & + \frac{l}{2} \cdot \ln\left(2\pi \cdot \sigma^{2q}\right) + \frac{1}{2\sigma^2} \cdot \sum_{k=1}^{l} [y_k(t) - \hat{y}_k(t)]^T \\ & \cdot [y_k(t) - \hat{y}_k(t)] \end{aligned} \quad (2.8)$$

(2.8) is the objective function of the covariance-based MAP method. The off-diagonal entries of Ω cannot be neglected because the parameters of the PBPK model are correlated, and are difficult to calculate with a small number of data sets.

Off-diagonal entries of the covariance matrix represent correlations between model parameters, while the diagonal entries represent the variance of each parameter. The variances can be obtained from the prior distribution of the parameters. Since the parameter correlations are determined by the model structure and are independent of observations or the estimation method used, this study proposes to calculate the off-diagonal entries using a large number of randomly generated data sets, and their corresponding estimates. Based on experimental data, a large number of random data sets can be generated by adding uniformly distributed random noise. With these random data sets, the parameters of the PBPK model can be estimated by the least squares method. From the estimation results, the off-diagonal entries of the covariance matrix can be estimated.

In conventional MAP methods, the off diagonal entries are neglected, and only the variance of each parameter is used. In this case, the objective function becomes:

$$J_{map}[\hat{y}(t),\theta] = \frac{1}{2}\cdot\sum_{i=1}^{m}\ln\left[2\pi\cdot\omega_i\right] + \frac{1}{2}\cdot\prod_{i=1}^{m}\omega_i^{-1}\cdot[\theta-\bar{\theta}]^T\cdot[\theta-\bar{\theta}] + \frac{l}{2}\cdot\ln\left(2\pi\cdot\sigma^{2q}\right) + \frac{1}{2\cdot\sigma^2}\cdot\sum_{k=1}^{l}[y_k(t)-\hat{y}_k(t)]^T \cdot[y_k(t)-\hat{y}_k(t)] \tag{2.9}$$

where ω_i is the variance of the i^{th} parameter.

In cases where there is more than one data set, the previous

estimation results can be used as a priori information for the subsequent estimation. In order to use posterior distribution as a prior distribution for the next estimation, the posterior probability density function needs to be obtained. One can use a Markov Chain Monte Carlo (MCMC) sampling method to find the posterior distribution for small-sized problems [42]. However, since PBPK models often include a large number of parameters, the posterior distribution consists of a large number of joint distributions between each parameter. For example, the Tegafur problem has 12 parameters, which involves 66 joint distributions. Therefore, it is computationally prohibitive to find the parameter set minimizing the posterior distribution. If the posterior distribution follows a normal distribution, the covariance matrix of posterior distribution can be determined without much computational burden since the prior distribution can be assumed to follow a normal distribution. Then, if there is a k^th estimation result $\hat{\theta}_k$ minimizing the posterior distribution, $P_k(\theta|X_k)$, the mean and covariance matrix of the posterior distribution can be calculated as:

$$\bar{\theta}_{k+1}(i) = \sum_{j=1}^{k} \frac{\hat{\theta}_j(i)}{k} \tag{2.10}$$

$$\Omega_{k+1}(i,i) = \sum_{j=1}^{k} \frac{[\hat{\theta}_j(i) - \bar{\theta}_{k+1}(i)]^2}{k-1} \tag{2.11}$$

where $\bar{\theta}_{k+1}(i)$ is the mean of the i^{th} parameter of prior distribution in the $(k+1)^{th}$ estimation, $\hat{\theta}_j(i)$ is the i^{th} parameter of the j^{th} estimation result, and $\Omega_{k+1}(i,i)$ is the (i,i) element of the covariance matrix of the k^{th} posterior distribution, which will be used as a prior distribution in the $(k+1)^{th}$ estimation. The off-diagonal

entries remain the same because they are converged values from simulations. With this information, the equation of the prior distribution in the $(k+1)^{th}$ estimation, $P_{k+1}(\theta)$, can be obtained, and the $(k+1)^{th}$ estimation result minimizing the objective function can also be determined.

2.2 Multivariate analysis methods

2.2.1 Principal component analysis

PCA is a statistical conversion method which uses an orthogonal transformation. A set of observations of correlated variables is converted into a set of values of uncorrelated variables. The set of values of uncorrelated variables called principal component (PC) is used as an axis of new coordinate called latent variables space. The first PC of the orthogonal transformation has the largest variance with the data, and each following PC has the largest variance under the orthogonal condition of previous PC. Therefore, the resulting each PC is a vector which is uncorrelated orthogonal basis set. Let X be a n-by-p data matrix with column-wise zero mean. n rows means system observations and p column means system variables. The orthogonal transformation is defined by a set of p dimensional loading vectors, ω, that map row vectors, x, of X into score vectors, t, given by:

$$t_{k,i} = x_i \cdot \omega_k \tag{2.12}$$

for i is 1 to n and k is 1 to $m < p$. To find first PC which has the

largest variance, the first loading vector satisfies:

$$\omega_1 = arg \max_{\|\omega\|=1} \sum_i t_{1,i}^2 \tag{2.13}$$

$$= arg \max_{\|\omega\|=1} \sum_i (x_i \cdot \omega)^2 \tag{2.14}$$

$$= arg \max_{\|\omega\|=1} \| X\omega \|^2 \tag{2.15}$$

$$= arg \max_{\|\omega\|=1} \| \omega^T X X \omega^T \| \tag{2.16}$$

$$= arg \max \left\{ \frac{\omega^T X^t X \omega}{\omega^T \omega} \right\} \tag{2.17}$$

the (2.17) can be maximized with Rayleigh quotient. Following PCs can be calculated to maximize the variance under the orthogonal condition with the first PC. The k-th PC can be found as:

$$X_k = X - \sum_{j=1}^{k-1} X \omega_j \omega_j^T \tag{2.18}$$

Then, the k-th loading vector which has the maximum variance with X_k can be calculated as:

$$\omega_k = arg \max \left\{ \frac{\omega^T X_k^T X_k \omega}{\omega^T \omega} \right\} \tag{2.19}$$

Finally, full loading matrix and score matrix which columns are loading and score vectors can be calculated and data matrix X can be described by:

$$X = TP^T + E \tag{2.20}$$

where $T = [t_1\ t_2\ \cdots\ t_m]$ is n-by-m score matrix and $P = [\omega_1\ \omega_2\ \cdots\ \omega_m]$ is p-by-m loading matrix. E is n-by-p residual matrix generated from the dimension reduction of X when the X was projected to the PCs' coordination.

2.2.2 Partial least squares

PLS is same idea as PCA, however, the process data matrix was divided into state data matrix X and output data matrix Y. Therefore, loading matrices and score matrices of X and Y can be calculated respectively with same methodology of PCA. Therefore, the result of PLS can be described as:

$$X = TP^T + E \tag{2.21}$$

$$Y = UQ^T + F \tag{2.22}$$

where X is n-by-p state data matrix and Y is n-by-q output data matrix. T is n-by-m score matrix of X, P is p-by-m loading matrix of X and E is n-by-p residual matrix which is the result of dimension reduction ($p \rightarrow m$) of X. U is n-by-l score matrix of Y, P is q-by-l loading matrix of Y and F is n-by-q residual matrix which is also the result of dimension reduction ($q \rightarrow l$) of Y. Because data matrix is

divided into X and Y, PLS can focus on the relation between state variables and output variables which is mostly neglected in with PCA.

2.2.3 Hotelling's T-squared and squared prediction error

Hotelling's T-Squared and squared prediction error (SPE) is the multivariate analysis in projected coordination of PCA/PLS. The graphical meaning of the value of Hotelling's T-squared is Euclidean distance from the origin of PCs to the projected data point. Since every PCs are determined with maximized variance, most of the data points projected around the origin of PCs. Therefore, if a certain data point is projected far from the origin, the data can be abnormal state compared with others. Similarly, the graphical meaning of SPE is Euclidean distance from the PCs' hyperplane to the data point in original coordination. Therefore, if the value of SPE of the data is large, this data point can be considered that it is departed from the normal tendency of the other data. In Figure 2.1, the T-squared and SPE are described in three dimensional space. The T-squared and SPE can be calculated as below:

$$T^2 = x^T P \Lambda_a^{-1} P^T x \tag{2.23}$$

$$SPE = x^T (I - PP^T)^T (I - PP^T) x \tag{2.24}$$

where x is new observation, P is loading matrix of X, I is iden-

tity matrix and Λ_a contains the non-negative real eigenvalues corresponding to the a principal components. The upper limit of T-squared is defined by F-distribution.

$$T^2_{a,n,\alpha} = \frac{a(n-1)}{n-a} F_{(a,n-a,\alpha)} \tag{2.25}$$

where n is the number of samples of the new observation and α is the level of significance of F-distribution. In addition, the upper limit of SPE is also defined by approximate distribution.

$$SPE_\alpha = \theta_1 \left(\frac{h_0 c_\alpha \sqrt{2\theta_2}}{\theta_1} + 1 + \frac{\theta_2 h_0 (h_0 - 1)}{\theta_1^2}\right)^{\frac{1}{h_0}} \tag{2.26}$$

where c_α is the value of the normal distribution with the level of significance α. θ_i and h_0 are calculated below:

$$\theta_i = \sum_{j=a+1}^{m} \lambda_j^i \tag{2.27}$$

$$h_0 = 1 - \frac{2\theta_1 \theta_3}{3{\theta_2}^2} \tag{2.28}$$

where λ_j is j-th largest eigenvalue of data matrix X. If the T-squared and SPE values exceed the upper limit, fault detection of new observations is possible.

In addition, residuals of each variable to the PCA plane and from

the origin of the PCA plane also can be calculated. The sum of residuals of each variable can describe that how much the variable contribute the T-squared and SPE value. The bar plot of the sum of the residuals at certain time called contribution plot [84]. With the contribution plot, the variable which had the largest absolute contribution can be determined as the root cause of the detected fault by T-squared and SPE.

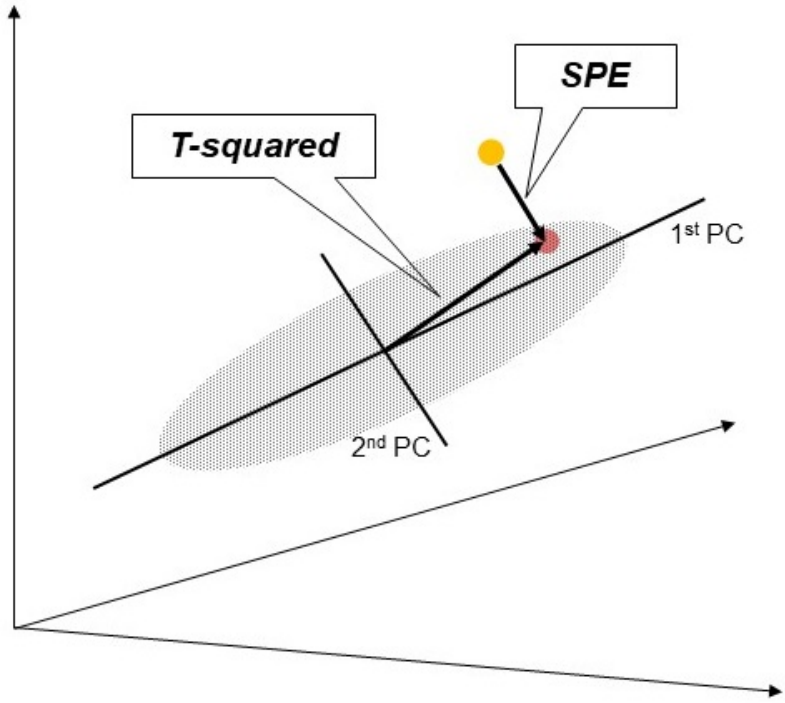

Figure 2.1: Graphcal description of T-squared and SPE value in three dimensional space. Original observation is projected to the PC plane. The distance to the plane is SPE value and the distance from the origin of the PC plane to the projected point is T-squared value. If T-squared or SPE value is larger than upper limit, the new observation can be classified as a fault data.

2.3 Machine learning techniques

2.3.1 Support vector machine

The SVM-based classifier can be constructed by solving the following optimization problem that minimizes the distance between hyperplane and data points. The objective function for calculating SVM hyperplane is:

$$\max \tilde{L}(\alpha) = \{\sum_{i=1}^{n} \alpha_i - 0.5 \sum_{i,j} \alpha_i \alpha_j y_i y_j K(x_i, x_j)\} \tag{2.29}$$

where n is the number of data points, α is the Largrange multiplier such that $\sum_{i=1}^{n} \alpha_i y_i = 0$ and $\alpha_i \geq 0$ for all i, x is the data points, y_i are either 1 or -1 each indicating the class of data point and K is a kernel function. In the objective function, Largrange multiplier and class of data should satisfy the following constraint from Karush-Kuhn-Tucker (KKT) condition.

$$\sum_{i=1}^{n} \alpha_i y_i = 0 \tag{2.30}$$

By solving the objective function, Largrange multiplier for the hyperplane can be determined, and from the α, the parameters for the hyperplane can be obtained as;

$$\omega = \sum_{i=1}^{n} \alpha_i y_i x_i \tag{2.31}$$

$$b = \frac{1}{N_{SV}} \sum_{i=1}^{N_{SV}} \left(K(\omega, x_i) - y_i \right) \tag{2.32}$$

where N_{SV} is the number of all support vectors. To construct SVM, determination of kernel function is necessary. There are some common kernel functions such as linear, polynomial and radial basis. If data points can be separated with linear hyperplanes, a linear kernel function $K(w, x) = b$ can be used. However, the data points of which class determined as 'peak' from the peak search algorithm cannot be separated by linear hyperplane because the peak region is two-dimensional elliptical region in the frequency and amplitude space as in Figure 2.2. Therefore, a nonlinear kernel function should be introduced [85]. The radial basis function, called Gaussian kernel function, is a widely used nonlinear kernel function to construct SVM. Because Gaussian kernel function can separate different class data with nonlinear hyperplanes in the form of hyper-spheres [86], this can be applied to separate the peak data points. The Guassian kernel function is given by;

$$K(x, y) = \exp\left(-\frac{\| x - y \|^2}{2\sigma^2} \right) \tag{2.33}$$

This work employed the 'svmtrain' function available in the Statistics and Machine Learning Toolbox in MATLAB to construct the classifier. When the pressure signal is obtained and converted to the frequency domain, the peaks of each signal are identified by the peak search algorithm. SVM structures can be trained with the information of peak location and the state of the water pipe can be classified by

the trained SVM classifier.

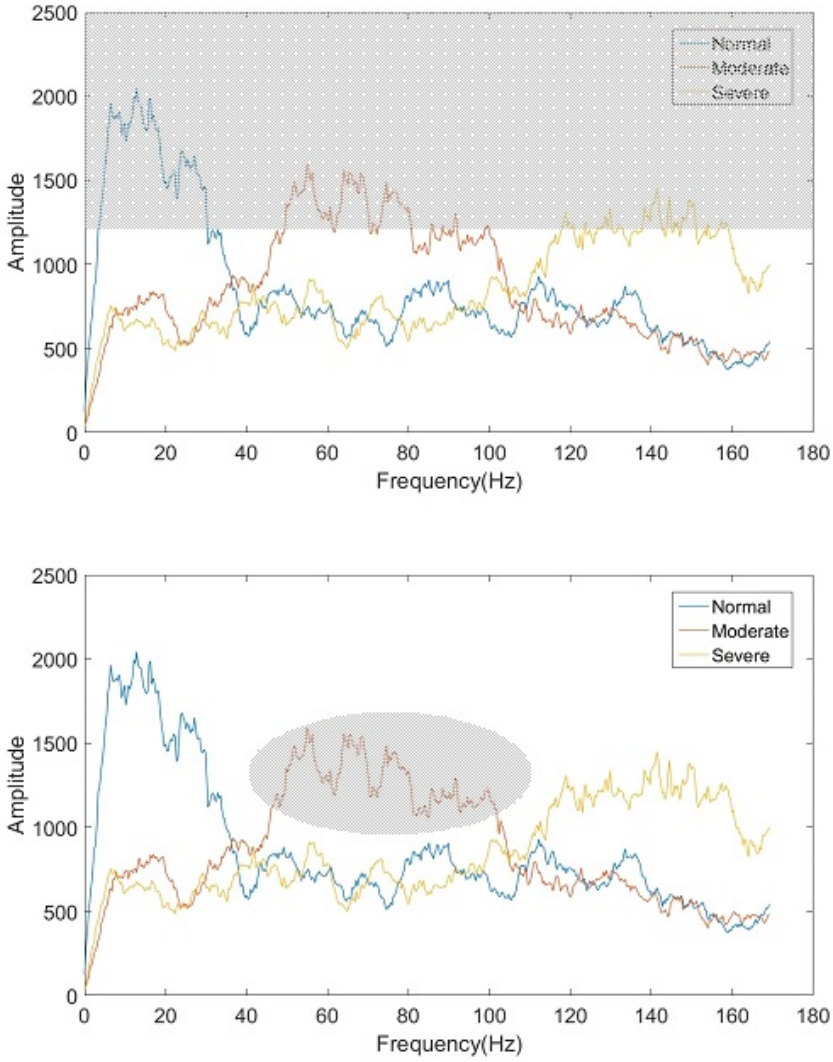

Figure 2.2: Example of peak region separation with linear (upper) and radial basis kernel function (lower). The peak region of orange line cannot be separated from other line. However, spherical SVD structure by nonlinear kernel function can separate peak region of orange line from others effectively.

2.3.2 Bayesian belief network

Bayesian belief network (BBN) is a graphic model combined with graphic theory and stochastic theory [73, 87]. Measured variables are nodes of BBN and the correlation between the nodes is called edge. Figure 2.3 is an example of a BBN. In Figure 2.34, the direction of edge from A to B is described cause and effect. Therefore, The cause A is called 'parent' and effected B is called 'child' of the BBN. A BBN is a triplet $\{G, E, D\}$ where $G = (N, A)$ is a directed acyclic graph with nodes N and edges A. E is a probabilistic space with non-empty space Ω, its subspace Z and probability p where $p(\Omega) = 1$. D is a set of random variables associated to N and defined on E such that:

$$p(N_1, N_2, \cdots, N_n) = \prod_{i=1}^{n} p(V_i|parents(V_i)) \tag{2.34}$$

To build BBN for a target system, the structure of BBN and the parameters of BBN is necessary. The structure is location of edges and those direction between the measurements and the parameter is associated with D which describe a size of causality of each edge. Finding the structure and parameters from the data is called 'structure learning' and 'parameter learning'. For structure learning, the Granger causality is introduced [88]. The Granger causality is statistical hypothesis test that the reliability of prediction for one time series data from the other time series data. If prediction of the data of variable B based on the past data of variable A and on its own past data is better than the prediction of the data of variable B based on the past data, edge and its direction are determined as Figure 2.3. Simi-

larly, if there exist two time series data x and y, x can be predicted based on its past data or on the past data of y and its past data. The former can be described with a restricted model and the later can be described with an unrestricted model as below [89]:

$$x_k = \sum_{i=1}^{p} \mu_i x_{k-i} + \epsilon_{xk} \tag{2.35}$$

$$x_k = \sum_{i=1}^{p} \alpha_i x_{k-i} + \sum_{j=1}^{q} \beta_j y_{k-j} + \eta_{xk} \tag{2.36}$$

where k is the current time, x_{k-i} is the i-lagged value of x, y_{k-j} is the j-lagged value of y. μ, α and β are the model parameters, ϵ_{xk} is the residual at time k of the restricted model and η_{xk} is the residual at time k of the unrestricted model. To determine model order, Akaike information criterion (AIC) or Bayesian information criterion (BIC) are used.

$$AIC(p) = ln(det(\Sigma)) + \frac{2pn^2}{T} \tag{2.37}$$

$$BIC(p) = ln(det(\Sigma)) + \frac{ln(T)pn^2}{T} \tag{2.38}$$

where Σ is the covariance matrix of noise, n is the number of variables and T is the number of observations. The model order is p which maximizes AIC or BIC. If the η_{xk} is significantly smaller than ϵ_{xk}, it means that the prediction of x based on the past data x and y is more accurate. Therefore, since y is said to Granger cause x, y is the

parent node of x in BBN structure. The threshold of residual can be driven from the F-distribution.

However, BBN structure learned with Granger causality is incomplete because the most of process data cannot show the process characteristic perfectly. If y is not Granger cause x, the result is incorrect since the data which describe the causality between x and y can be missing. To overcome the data incompletion, Markov chain Monte Carlo (MCMC) simulation is necessary to converse to the proper BBN structure. Nevertheless, MCMC simulation is difficult to apply the BBN leaning because the structure and parameter learning of BBN is NP-hard problem [90, 91]. The large computation time of NP-hard problem make hard to update BBN when the target process needs frequent operation condition change or short operation time such as the batch process. Therefore, the weighted Granger causality for BBN learning is proposed. From the knowledge-based approach, such as SDG, correlation between the process variables can be determined. From the SDG model, weighting matrix ω for the link between the process variables can be determined. The threshold of Granger causality with weighting matrix can be described below:

$$F = \frac{(w_{xy} RSS_r - RSS_{ur})/q}{RSS_{ur}/(T-p-q-1)} \sim F(p, T-p-q-1) \qquad (2.39)$$

where RSS_r is sum of the squares of the residuals of the restricted model, RSS_{ur} is sum of the squares of the residuals of the unrestricted model and w_{xy} is the entry of weighting matrix of vari-

able x and y. If $w_{xy} > 1$, the correlation between x and y is guaranteed by knowledge-based model, else if $0 < w_{xy} < 1$, x and y is uncorrelated, else if $w_{xy} = 1$, the edge between x and y cannot exist in BBN structure. In this manner, not only the prior knowledge can reflect correlations which are not described in the learning data, but the Granger causality can verify the prior knowledge by finding evidences in the learning data. This work used SDG model to determine weighting matrix for structure learning of BBN.

The BBN edges have its own values which describe the magnitude of causality between parents nodes and child nodes. If the cause and effect probability is described as $p(parents(V_i)|V_i)$, the parameter learning is finding likelihood function $p(V_i|parents(V_i))$. In this thesis, the fault cause and effect was described with normal distribution and the likelihood function of each node is learned by using 'fitdist' function in MATLAB.

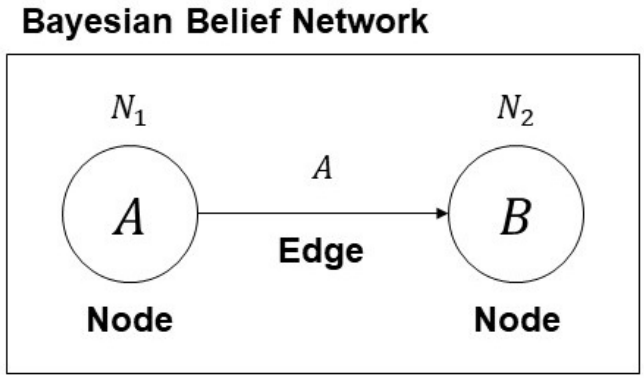

Figure 2.3: Simple example of BBN. Circles A and B are measured variables called 'node', the link between A and B is described the causality between the variables called 'edge' and the form of a directed acyclic graph is called Bayesian belief network.

Chapter 3

Model description & Simulation

3.1 Model-based approach for drug delivery system

3.1.1 Model description

The PBPK model can be constructed by three kinds of differential equations. The first describes the transportation and metabolism of the medicine, based on mass balance:

$$V \cdot \frac{dC}{dt} = Q \cdot (C_{in} - \frac{C}{P}) - R_e \tag{3.1}$$

where V is the organ volume, Q is the volumetric flow rate of blood in the organ, C_{in} is the drug concentration going into the organ, and C is the drug concentration in the organ. P is the tissue/blood partition coefficient of the organ, and describes the proportion of blood volume in the organ. R_e is the consumption term due to the metabolism in the organ, such as degradation, transformation and excretion. Since a drug is transformed by enzymes, transformation metabolism can be described by the Michaelis-Menten equation [92], as follows:

$$v = \frac{V_{max} \cdot C}{K_m + C} \tag{3.2}$$

where v is the reaction rate, V_{max} is the maximum reaction rate, and K_m is the Michaelis-Menten constant, describing the substrate concentration when the reaction rate is $V_{max}/2$. The clearance effect is another component of metabolism, defined as the rate of drug elimination in the blood vessel [93]:

$$\nu_{cl} = CL \cdot C \tag{3.3}$$

where $\nu_c l$ is the rate of elimination, and CL is the kinetic parameter of the clearance effect. The consumption term, R_e, can be expressed with those two equations. The second equation for constructing a PBPK model describes the absorption of medicine in the capillary blood vessel. The drug is absorbed into the organ from the capillary blood vessel, and then circulates around the body. This kind of absorption dynamics can be described as:

$$\frac{dC}{dt} = k_{abs} \cdot C \tag{3.4}$$

where k_{abs} is the absorption rate of the medicine into the organ from the capillary blood vessel.

Finally, inclusion of the drug dissolution model is necessary to develop a PBPK model for orally administered drugs, including Tega-

fur. The following Noyes-Whitney equation can be used for describing the drug dissolution mechanism [83]:

$$\frac{dW}{dt} = \frac{D \cdot A \cdot (C_s - C_b)}{L} \tag{3.5}$$

where W is the mass of the drug, D is the diffusion coefficient, A is the surface area of the drug, L is the diffusion layer thickness, C_s is the drug concentration at the interface between blood and organ, and C_b is the drug concentration in the bulk phase of the blood. W can also be written as:

$$\frac{d(M_w \cdot n)}{dt} = M_w \cdot \frac{dn}{dt} \tag{3.6}$$

where n is the number of moles of drug and M_w is the molecular weight of the drug. To observe the concentration change of Tegafur tablet, the Tegafur is assumed to exist in the interface and dissolved into the bulk phase. Therefore, an arbitrary constant volume of interface, V_d, is introduced. Substitution of (3.6) into (3.5) and multiplication of $1/V_d$ by both hand sides yield:

$$\frac{dC_s}{dt} = \frac{D \cdot A \cdot (C_s - C_b)}{V_d \cdot M_w \cdot L} \tag{3.7}$$

Since the other terms except C_s and C_b are constants, (3.7) can be recast as:

$$\frac{dC_s}{dt} = K_d \cdot (C_s - C_b) \tag{3.8}$$

where K_d is a lumped parameter related to dissolution dynamics.

By describing the dynamics in each organ in Figure 3.1 with these three kinds of equations, the PBPK model, consisting of fourteen differential equations, was developed as follows:

$$\frac{dC_{tab}}{dt} = -K_d \cdot (C_{tab} - C_{lmn}) \tag{3.9}$$

$$\frac{dC_{lmn}}{dt} = K_d \cdot (C_{tab} - C_{lmn}) - C_{lmn} - k_{abs} \cdot C_{lmn} \cdot V_{lmn} \tag{3.10}$$

$$V_g \cdot \frac{dC_{g,T}}{dt} = k_{abs} \cdot V_{lmn} \cdot C_l mn + Q_g \cdot C_{b,T} - Q_g \cdot \frac{C_{g,T}}{P_{g,T}} \tag{3.11}$$

$$\begin{aligned} V_l \cdot \frac{dC_{l,T}}{dt} =& (Q_l - Q_g) \cdot C_{b,T} - Q_l \cdot \frac{C_{l,T}}{P_{l,T}} + Q_g \cdot \frac{C_{g,T}}{P_{g,T}} \\ & - \frac{V_{ml,T} \cdot C_{l,T} \cdot V_l}{K_{ml,T} + C_{l,T}} \end{aligned} \tag{3.12}$$

$$V_t \cdot \frac{dC_{t,T}}{dt} = Q_t \cdot C_{b,T} - Q_t \cdot \frac{C_{t,T}}{P_{t,T}} - \frac{V_{mt,T} \cdot C_{l,T} \cdot V_t}{K_{ml,T} + C_{l,T}} \tag{3.13}$$

$$V_w \cdot \frac{dC_{w,T}}{dt} = Q_w \cdot C_{b,T} - Q_w \cdot \frac{C_{w,T}}{P_{w,T}} \tag{3.14}$$

$$V_p \cdot \frac{dC_{p,T}}{dt} = Q_p \cdot C_{b,T} - Q_p \cdot \frac{C_{p,T}}{P_{p,T}} \tag{3.15}$$

$$\begin{aligned} V_b \cdot \frac{dC_{b,T}}{dt} =& Q_l \cdot \frac{C_{l,T}}{P_{l,T}} + Q_t \cdot \frac{C_{t,T}}{P_{t,T}} + Q_w \cdot \frac{C_{w,T}}{P_{w,T}} + Q_p \cdot \frac{C_{p,T}}{P_{p,T}} \\ & - Q_b \cdot C_{b,T} - CL_T \cdot C_{b,T} \end{aligned} \tag{3.16}$$

$$V_g ß \frac{dC_{g,FU}}{dt} = Q_g \cdot C_{b,FU} - Q_g \cdot \frac{C_{g,FU}}{P_{g,FU}} \tag{3.17}$$

$$\begin{aligned} V_l \cdot \frac{dC_{l,FU}}{dt} =& (Q_l - Q_g) \cdot C_{b,FU} - Q_l \cdot \frac{C_{l,FU}}{P_{l,FU}} + Q_g \cdot \frac{C_{g,FU}}{P_{g,FU}} \\ & + \frac{V_{ml,T} \cdot C_{l,T} \cdot V_l}{K_{ml,T} + C_{l,T}} - \frac{V_{ml,FU} \cdot C_{l,FU} \cdot V_l}{K_{ml,FU} + C_{l,FU}} \end{aligned} \tag{3.18}$$

$$\begin{aligned} V_t \cdot \frac{dC_{t,FU}}{dt} =& Q_t \cdot C_{b,FU} - Q_t \cdot \frac{C_{t,FU}}{P_{t,FU}} + \frac{V_{mt,T} \cdot C_{t,T} \cdot V_t}{K_{mt,T} + C_{t,T}} \\ & - \frac{V_{mt,FU} \cdot C_{t,FU} \cdot V_t}{K_{ml,FU} + C_{t,FU}} \end{aligned} \tag{3.19}$$

$$V_w \cdot \frac{dC_{w,FU}}{dt} = Q_w \cdot C_{b,FU} - Q_w \cdot \frac{C_{w,FU}}{P_{w,FU}} \tag{3.20}$$

$$V_p \cdot \frac{dC_{p,FU}}{dt} = Q_p \cdot C_{b,FU} - Q_p \cdot \frac{C_{p,FU}}{P_{p,FU}} \tag{3.21}$$

$$\begin{aligned} V_b \cdot \frac{dC_{b,FU}}{dt} =& Q_l \cdot \frac{C_{l,FU}}{P_{l,FU}} + Q_t \cdot \frac{C_{t,FU}}{P_{t,FU}} + Q_w \cdot \frac{C_{w,FU}}{P_{w,FU}} \\ &+ Q_p \cdot \frac{C_{p,FU}}{P_{p,FU}} - Q_b \cdot C_{b,FU} - CL_{FU} \cdot C_{b,FU} \end{aligned} \tag{3.22}$$

where C_{tab} and C_{lmn} are the concentrations of Tegafur tablet and drug at lumen, respectively. Other notations are given in Tables 3.1-3.3. The undetermined kinetic parameters are listed in Table 3.1, and the physical parameters are in Tables 3.2 and 3.3. The nominal values for the physical parameters are adapted from [94] and given in Tables 3.4 and 3.5.

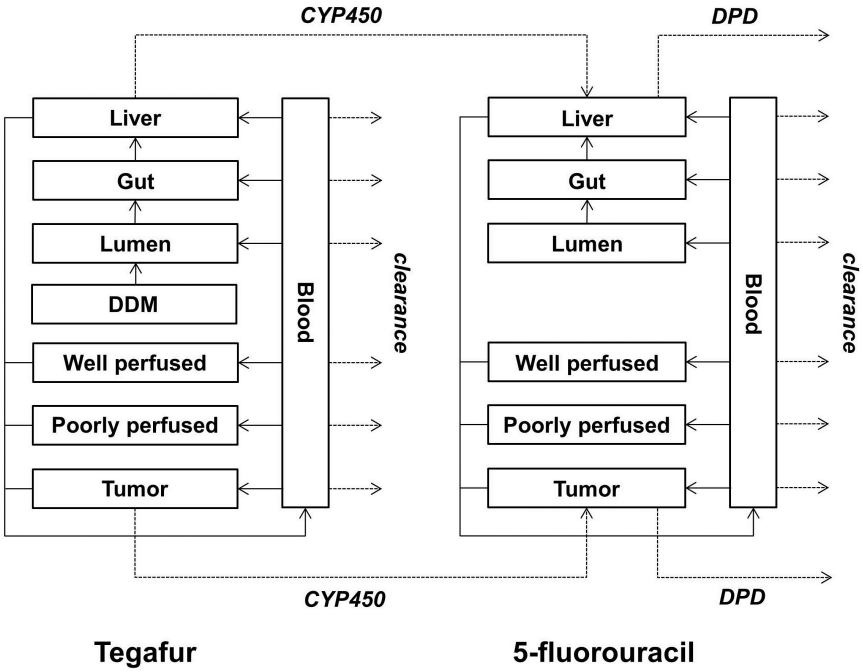

Figure 3.1: The PBPK scheme of Tegafur and 5-fluorouracil. Tegafur is orally administrated and absorbed into the body. The dissolution of Tegafur is expressed by drug dissolution model (DDM). Tegafur is converted to 5-fluorouracil by CYP450 in liver and tumor. The 5-fluorouracil is degraded by DPD in liver and tumor. Both Tegafur and 5-fluorouracil are also cleared in blood vessel.

Table 3.1: Notations of unknown parameters

Parameter	Description
$K_{ml,T}$ (nmol/min/g tissue)	V_{max} for CYP450 enzyme in liver
$V_{ml,T}$ (nmol/ml)	Michaelis-Menten constant for CYP450 enzyme in liver
$K_{mt,T}$ (nmol/min/g tissue)	V_{max} for CYP450 enzyme in tumour
$V_{mt,T}$ (nmol/ml)	Michaelis Menten-constant for CYP450 enzyme in tumour
$K_{ml,FU}$ (nmol/min/g tissue)	V_{max} for DPD enzyme in liver
$V_{ml,FU}$ (nmol/ml)	Michaelis-Menten constant for DPD enzyme in liver
$K_{mt,FU}$ (nmol/min/g tissue)	V_{max} for DPD enzyme in liver
$V_{mt,FU}$ (nmol/ml)	Michaelis Menten-constant for DPD enzyme in tumour
k_{abs} $(\min^{-1})$	Absorption coefficient of Tegafur
K_d $(\min^{-1})$	Dissolution coefficient of Tegafur
CL_T (ml/min)	Clearance rate of Tegafur from plasma
CL_{FU} (ml/min)	Clearance rate of 5-flourouracil from plasma

Table 3.2: Notations for organ volume and blood flow rate

Organ	Organ volume(ml)	Blood flow rate(ml/min)
Blood	V_b	Q_b
Gut	V_g	Q_g
Liver	V_l	Q_l
Tumour	V_t	Q_t
Well perfused organs	V_w	Q_w
Poorly perfused organs	V_p	Q_p

Table 3.3: Notations for tissue/blood partition coefficient

Organ	Tegafur(T)	5-fluorouacil(FU)
Blood	$P_{b,T}$	$P_{b,FU}$
Gut	$P_{g,T}$	$P_{g,FU}$
Liver	$P_{l,T}$	$P_{l,FU}$
Tumour	$P_{t,T}$	$P_{t,FU}$
Well perfused organs	$P_{w,T}$	$P_{w,FU}$
Poorly perfused organs	$P_{p,T}$	$P_{p,FU}$

Table 3.4: Organ volumes and blood volumetric flow rates

Organ	Volume (ml)	Blood flow rate (ml/min)
Blood (V_b, Q_b)	13.2	76.45
Gut (V_g, Q_g)	7.92	17.1
Liver (V_l, Q_l)	8.8	19
Tumour (V_t, Q_t)	1.0	0.25
Well perfused organs (V_w, Q_w)	8.5	38.9
Poorly perfused organs (V_p, Q_p)	165	18.3

Table 3.5: Tissue/blood partition coefficients

Organ	Tegafur (T)	5-fluorouacil (FU)
Blood (P_b)	0.808	0.794
Gut (P_g)	0.768	0.759
Liver (P_l)	0.895	0.5
Tumour (P_t)	0.336	0.169
Well perfused organs (P_w)	0.834	0.826
Poorly perfused organs (P_p)	0.8	0.795

3.1.2 Simulation

800 random data sets were generated based on the real experimental data set adapted from [31]. Random noises from the uniform distribution with the bounds of 50% of the maximum value of the data were added to each data point. The off-diagonal entries of covariance matrix were calculated using these data sets. In addition, 800 parameter estimates were calculated by the least squares estimation in (2.2). From these parameter sets, the covariance matrix was calculated as:

$$\omega_{ij} = \frac{1}{N-1} \cdot (\theta_{ij} - \bar{\theta}_j) \cdot (\theta_{ik} - \bar{\theta}_k) \tag{3.23}$$

where N is the number of parameter estimates, ω_{ij} is the (i, j) entry of the covariance matrix, $\bar{\theta}_j$ is the mean of the j^{th} parameter, and $\bar{\theta}_{ij}$ is the i^{th} estimate of the jth parameter. The off-diagonal entries were used in (2.8), and the diagonal entries were replaced by prior knowledge of the parameters. Since only one in vivo data set was available in this thesis, the prior information was assumed to be the result of simulation.

The 12 parameters of the PBPK model for the Tegafur delivery system were estimated by least squares estimation, covariance-based MAP (Cov-MAP), and variance-based MAP (Var-MAP) methods with 100 random data sets, generated as described above. The Cov-MAP method can use the estimation results from former data sets as prior information for the next estimation step with a new data set. The least squares method does not have an information update step, and the prior variances of parameters in the Var-MAP method

are the same as the diagonal entries of the covariance matrix from simulation.

Finally, with the real in vivo experimental data, the parameters of the PBPK model for the Tegafur delivery system were calculated by the Cov-MAP method. With the parameter estimates, the concentration profile in the gut, liver, tumour cells and blood vessels were predicted. In addition, improvement of the prediction performance by integrating the drug dissolution model is also demonstrated.

3.2 Data-driven approach for water pipe network

3.2.1 Water pipe network

To simulate partial blockage in water pipe network, lattice type test network with ductile cast iron pipes (DCIP) was constructed as shown in Figure 3.2. The length of outer lattice was 44 m, and the inner lattice was 22 m. In this thesis, the pipeline shown in blue in Figure 3.3 was used, and the test pipe network is essentially the same as a single pipeline. The outer diameter of pipe was 100 mm and inner diameter was 80 mm . To supply the continuous water flow, two water tanks that can contain $2\mathrm{m}^3$ of water were installed on the rooftop of the building. Test water pipe network had four pressure sensors and one valve that can generate leakage of water. In Figure 3.3, the pipeline shown in red line can be replaced by another pipeline. Therefore, the test network can simulate blockage of different size by replacing the pipes.

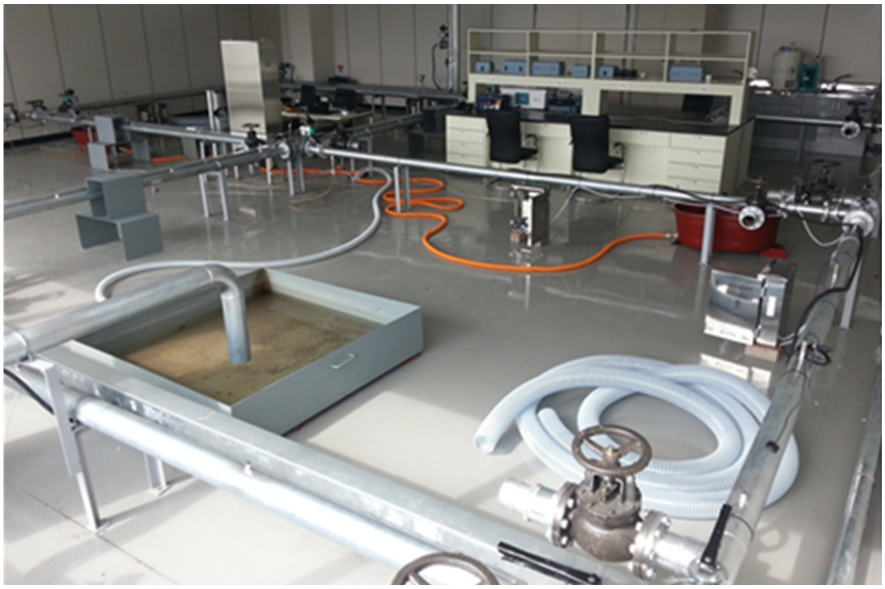

Figure 3.2: Test pipe network for transient test.

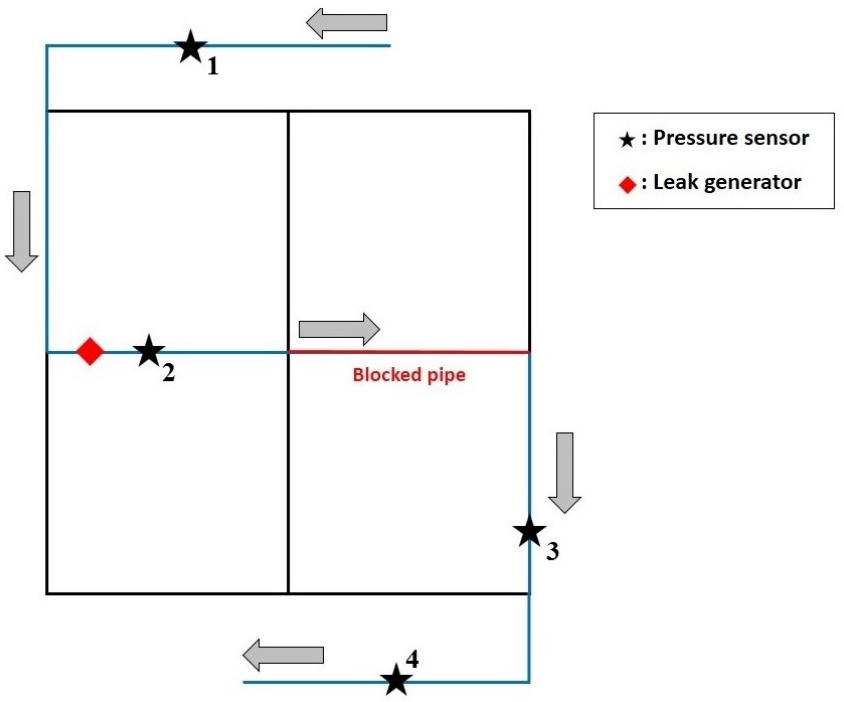

Figure 3.3: The location of pressure sensor and leak generator in test pipe network. The pipe shown in red line can be replaced by another pipe with a different size of blockage.

3.2.2 Experiments & Simulation

The detection of partial blockage and estimation of its size are becoming possible by using the hydraulic head data of the normal and abnormal cases with various sizes of blockage. The disturbance in the pressure signals by the partial blockage can be better characterized in the frequency domain than in the time domain under well-controlled experiments [6, 44, 45]. The oscillatory flows generated from the water hammer is widely used to make a transient flow for the frequency domain analysis [44, 6, 95]. However, generating oscillatory flows is not practical for the real water pipe network because the installation cost is prohibitive and the water hammering can cause serious damage to the pipe and other equipment such as hydrants or pumps [96, 97]. Therefore, a proper method for real water pipe network without installing additional equipment or potential risk of damaging water pipe is necessary.

Artificial leakage can also generate a transient flow because the most of head loss is recovered to the original by the supply pressure to maintain the supply flowrate in real pipeline [98, 99, 100]. This method does not significantly damage pipes compared with water hammering. Artificial leakage is also easy to implement using hydrants installed in the field. However, since the transient flow generated by artificial leakage has only one transition phase, detection of a faulty state is not an easy task, compared with oscillatory flow methods. Therefore, improved water pipe monitoring scheme to identify pipe state and its blockage size with noisy and irregular signals generated by artificial leakage in real water pipe needs to be developed.

Three different blocked pipes were used in the experiment. Figure 3.4 shows the cross sectional areas of blockage. To incur transient flow, artificial leakage was generated by opening the valve. The pressure signal was collected from the pressure sensors with the sampling frequency of 1000 Hz. The pressure signal from sensor 2 was used to estimate blockage size. Although the other sensor data can be used to estimate the location of pipe fault [101], only the sensor 2 data were used because this study is more focused on the diagnosis of blockage size. The tests were repeated 8 times for each blockage size. In addition, the same tests were repeated for each blockage size to verify the detection performance of the proposed scheme.

If there exists partial blockage inside the pipe, the pressure signal will show different trends compared with that of the normal pipe because of the pressure waves reflecting from the partial blockage as shown in Figure 1.3. However, since the pressure signal is generally of high frequency and contaminated with measurement noise, the pressure differences between the normal and partially-blocked pipes are difficult to distinguish in the time domain. Therefore, the time domain data are converted to the frequency domain signals.

Fourier transform has been widely used for frequency domain analysis in fault detection of water pipe network [102, 103, 104]. In Figure 1.3, small differences of the pressure signal from blocked pipe are difficult to distinguish in the time domain with significant measurement noise. However, conversion of the signal to the frequency domain using Fourier transform can reveal the peaks at relevant fre-

quencies, whose ranges are not shifted by measurement noise. Therefore, we propose to apply fast Fourier transform to the pressure signals obtained from transient test to detect the partial blockage inside the pipe and estimate the size of blockage in the radial direction of a pipe.

In frequency domain, a large amplitude at a particular frequency range for each blockage size should be distinguished. We refer to the peaks occurring at this particular frequency range for each blockage size as FC-peak for a particular blockage size. The high-frequency components of the signal and measurement noise make it difficult to distinguish the FC-peaks from other signals of large amplitude. In Figure 3.5, there are signal components with large amplitude in certain frequency ranges for each blockage size. For accurate diagnosis, the FC-peak should be distinguished with objective and quantitative criteria. Therefore, a peak detection algorithm is necessary to identify the FC-peaks for each blockage size. There exist several algorithms for peak search. Because frequency domain pressure signal is noisy, robustness is also required for the search algorithm. This study employs the value F which averages the average distances from left and right neighbours as in (3.24) [105].

$$F = \frac{\frac{(x_i - x_{i-1} + x_i - x_{i-2} + \cdots + x_i - x_{i-k})}{k} + \frac{(x_i - x_{i+1} + x_i - x_{i+2} + \cdots + x_i - x_{i+k})}{k}}{2} \tag{3.24}$$

Let X be a vector of amplitudes of pressure signal in the frequency domain. X was converted from the pressure signal sampled

in the time domain by fast Fourier transform. Let x_i be the i^{th} element of X. To determine if x_i is a relevant peak or not, the average of differences between x_i and k-left neighbouring points (left-differences average) and the average of differences between x_i and k right neighbouring points (right-differences average) are used in (3.24).

In Figure 3.5, the amplitudes of frequency domain signal have high values within a certain frequency range in all the cases. However, although the averaging filters out measurement noise in the data and thus ignores 'noisy peaks', the algorithm can still have a possibility of wrong identification when the 'noisy peak' does not have enough signal amplitude which is located out of the certain frequency range. Therefore, this work introduces the minimum amplitude of FC-peak, f. If the average of left and right differences averages is larger than a threshold value, h, and x_i is larger than the minimum amplitude of FC-peak, x_i is determined as a peak. The peak detection algorithm is summarized in the following:

1. Determine the threshold value, h, and minimum amplitude of peak, f
2. Calculate F for x_i in (3.24).
3. If $F > h$ and $x_i > f$, x_i is determined as a peak. Otherwise, x_i is not a peak.
4. Set $i = i + 1$ and return to step 2.

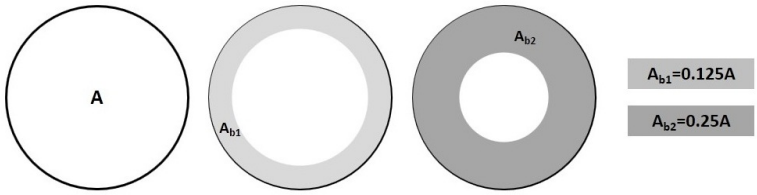

Figure 3.4: Cross-sectional view of pipes with different blockage size. Normal pipe had no blockage with the cross-sectional area of A.

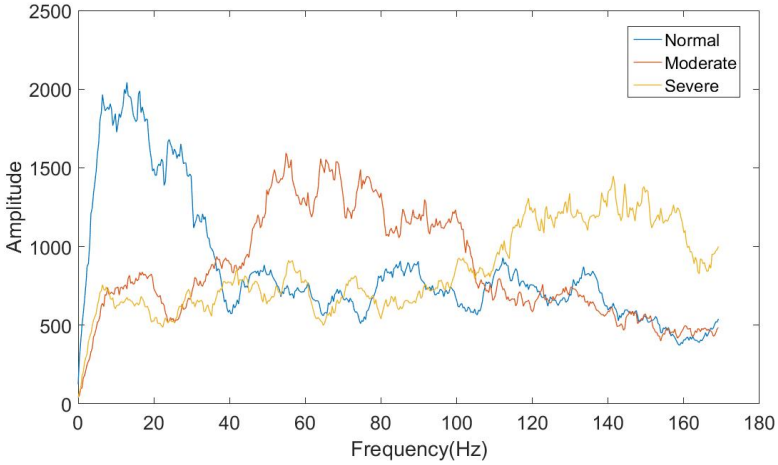

Figure 3.5: Frequency domain signals for different blockage sizes. The blue signal is for the frequency domain signal from normal pipe, the red from the blocked pipe of size $0.125A$ and the yellow from that of size $0.25A$. The high amplitude region for each signal was identified over different frequency ranges.

3.3 Data-driven approach using Bayesian network

3.3.1 Continuous stirred-tank reactors

3.3.1.1 Model description

CSTR process have been widely used in chemical industry. Traditionally, FDD of CSTR used knowledge-based model and multivariate analysis [106, 107]. Therefore, BBN-based approach was applied to example CSTR model and its results was compared with PCA-based approach which is typical FDD method used for CSTR process. Four CSTRs are used for an example system to verify the FDD performance of the proposed BBN-based approach. Process flow diagram of the example system is in Figure 3.6. Level of each CSTR is controlled by PI controller and reactions in each CSTR are described below:

- CSTR 1
  - $A + B = C + D$ with $r_1 = k_1 C_A C_B$
- CSTR 2
  - $A + B = C + D$ with $r_1 = k_1 C_A C_B$
  - $C + E = G$ with $r_2 = k_2 C_C C_E^2$
- CSTR 3
  - $A + B = C + D$ with $r_1 = k_1 C_A C_B$
  - $D + F = A$ with $r_3 = k_3 C_D C_F$
- CSTR 4

- $A + B = C + D$ with $r_1 = k_1 C_A C_B$
- $C + E = G$ with $r_2 = k_2 C_C C_E^2$
- $D + F = A$ with $r_3 = k_3 C_D C_F$

where G is the product, r_i is the reaction rate, k_i is the reaction constant and C_j is the concentration of j material. The concentration of G is off-line measurement observed in mid-product flow and final product flow. With the process model and various fault scenarios, time domain data of measurements are collected. The measurements of example process are in Table 3.6.

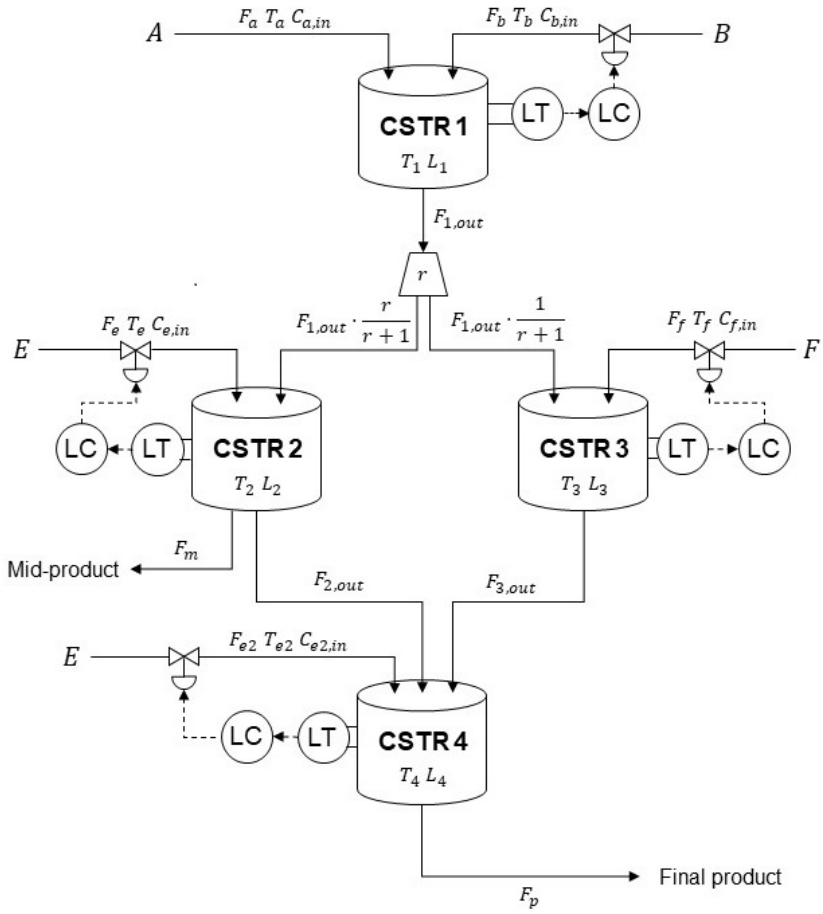

Figure 3.6: Example system for FDD approach with Bayesian network.

Table 3.6: On-line and off-line measurements the example system in Figure 3.6

On-line measurements	Description
T_1	Temperature in CSTR 1
L_1	Level in CSTR 1
T_2	Temperature in CSTR 2
L_2	Level in CSTR 2
T_3	Temperature in CSTR 3
L_3	Level in CSTR 3
T_4	Temperature in CSTR 4
L_4	Level in CSTR 4
Off-line measurements	**Description**
$C_{p,mid}$	Concentration of the product G in mid-product flow
C_p	Concentration of the product G in final product flow

3.3.1.2 Simulation

To simulate fault condition of the example system, five fault scenarios were defined as below:

- Fault 1: Leakage was generated in CSTR 1.
- Fault 2: Failure of level transmitter disturbed to observe accurate tank level of CSTR 1.
- Fault 3: Undesired heat flux was generated into the CSTR 2 that increase the temperature.
- Fault 4: Leakage was generated in CSTR 3.
- Fault 5: Undesired heat flux was generated into the CSTR 4 that decrease the temperature.

The CSTRs model and fault scenarios were built in MATLAB and the model was simulated with each fault scenario. The data sets for structure and parameters learning of BBN are generated with five step input changes. Every fault scenarios were activated respectively to collect single fault case data. In addition, to collect multiple faults data sets, two kinds of faults are activated at the same time. Therefore, total 15 fault data sets were generated and the magnitude of fault was randomly determined for each simulation. 30 simulation was conducted for each fault case and both of PCA-based and BBN-based fault diagnosis methods were applied to find root causes of the faults.

3.3.2 Wet gas compressor

3.3.2.1 Model description

Wet gas compressor is widely used process in chemical industry. To apply the proposed FDD scheme, dynamic model of WGC was built using Aspen HYSYS. Simple description about the process is in Figure 3.7. Aspen HYSYS dynamic model of WGC was built to follow the reference sensor data from a real refinery plant. To maintain the limitation of the equipment settlement in the real plant, only seven temperature sensors, which are also installed in the plant, measured the real time temperature data in Table 3.7. With the seven measurements, BBN-based FDD scheme was applied to the example WGC process under different kinds of fault scenarios.

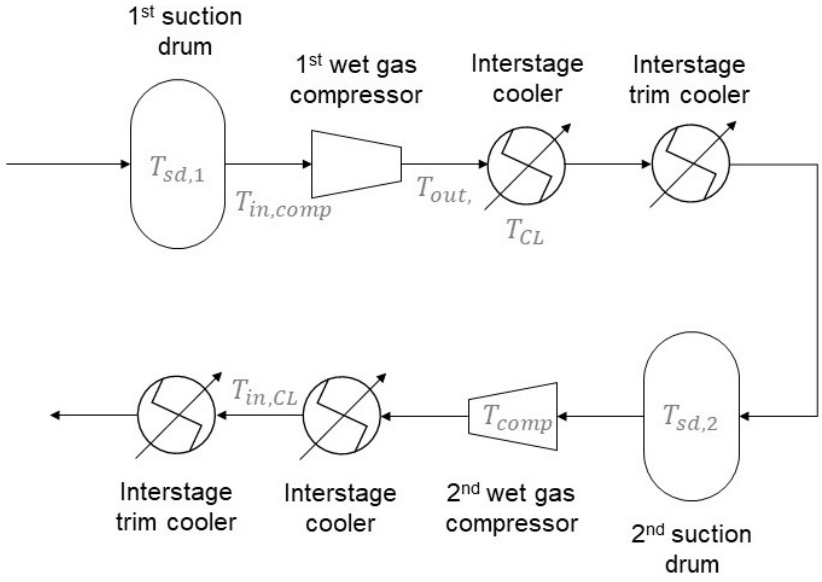

Figure 3.7: Process description of example Wet gas compressor system. Eight different units consist the process and seven measurements exist.

Table 3.7: On-line measurements of the WGC process.

On-line measurements	Description
$T_{sd,1}$	Temperature in the first suction drum
$T_{in,comp}$	Temperature of inlet flow to the first WGC
T_{out}	Temperature of outlet flow from the first WGC
T_{CL}	Temperature in the first interstage cooler
$T_{sd,2}$	Temperature in the second suction drum
T_{comp}	Temperature in the second WGC
$T_{in,CL}$	Temperature of outlet flow from the second interstage cooler

3.3.2.2 Simulation

The learning data for structure and parameter learning of BBN were generated from the Aspen HYSYS dynamic model of WGC with 5 different input compositions. To describe real process data, 5 signal-to-ratio white Gaussian noise was added to the learning, normal and fault data. The normal stated data was used to construct PCA model for fault detection system.

To simulate fault condition of the example WGC process, four fault scenarios, which were interesting issues in the real refinery plant, were defined as below:

- Fault 1: Pressure is decreased in the first suction drum because of crack or air leakage of suction drum.
- Fault 2: Efficiency of the second WGC is decreased because of unknown failure.
- Fault 3: Pressure is decreased in the second suction drum because of crack or air leakage of suction drum.
- Fault 4: Efficiency of the first WGC is decreased because of unknown failure.

From the four single fault scenarios, six multiple fault scenarios were generated and total ten fault scenarios were used to verify the performance of the proposed method. The traditional PCA-based method also used for fault diagnosis to compare the results. Two fault diagnosis methods were applied to each fault scenario with

pre-defined fault size.

3.3.3 Penicillin batch process

3.3.3.1 Model description

The batch process model is one of the challenging problem of process monitoring because of it nonlinearity. Moreover, the real-time monitoring of bioreactors is difficult since most of system statements are off-line measurements. Therefore, FDD scheme for batch process with limited observations is necessary.

Goldrick et al. developed example batch process model which produce penicillin for computer simulation [108]. The model has eight input variables, three on-line measurements and three off-line measurements. The product is Penicillin and its concentration can be measured in off-line. The model description is in Figure 3.8 and parameters of the batch process are in Table 3.8. PID controllers were used to control the batch temperature and pH values. Manipulated variable of temperature controller was cooling water flow rate, F_c, and manipulated variables of pH controller were acid and based flow rate, F_a and F_b. In real situation, off-line measurements were measured by discrete sampling points. Therefore, those variables were neglected because the proposed scheme is used for on-line monitoring.

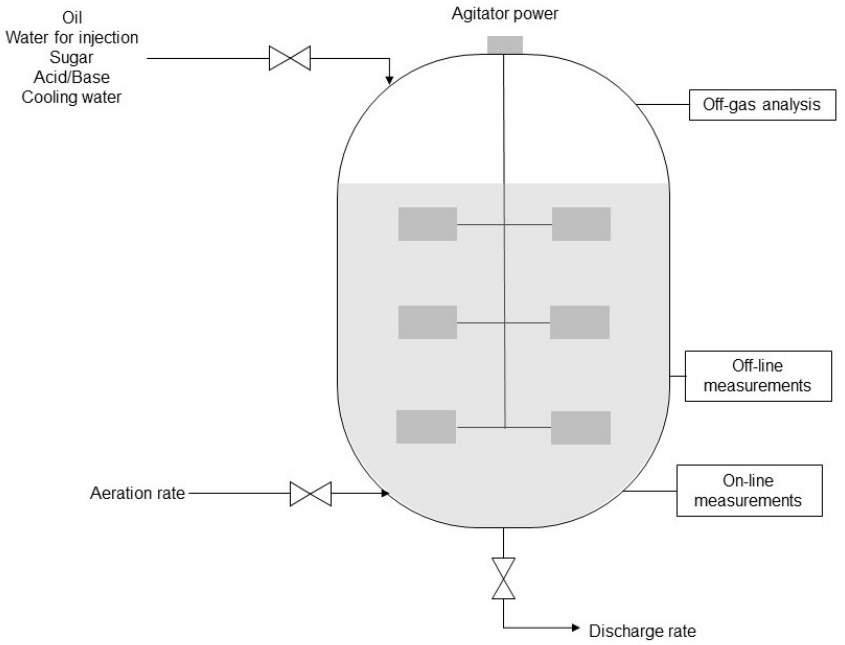

Figure 3.8: Process description of Penicillin batch process model.

Table 3.8: Input variables, On-line and off-line measurements the Penicillin batch process.

Input variables	Description
F_{oil}	Oil flow rate
F_w	Water for injection flow rate
F_s	Sugar flow rate
F_a	Acid flow rate
F_b	Base flow rate
F_c	Cooling water flow rate
F_g	Aeration rate
RPM	agitator rpm
On-line measurements	Description
T	Temperature in the batch reactor
pH	pH in the batch reactor
DO_2	Dissolved oxygen concentration in the batch reactor
Off-line measurements	Description
$C_{p,mid}$	Penicillin concentration
C_p	Phenylacetic acid concentration
$C_{p,mid}$	NH_3 concentration

3.3.3.2 Simulation

Seven different fault scenarios were defined in the Penicillin batch process model [108] described below:

- Fault 1: Aeration rate fault.
- Fault 2: Vessel back pressure fault.
- Fault 3: Substrate feed rate fault.
- Fault 4: Base flow rate fault.
- Fault 5: Coolant flow rate fault.
- Fault 6: Temperature sensor error.
- Fault 7: pH sensor error.

With seven fault scenarios, eight multiple faults scenarios were defined with possible combinations of seven single faults. The other combinations for multiple faults scenarios cannot be used because the MATLAB model cannot be conversed when the other scenarios were applied.

To collect learning data for BBN, various sizes of step inputs were necessary. However, because F_a, F_b and F_c were manipulated variable, those values cannot changed arbitrary. Therefore, 5 signal-to-noise ratio white Gaussian noise was added to set point of temperature and pH. The other input variables had 5 random size step changes to collect learning data. Based on the MATLAB model, the normal stated data was generated to construct PCA model for fault detection

system. In addition, 50 simulations for each fault scenario with random fault size were conducted to verify the FDD performance of the proposed method and the traditional PCA-based method was used for comparison.

Chapter 4

Simulation results

4.1 Robust parameter estimation for drug delivery system

Robustness refers to the ability of a parameter estimation method to be "insensitive" to measurement noise or other unknown disturbances. While a large number of data will reduce the variance with convergence, an increased number of sampled data less than a certain value withdrawn from uniform distribution can show a larger variance owing to the increasing randomness in the data. In this thesis, the variance kept increasing if the number of data points was less than 100. We would like to note that the proposed method aims at estimating parameters robustly under a small data set. Hence, we limit the number of data points less than 100. For this reason, the unknown parameters were estimated with 100 random data sets. For each estimation scheme, the variance of the parameters was calculated and compared. The variances changed when the estimation scheme was repeated with new data sets and showed different trends for each parameter. If a parameter estimate shows a smaller variance, the robustness of the estimation method is better than the others.

For the parameters in Figure 4.1, there were no significant differences between the estimation methods. However, for the parameters in Figure 4.2, the variances of the parameter estimates using the Cov-MAP method were smaller than those of the other methods. The benefit of the Cov-MAP method was not observed for the two parameters in Figure 4.3.

The predicted concentration profiles in each organ are shown in Figures 4.4-4.5, using the Cov-MAP estimation scheme. The solid line represents the proposed PBPK model, the dashed line represents the PBPK model without inclusion of drug dissolution dynamics, and the circles show in vivo experimental data. Figures 4.4-4.5 demonstrate that the results of the PBPK model including drug dissolution model (DDM) were closer to the experimental observations. The mean squared error of the proposed PBPK model was 1.817, and that of the model without drug dissolution dynamics was 6.765 in the log scale. Moreover, Table 4.1 shows that the mean squared error of the proposed PBPK model and estimation scheme was less than that of the conventional approaches for the concentration in every measured organ. This implies that the model structure including DDM describes experimental PBPK data more accurately than the same model without DDM.

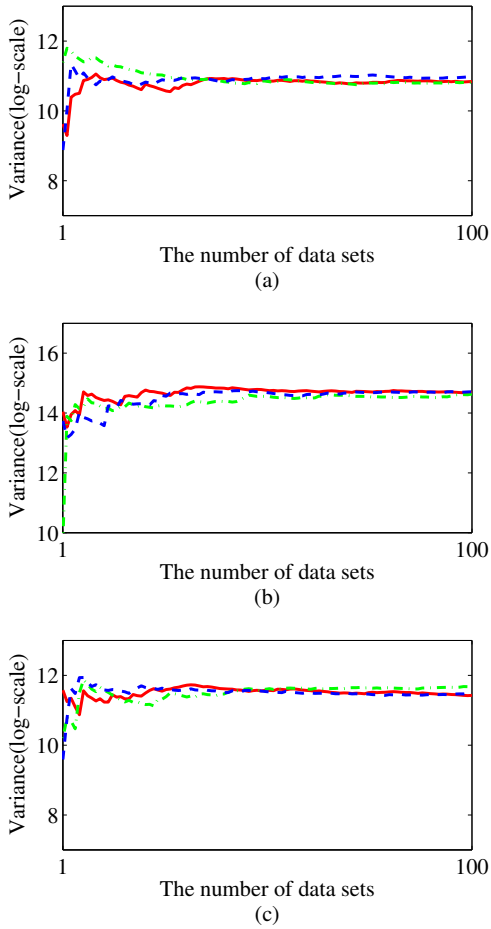

Figure 4.1: Changes of the variance of parameter value observed when the estimation scheme was repeated with the number of new data sets. The red solid line is the result with Cov-MAP method, the green dash-dot line is the result with least squares method, and the blue dashed line is the result with Var-MAP method. No significant differences were observed between the three estimation methods for these parameters. (a) Diffusion coefficient, K_d, (b) K_m value of CYP450 enzyme in liver cell, $K_{ml,T}$, and (c) K_m value of DPD enzyme in liver cell, $K_{ml,FU}$.

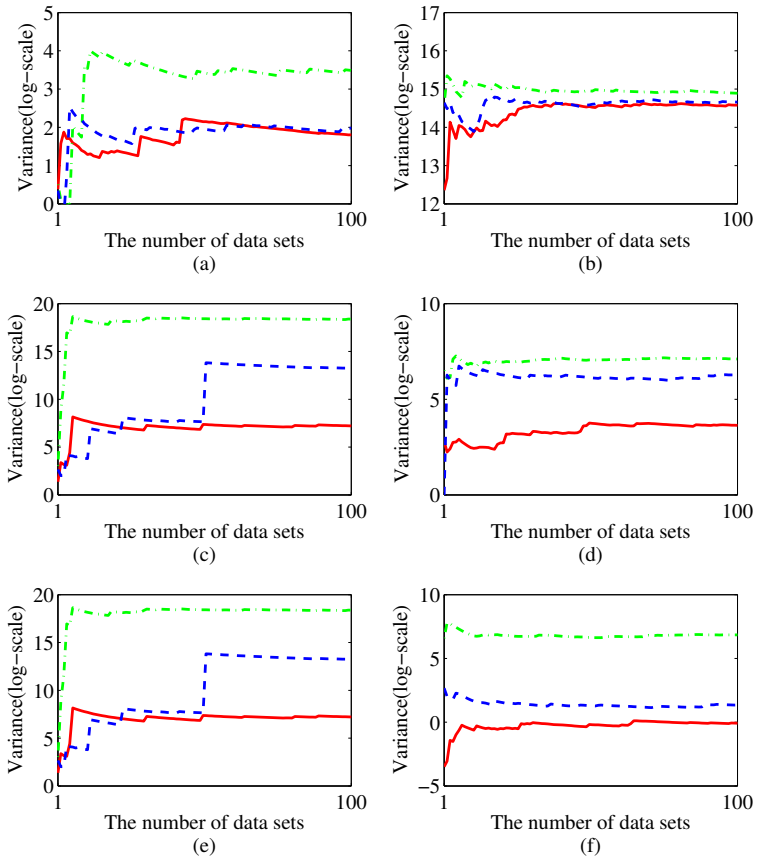

Figure 4.2: Changes of the variance of parameter value observed when the estimation scheme was repeated with the number of new data sets. The red solid line is the result with Cov-MAP method, the green dash-dot line is the result with least squares method, and the blue dashed line is the result with Var-MAP method. The Cov-MAP method was found to perform the most robust estimation for these parameters: (a) The absorption coefficient, k_abs, (b) the V_m value of CYP450 enzyme in tumour cells, $V_{mt,T}$, (c) the K_m value of CYP450 enzyme in tumour cells, $K_{mt,T}$, (d) the V_m value of DPD enzyme in tumour cells, $V_{mt,FU}$, (e) the K_m value of DPD enzyme in tumour cells, $K_{mt,FU}$ and (f) the clearance rate of 5-fluorouracil, CL_{FU}.

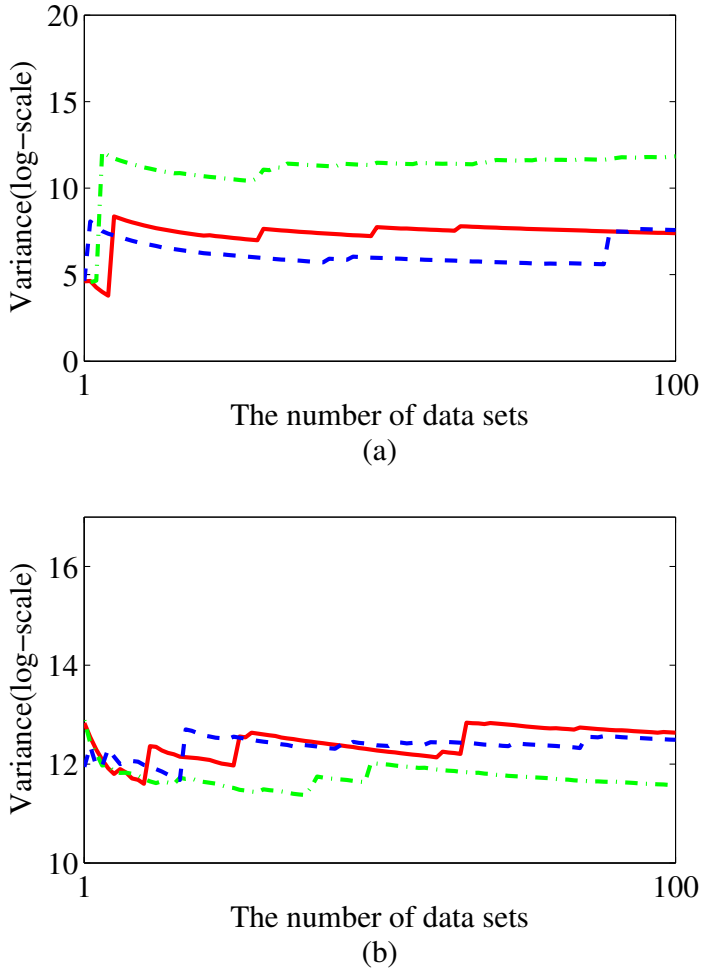

Figure 4.3: Changes of the variance of parameter value observed when the estimation scheme was repeated with the number of new data sets. The red solid line is the result with Cov-MAP method, the green dash-dot line is the result with least squares method, and the blue dashed line is the result with Var-MAP method. For these parameters, the results of the Cov-MAP method were less robust than the other estimation methods. (a) The V_m value of DPD enzyme in liver cells, $V_{ml,FU}$, and (b) the V_m value of CYP450 enzyme in liver cells, $V_{ml,T}$.

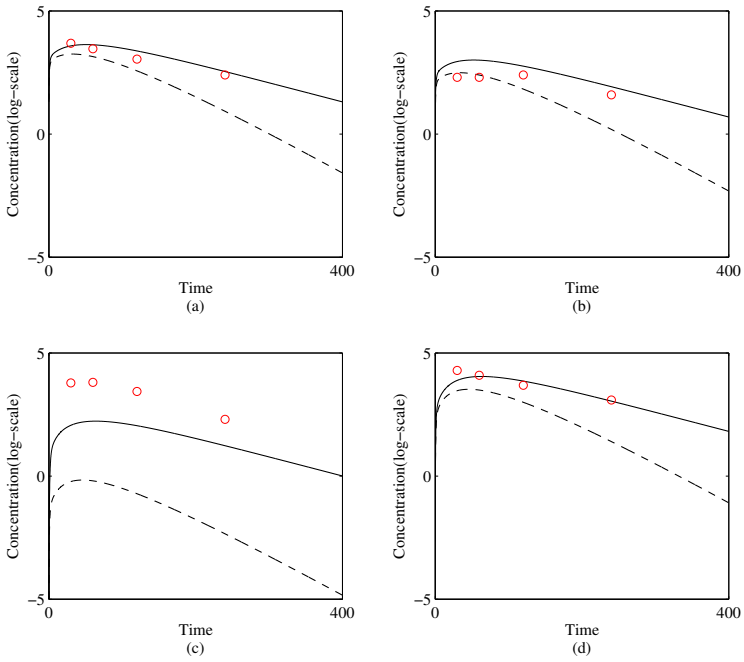

Figure 4.4: Estimated Tegafur concentration profile. (a) is the estimated Tegafur concentration at gut, (b) is at liver, (c) is at tumour, and (d) is at blood.

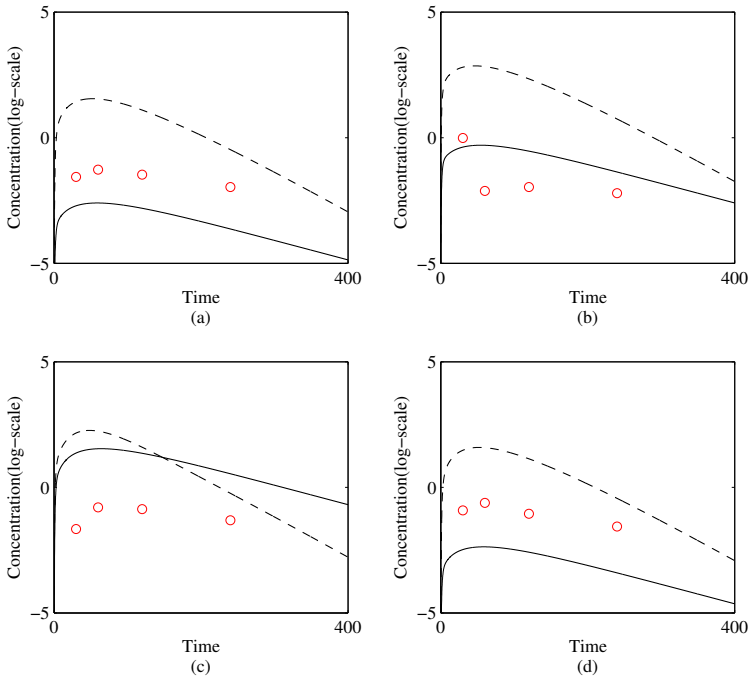

Figure 4.5: Estimated 5-fluorouracil concentration profile. (a) is the estimated 5-fluorouracil concentration at gut, (b) is at liver, (c) is at tumour, and (d) is at blood.

Table 4.1: The log scaled mean squared error of the estimation results in each organ

	Proposed PBPK model	Conventional PBPK model
Tegafur in gut	0.042	0.674
Tegafur in liver	0.285	0.582
Tegafur in tumour	2.131	17.719
Tegafur in blood	0.049	1.093
5-fluorouacil in gut	1.915	6.428
5-fluorouacil in liver	1.548	14.917
5-fluorouacil in tumour	1.748	7.767
5-fluorouacil in blood	2.814	4.937

4.2 Diagnosis of partial blockage in water pipe network

Three experimental pressure data sets were obtained for each blockage size. Therefore, total 9 sets were used to construct the SVM classifier. The time domain pressure data were converted into frequency domain using fast Fourier transform. Figure 3.5 shows the frequency domain data for each blockage size. Each blockage shows its own distinctive region in the amplitude space. In order to identify the region of distinctive and significant amplitudes for each blockage size, the proposed peak search algorithm was applied. Because, in Figure 3.5, the amplitude larger than 1200 is distinguishable for each blockage case, f was set to be 1200 in the proposed algorithm. In addition, to choose h, standard points which has suitable to be classified as peaks was selected and the value of h was determined to classify every standard points as peaks. Typically, standard points can be determined as $u_F + 1 < F_i < u_F + 3$ where F_i was F value of each standard point and u_F was mean of all F values of signal in (3.24). In this case, the user-defined values of k and h were set as 15 and 700, respectively. Figure 4.6 shows the result of applying the peak search algorithm. The peak of normal pipe is located in low frequency region whereas that of the pipe with severe blockage is located in the relatively high frequency region. The peak of blocked pipe with moderate blockage is located between the frequency region of normal and severely blocked pipes.

To construct SVM classifier for each peak, the class of each peak was specified as class 1 and the class of other data points was deter-

mined as class 2. Only the identified peaks were used for the training of SVM classifier. In Figure 4.7, the shaded regions represent class 1 for each blockage by the trained SVM classifier. Therefore, if a pressure data set to be diagnosed is obtained, the blockage size can be estimated by the number of class 1 data points classified by the SVM classifier. If the SVM classifier for normal pipe yields the largest number of class 1 for a given data set, the pipe is then diagnosed to be at the normal state. To verify the diagnosis performance of proposed scheme, other 24 data sets were obtained from the pipes of different blockage size. Table 1 shows the results of size estimation. H-score means how many data points are located in the SVM classifier of Figure 4.7-(c). M-score means how many data points are located in the SVM classifier of Figure 4.7-(b) and N-score means how many data points are located in the SVM classifier of Figure 4.7-(a). H is the state with $0.25A$ blockage size, M is the state with the blockage size of $0.125A$ blockage size, N is the normal state, F means failed diagnosis and S means successful diagnosis. Data sets 1 to 8 were obtained from the blocked pipeline with the blockage size of $0.25A$, 9 to 16 were obtained from the blocked pipeline with the blockage size of $0.125A$ and 17 to 24 were obtained from normal pipeline. In Table 4.2, where 24 data sets were diagnosed, blockage sizes of 20 data sets were diagnosed successfully. Only 4 data sets failed to diagnose correctly.

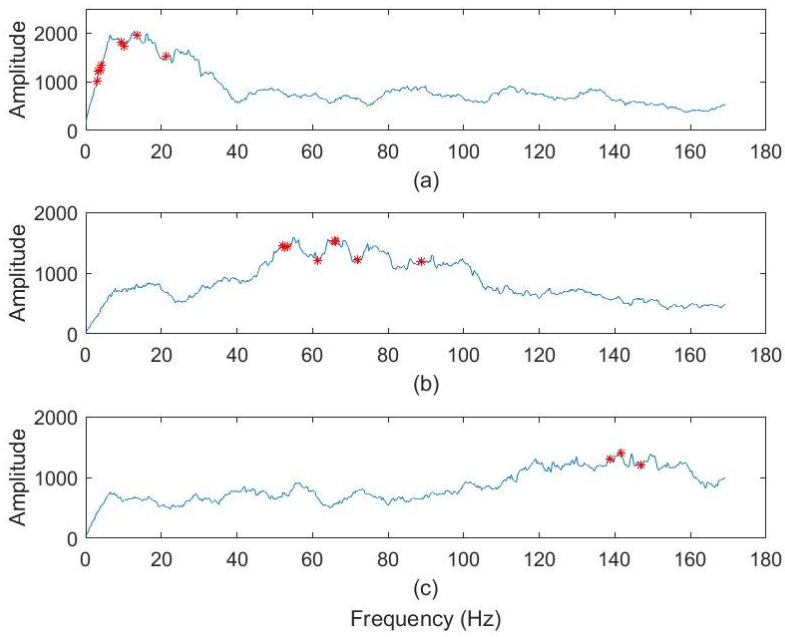

Figure 4.6: FC-peaks for each blockage case identified by the peak search algorithm. The red dots indicated the location of peaks in each signal. (a) is the frequency domain signal of normal state pipe, (b) is the signal of moderate blocked pipe with $0.125A$ blockage size and (c) is the signal of severe blocked pipe with $0.25A$ blockage size.

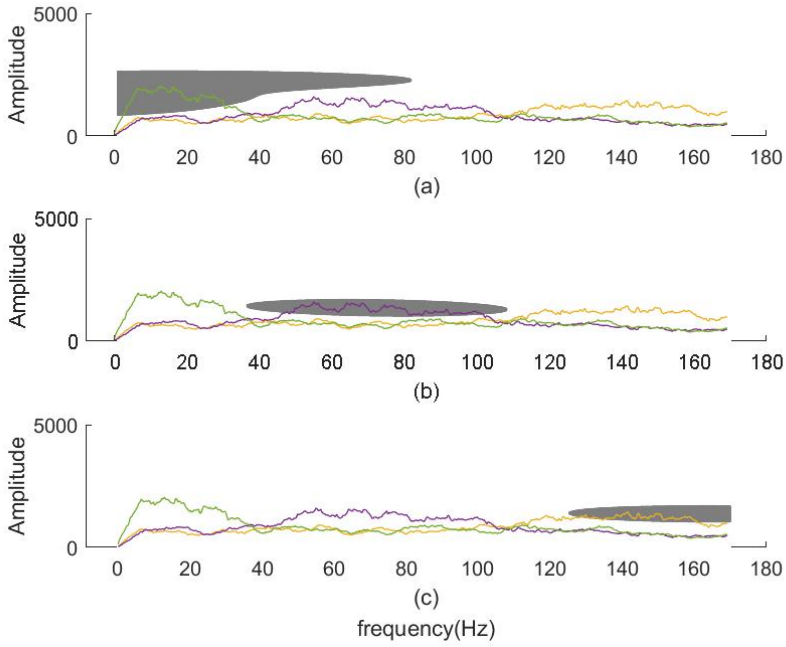

Figure 4.7: Trained SVM structure by frequency domain signal from test water pipe network. The gray area in (a) is SVM structure for normal pipe. The area in (b) is SVM structure for blocked pipe with $0.125A$ blockage. The area in (c) is SVM structure for blocked pipe with $0.25A$ blockage.

Table 4.2: SVM scores of 24 pressure data sets with trained SVM structure

Data set	H-score	M-score	N-score	Diagnosis	Result
1	10	35	0	M	F
2	79	0	0	H	S
3	71	0	19	H	S
4	82	0	0	H	S
5	85	0	0	H	S
6	96	0	0	H	S
7	81	0	0	H	S
8	100	0	0	H	S
9	0	56	76	N	F
10	0	88	60	M	S
11	0	139	0	M	S
12	0	91	61	M	S
13	0	49	98	N	F
14	0	97	0	M	S
15	0	129	0	M	S
16	0	71	0	M	S
17	0	18	100	N	S
18	0	15	90	N	S
19	0	87	62	M	F
20	0	37	89	N	S
21	0	10	88	N	S
22	0	0	59	N	S
23	0	23	100	N	S
24	0	1	83	N	S

4.3 Fault detection & diagnosis with Bayesian network

4.3.1 Continuous stirred tank reactors

From the simulated learning data and knowledge-based model, BBN structure and parameters was built in Figure 4.8. The parameters of edges of four case of faults and their child nodes were supposed by prior knowledge of the process. For example, the child node of fault 3 is variable 3, temperature of CSTR 2, because additional heat input occurred when the fault scenario 3 was activated. Hotelling's T-squared was applied to fault detection of the process and both of PCA-based contribution plot and BBN-based fault diagnosis method is used to find root causes of the faults. The results of 30 random fault magnitude simulation are in Table 4.3. T-squared method was effective for most of fault cases except the fault 3&5 case. Contribution plot succeed in finding root cause when the single fault was generated. However, diagnosis accuracy was poor when the multiple faults were generated. In contrast, BBN-based fault diagnosis was effective for both of single fault and multiple faults cases. In Figure 4.9, the contribution of CSTR 2 temperature had the largest value because fault 3 generate additional heat input to the CSTR 2. Therefore, the PCA-based fault diagnosis is effective for single fault 3. However, although fault 1 and 2 effect temperature and level of CSTR 1, the contribution of CSTR 4 temperature had the largest value. Thus, false alarm will be generated for CSTR 4 which was in normal condition. In contrast, In Figure 4.10, BBN-based fault diagnosis figured out the root causes of fault for both cases accurately.

However, for fault 2&4 and 3&5 cases, the accuracy of fault diag-

nosis is worse than the others, 48.7% and 52.9%, respectively. Figure 4.11 shows the result of BBN-based fault diagnosis of fault $2\&4$ and $3\&5$ cases. One of the root cause was diagnosed successfully for both cases, however the other root cause could not be figured out.

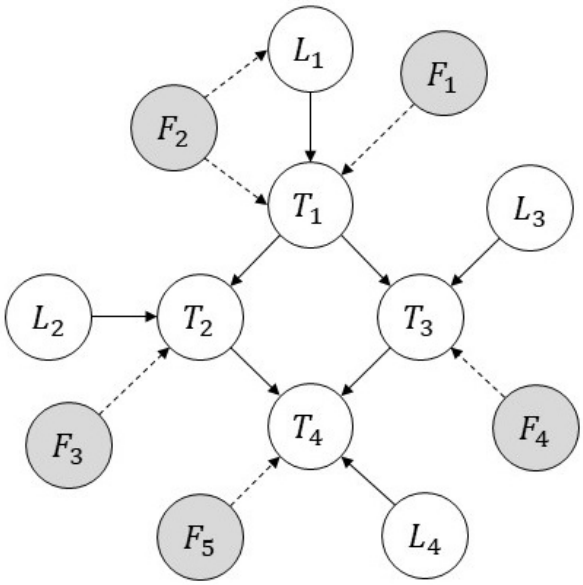

Figure 4.8: The result of Bayesian belief network learning for the example CSTRs process and fault scenarios with the process data sets.

Table 4.3: FDD rate of the traditional method and proposed method with various fault scenarios

Fault scenario	Detection rate	Diagnosis rate (PCA)	Diagnosis rate (BBN)
1	0.986	0.992	1.000
2	1.000	0.875	0.924
3	0.992	0.667	0.882
4	1.000	0.925	0.924
5	0.986	0.821	0.961
1&2	1.000	0.145	1.000
1&3	0.986	0.267	0.961
1&4	0.992	0.086	0.926
1&5	1.000	0.312	1.000
2&3	0.986	0.467	1.000
2&4	0.942	0.086	0.487
2&5	1.000	0.125	0.828
3&4	0.933	0.367	0.843
3&5	0.728	0.467	0.529
4&5	1.000	0.854	0.926

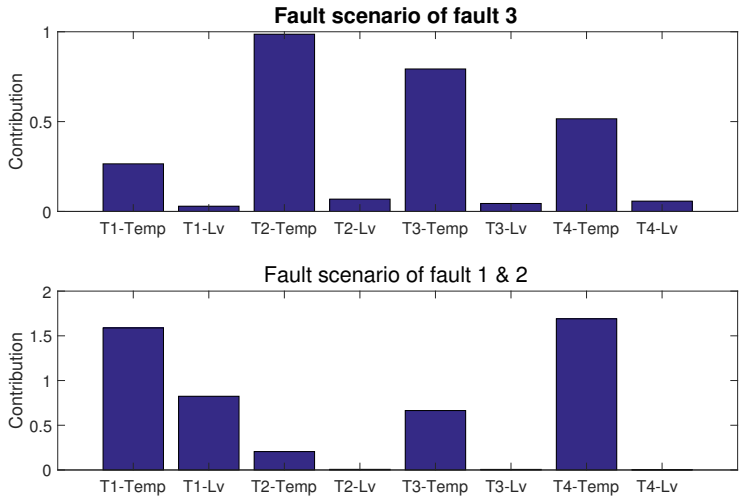

Figure 4.9: Example of the result of contribution plot from PCA-based fault diagnosis when single fault 3 case and multiple faults 1&2 case.

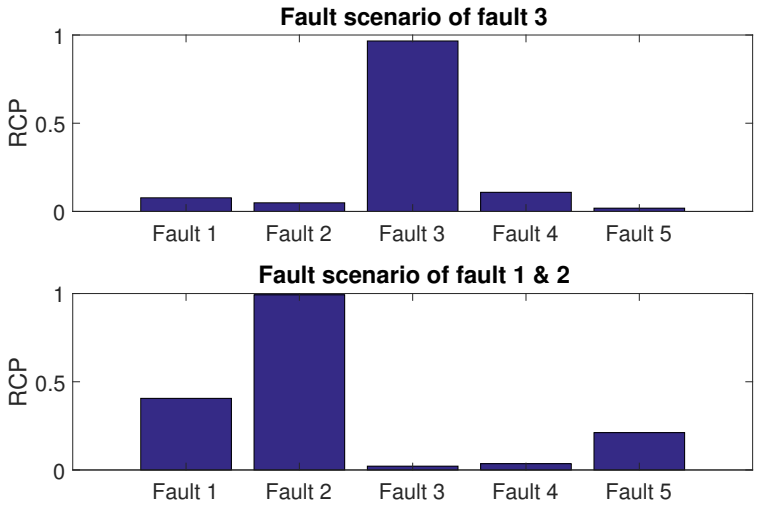

Figure 4.10: Example of the result of BBN-based fault diagnosis when single fault 3 case and multiple faults 1&2 case.

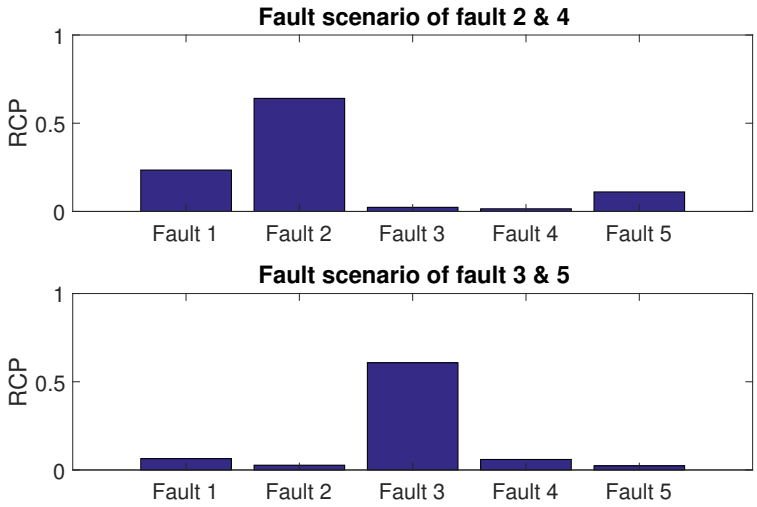

Figure 4.11: The result of BBN-based fault diagnosis when fault $2\&4$ case and multiple faults $3\&5$ case.

4.3.2 Wet gas compressor

The BBN for the WGC process learned with operation data is in Figure 4.12. The parameters of edges of four case of faults and their child nodes were supposed by prior knowledge of the process. Seven fault scenarios, four single fault cases and three multiple faults cases, were simulated to generate operation data sets based on HYSYS dynamic model. Hotelling's T-squared method was used to detect the process faults and traditional PCA-based FDD scheme and the proposed BBN-based FDD scheme were used to diagnose the root causes of the detected faults. For single fault case, Figure 4.14 shows the root cause diagnosis result of fault 4 scenario from the PCA-based scheme. If the contribution plot can diagnose the proper root cause, the variables which are most related to 2nd WGC have the largest values. However, every measured variables have almost same contributions for the Hotelling T-squared value at fault detected time. In contrast, since root cause probability of fault 4 from the proposed BBN-based scheme is the largest in Figure 4.13, the proposed scheme can diagnose the root cause of fault 4 accurately. For multiple faults case, Figure 4.16 shows that the PCA-based contribution plot cannot diagnose the root causes of the fault 1 & 3. However, the proposed scheme success to diagnose the root causes of Fault 1 and Fault 3. The overall results of FDD for the WGC process are in Table 4.4. Fault detection for WGC process was success for all fault scenarios. However, contribution plot of PCA-based root cause diagnosis was failed for all fault scenarios. The proposed scheme can diagnose the root causes of the fault scenarios accurately except fault 2 and fault 1& 2 cases.

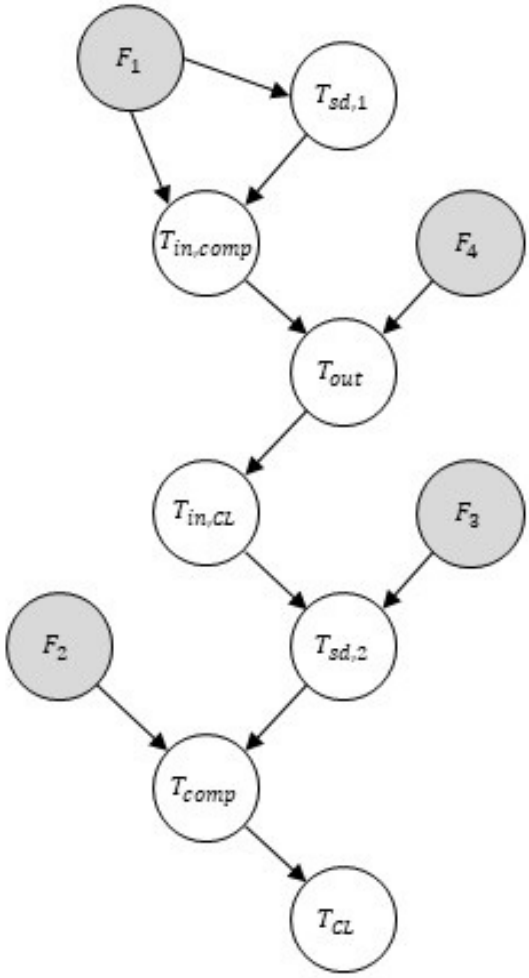

Figure 4.12: The result of Bayesian belief network learning for wet gas compressor and fault scenarios with the process data sets.

Table 4.4: Results of FDD with the traditional and proposed methods with seven fault scenarios

Fault scenario	Detection	PCA-based	BBN-based
1	S	F	S
2	S	F	F
3	S	F	S
4	S	F	S
1&2	S	F	F
1&3	S	F	S
3&4	S	F	S

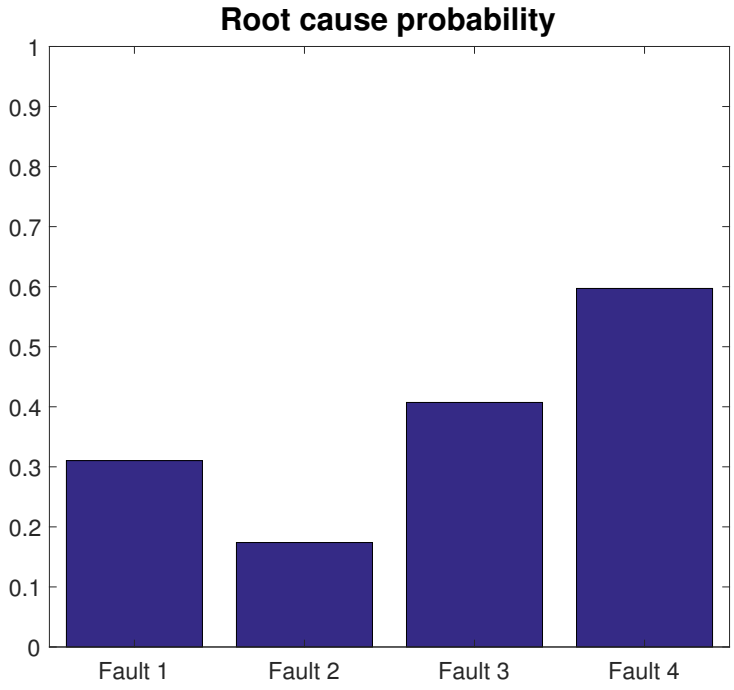

Figure 4.13: Root cause probability from the result of the proposed scheme under fault 4 condition.

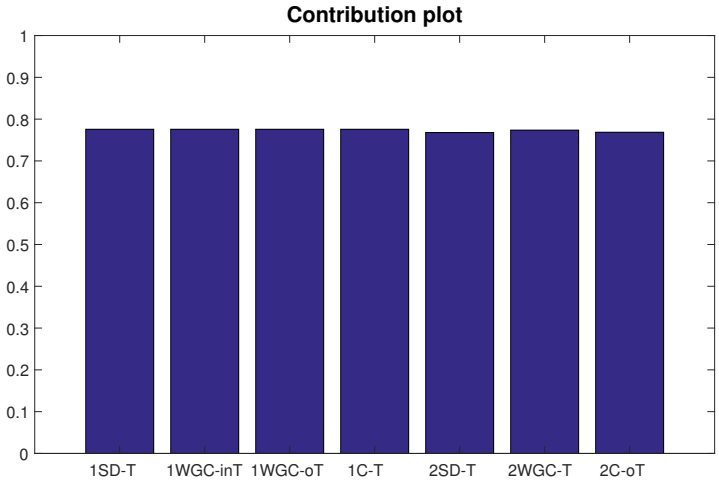

Figure 4.14: Contribution plot from the result of the traditional PCA under fault 4 condition.

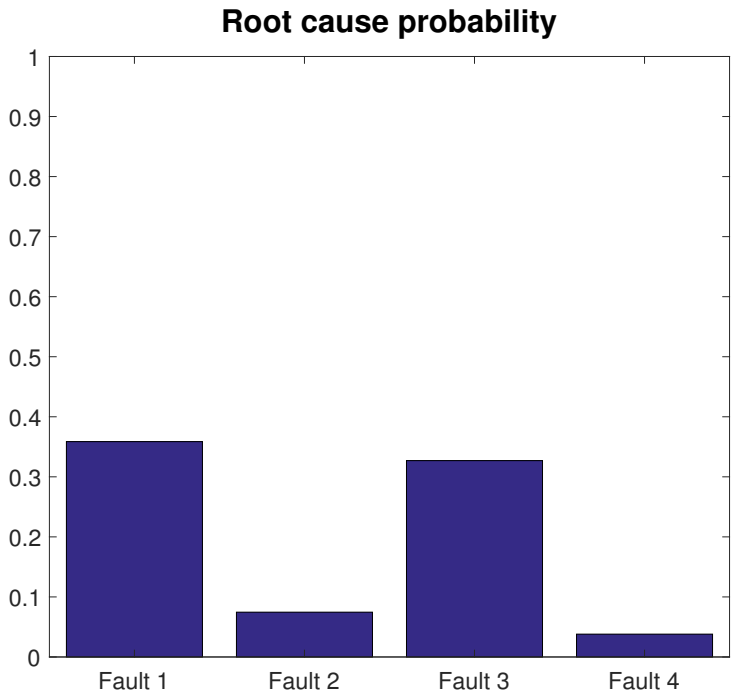

Figure 4.15: Root cause probability from the result of the proposed scheme under multiple faults condition (fault 1 & 3).

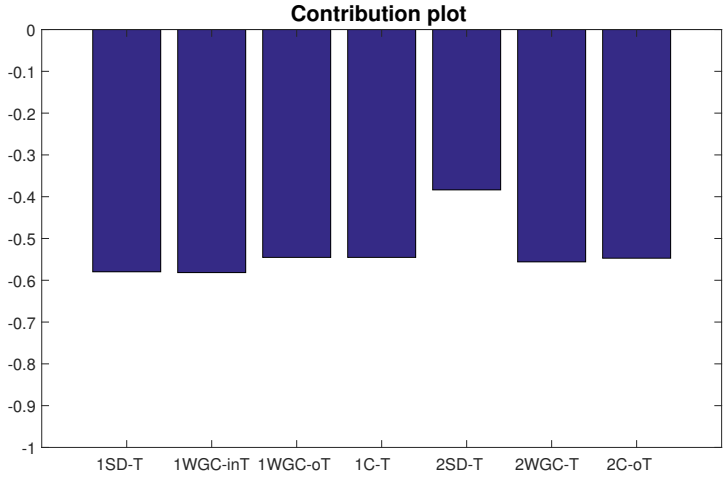

Figure 4.16: Contribution plot from the result of the proposed scheme under multiple faults condition (fault 1 & 3).

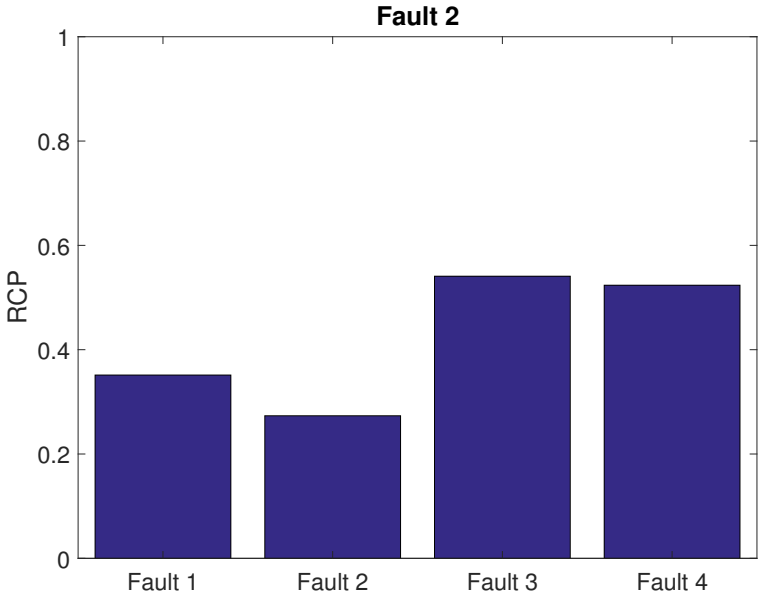

Figure 4.17: Root cause probability from the result of the proposed scheme under fault 2 scenario.

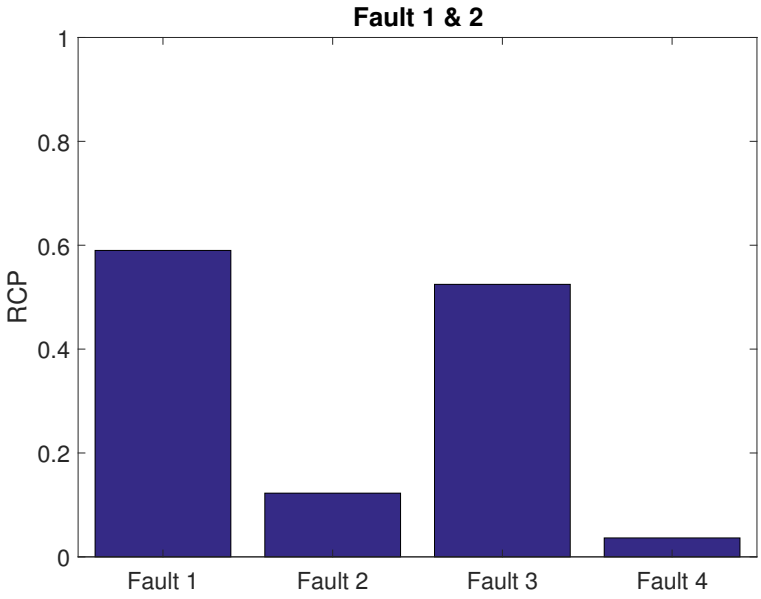

Figure 4.18: Root cause probability from the result of the proposed scheme under fault 1&2 scenario.

4.3.3 Penicillin batch process

With the learning data, BBN for the Penicillin batch process was built in Figure 4.19. To describe control loop, time lag l was introduced to define causality between manipulated variables and controlled variables. This kind of BBN is called dynamic Bayesian network (DBN) [109]. The time lag l was determined as 3 considering the time delay of the feedback controller from the Granger causality in this time. Threshold of weighted Granger causality was 1.144 when $l = 1$, 3.987 when $l = 2$ and 4.117 which is maximum when $l = 3$. With DBN of the batch process and 15 fault scenarios, FDD performances of the proposed method and PCA-based method were verified.

The simulation results of fault scenarios are in Table 4.5. For fault detection, PCA-based T-squared method can detect process faults of all fault scenarios. For fault diagnosis with PCA-based method, single fault scenario of fault 1, 2, 3 and 6 was accurate. However, for fault 4, 5 and 7, accurate fault diagnosis rate was 56.4%, 64.3% and 64.3% respectively. Although the PCA-based method guaranteed 50% above fault diagnosis rate, the fault diagnosis accuracy was poor under the double faults conditions. In figure 4.20, the example of successful fault diagnosis result of the PCA-based method. For fault 1 case, base flow rate, aeration rate and dissolved oxygen concentration had major contribution to T-squared value. Since the fault 1 was aeration rate fault, large F_g and DO_2 contribution can be an evidence of fault 1. Fault $2\&7$ case is related to vessel back pressure and pH sensor error. Therefore, the major contributions of F_b, pH and DO_2

showed that the fault 2 and 7 occur. Figure 4.21 shows the unsuccessful fault diagnosis result with PCA-based method. Although the fault 4 was base flow rate failure, F_c and DO_2 had major contribution of the T-squared value. For fault 5&7 case, regardless of the contribution plot diagnose the fault 7, there is no evidence of fault 5, coolant flow rate failure.

BBN-based fault diagnosis had better performance than the PCA-based method in all single fault scenarios. Figure 4.22 shows the fault diagnosis results of single fault cases related to input variables and Figure 4.23 shows the result of single fault cases related to on-line measurements. The faults related to measurements are difficult to find their root cause because those variables were highly correlated with input and hidden variables and the correlation is highly nonlinear. However, BBN-based fault diagnosis method can find the root cause not only for the fault related to the input variables but for the fault related to the measurements. In addition, BBN-based method was effective for double faults scenarios compared with the PCA-based method. For the fault 1&2, 2&7, 3&5 and 4&6 cases, BBN-based fault diagnosis method can figure out the root causes of the fault accurately. Fault diagnosis rate of BBN-based method was also better than the rate of the PCA-based method. However, for fault 3&4, 4&5 and 5&7, fault diagnosis performance of the proposed method was not enough to be used. Figure 4.24 shows the example of successful fault diagnosis results when the double faults occur. RCP of the correct root causes had the largest and the second largest value in all cases. Figure 4.25 shows the example of unsuccessful results of double fault cases. For the fault 3&4 case and 5&7 case, RCP of the one

root cause had the largest value but RCP value of the other root cause was not the second largest value. For fault $4\&5$ case, RCP values of the correct root causes were the second and third largest value.

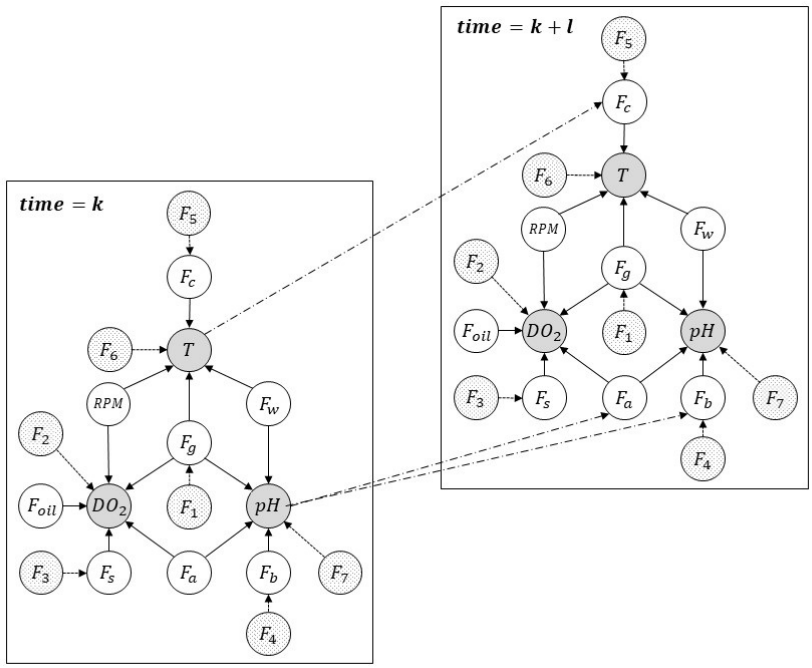

Figure 4.19: DBN for the Penicillin batch process with time lag l. Because of the PID control loop, measurement variables also effect to the input variables after l time. Therefore, the control loop should be described by DBN.

Table 4.5: FDD rate of the traditional method and proposed method with various fault scenarios of the Penicillin batch process

Fault scenario	Detection rate	Diagnosis rate (PCA)	Diagnosis rate (BBN)
1	1.000	0.912	1.000
2	1.000	0.925	0.954
3	1.000	0.912	0.975
4	1.000	0.564	0.625
5	1.000	0.643	0.812
6	1.000	0.956	1.000
7	1.000	0.643	0.728
1&2	1.000	0.351	0.812
2&3	1.000	0.086	0.487
2&7	1.000	0.218	0.954
3&4	1.000	0.028	0.260
3&5	1.000	0.250	0.728
4&5	1.000	0.000	0.125
4&6	1.000	0.312	0.900
5&7	1.000	0.028	0.320

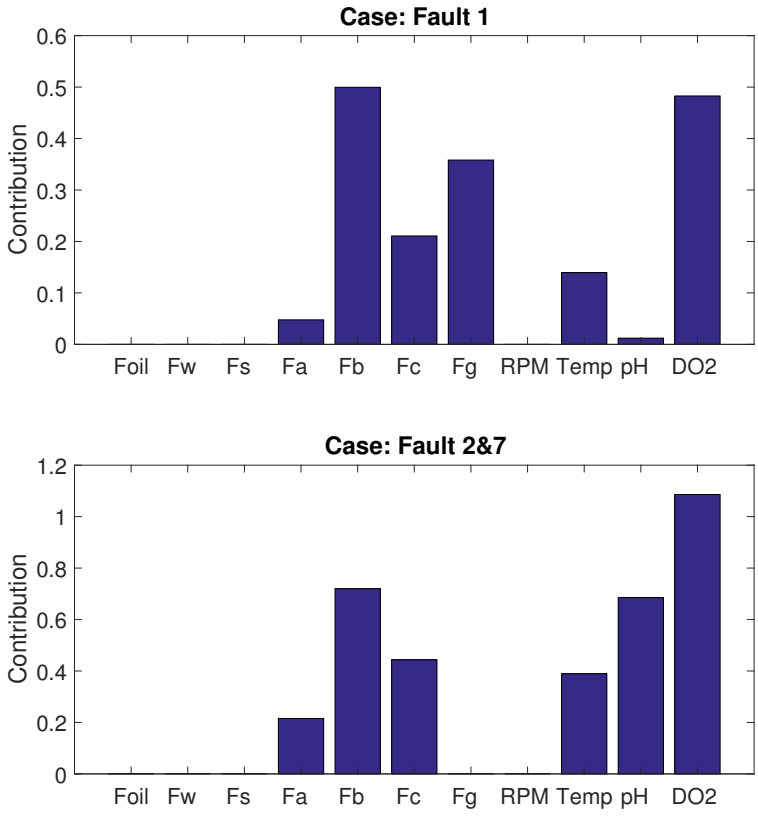

Figure 4.20: Successful fault diagnosis with PCA-based contribution plot under single fault condition

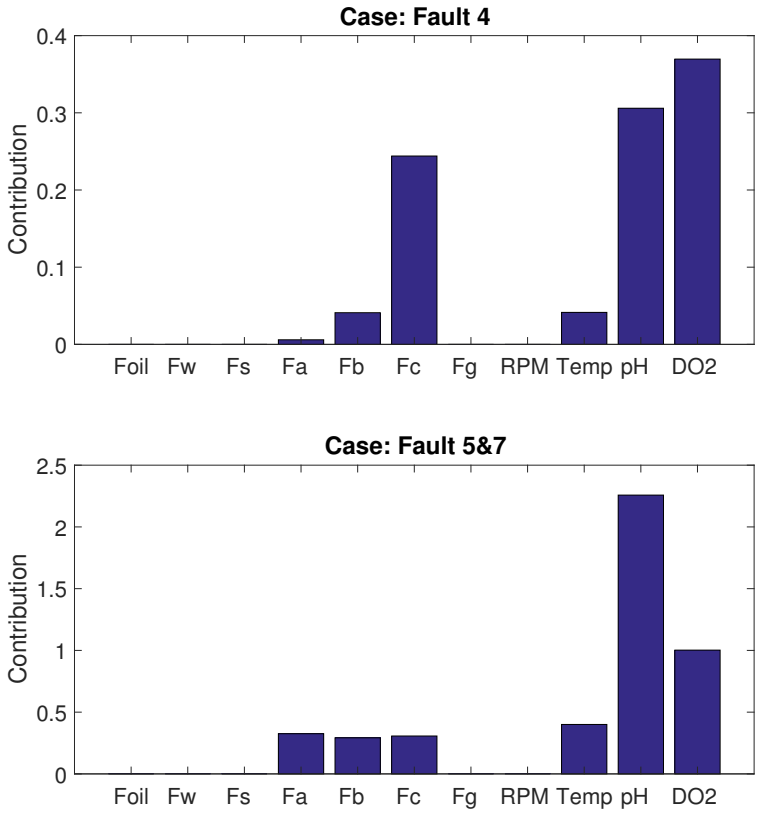

Figure 4.21: Unsuccessful fault diagnosis with PCA-based contribution plot under multiple faults condition

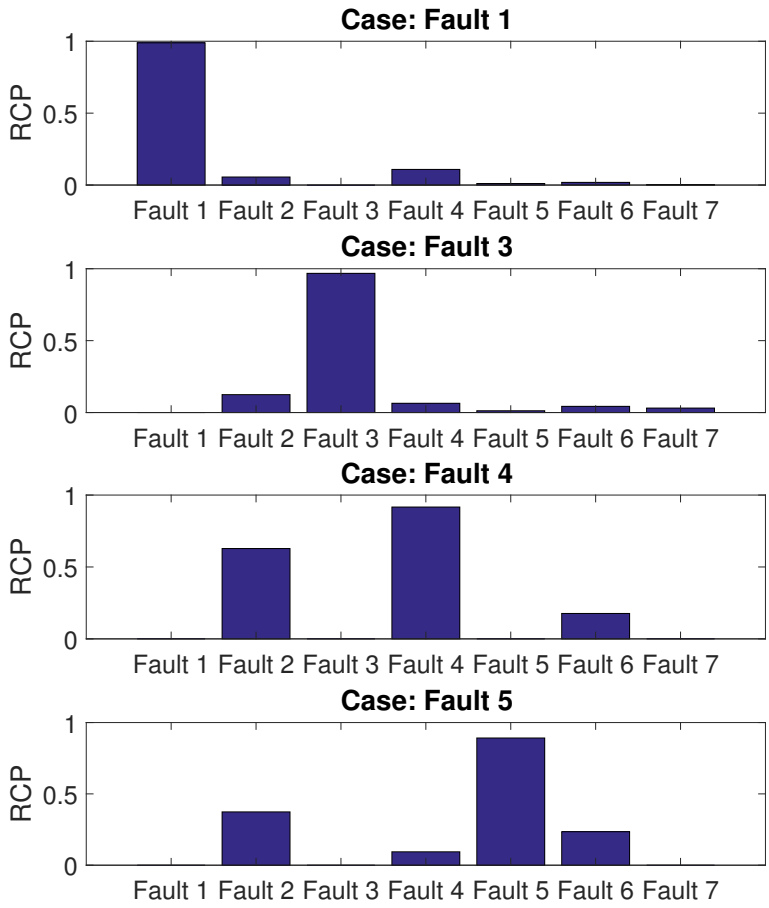

Figure 4.22: Root cause probability from the result of the proposed scheme under input variable failures.

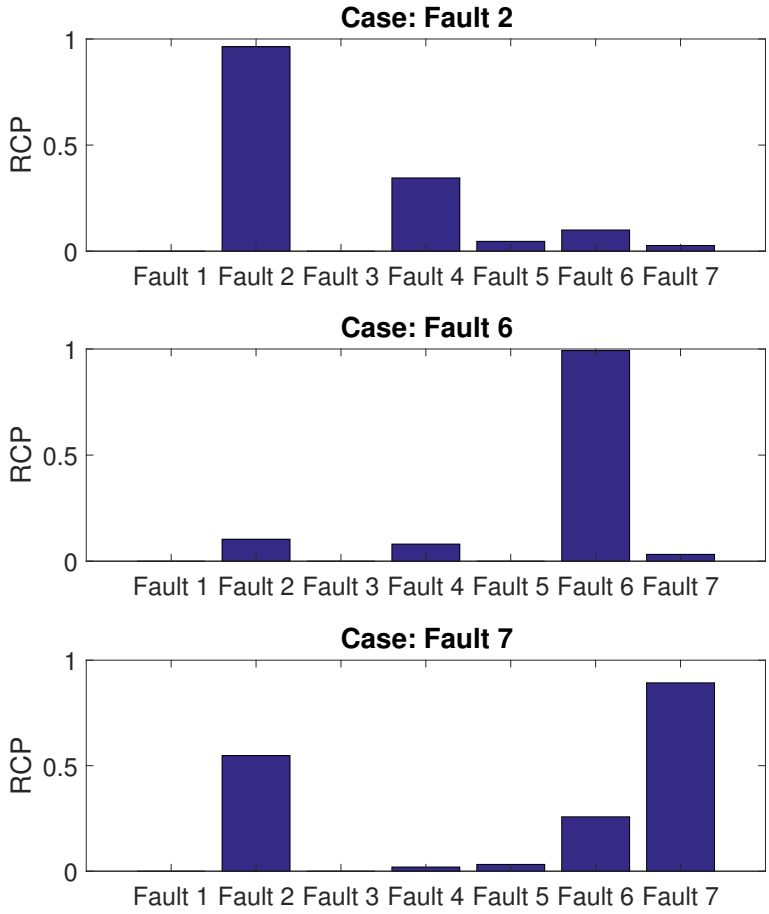

Figure 4.23: Root cause probability from the result of the proposed scheme under on-line measurement failures.

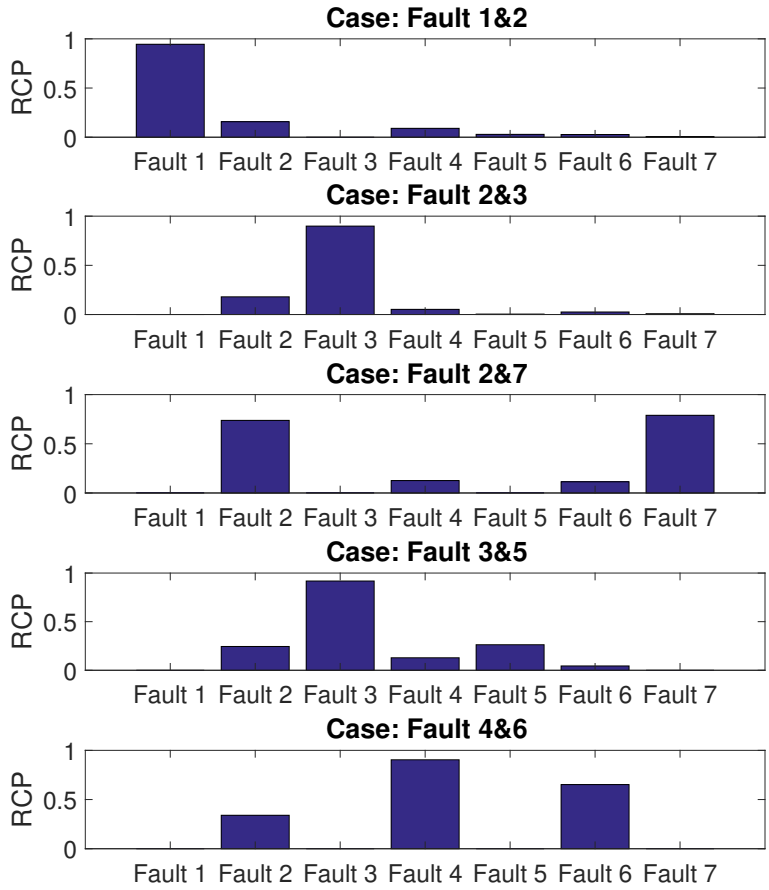

Figure 4.24: Successful fault diagnosis with the proposed scheme under multiple faults condition.

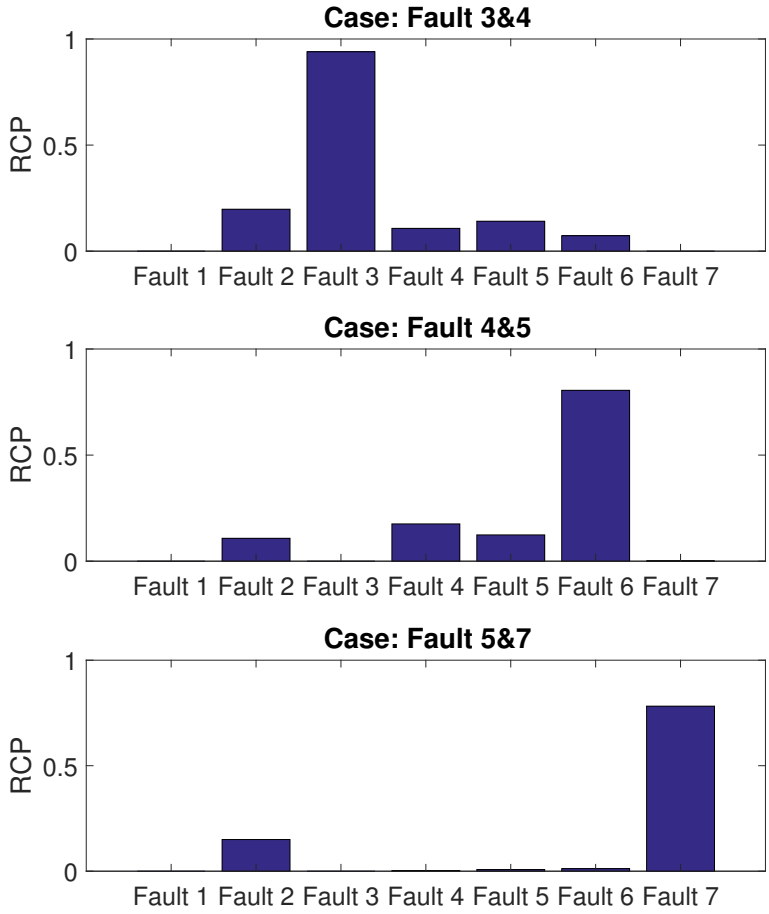

Figure 4.25: Unsuccessful fault diagnosis with the proposed scheme under multiple faults condition.

Chapter 5

Discussions & Concluding remarks

5.1 Robust parameter estimation for drug delivery system

Figures 4.1-4.3 show that most of the parameters estimated by the proposed MAP scheme have lower variances compared with the results of the least squares and Var-MAP method. This means that if PK data have large variations, estimation by the proposed method is more robust against these differences. In complicated PBPK models, a large number of parameters exist. The larger the number of parameters, the more difficult it is to find a global optimum of the objective function for estimation. In addition, because the in vivo data contains limited number of data sets and the Var-MAP method only considers the maximum probability estimation result, the values of the parameters are likely to yield different results depending on each data set. In contrast, because the covariance matrix calculated by simulation includes information on the correlations among the model parameters, the estimation result from the proposed method is also likely to follow the correlations. Therefore, if a different data set is used, the estimated parameters do not have significant differences, and can be identified within a parameter space confined by the correlations.

In Figure 4.2, the variances of the parameter estimates show stable trends as the update steps are repeated, owing to the recursive update of the prior distribution by the posterior distribution of the previous step. However, some significant changes can be observed for the other estimation methods, and the absolute values of the variances are higher because of the absence of such an update step. In some cases, because of the scarcity and randomness of the data set, the estimation performance of cov-MAP method was no better than the other methods. Since there is only one experimental data set, verification of robustness was conducted using data sets generated by sampling the parameters from uniform distributions described in Section $3.2.1$. These data sets had various trends with the signal-to-noise ratio up to 2 and some of them were quite different from the original in vivo data. To verify the robustness of cov-MAP method in a rigorous manner, these "uncommon" data sets were not excluded. With these data sets, the optimizer may find a solution different from the nominal value, which also affects the subsequent estimation step. For this reason, in Figure 4.3, estimation result for $V_{ml,FU}$ and $V_{ml,T}$, among 12 parameters, show larger variances compared with other methods. If a more number of data sets with proper filtering of outliers are available, the robustness of cov-MAP method will be further improved.

Although the same estimation scheme was used, the prediction performances were different for the different model structures. Mean squares error is commonly used for quantifying the fitting accuracy. If a value of MSE is small, this means that the result is well fitted. With the in vivo data on Tegafur, the mean squares error of the proposed

model structure was 1.817 in the log scale, which was 73% less error than that achieved by the PBPK model without dissolution dynamics. This means that including the dissolution dynamics can provide a better description of orally administrated drug delivery systems. While dissolution dynamics involves an additional parameter to determine, the Cov-MAP estimation scheme can effectively address the issue of the increased number of parameters.

5.2 Diagnosis of partial blockage in water pipe network

This study proposed a diagnosis scheme for partial blockage in water pipeline using transient test, peak search algorithm in the frequency domain, and SVM classifier. The transient test was performed in a real test bed of water pipe network, where transient flow was generated by opening and closing a valve. In blockage section of the network, different sizes of blocked pipes were installed. To obtain training data sets for construction of SVM classifier, transient test for each blockage size was performed. The time domain transient data sets were converted to the frequency domain, and the peaks with significant amplitudes were observed in different frequency ranges for each blockage size. Because the difference between frequency ranges of peaks can be used for the partial blockage diagnosis, a peak search algorithm was proposed to identify a peak location correctly. As a result, the FC-peaks of each blockage size were identified. With these FC-peaks, SVM classifier was constructed. To verify the performance of the proposed diagnosis scheme, several additional transient tests

were performed and 24 pressure data sets were obtained. With the new data sets, SVM structure could diagnose existence of the partial blockage in pipeline and its blockage size. Therefore, the proposed diagnosis scheme could diagnose the partial blockage in water pipeline within the framework of transient analysis.

However, diagnosis of four data sets failed. For data set 1, the SVM model detected partial blockage of pipe, but failed to diagnose its size. For data sets 9 and 13, the SVM classifier could not detect the blockage of $0.125A$. For data set 19, although the pipe did not have any blockage, the model generated a false alarm. Therefore, the success ratio of the proposed method was 87.5%, frequency of false alarm was 4.17%, and the ratio of detection failure was 8.33%. In addition, failures of detection and diagnosis of the SVM model occurred in moderate blockage cases. For data sets 10 and 12, the values of M-score were 88 and 91 which were not significant because H-scores of those were 60 and 61, respectively. However, the result can be considered that the blockage size of the pipeline is moderate, if any. Therefore even if the proposed scheme failed to diagnose the blockage sizes of data sets 10 and 12, it could still detect the blockage inside the pipe. Moreover, the false alarm occurred when the SVM classifier diagnosed a normal pipeline as moderately blocked pipeline. However, it is worthwhile to note that all the fault diagnosis algorithms are basically designed to strike a balance between false and missed alarms, and the proposed approach could successfully detect the severe blockage cases except one and the moderate blockage with low false alarm rates.

The proposed method does not require additional equipment like oscillation transient generator or water hammer, but provides effective diagnosis performance for water pipe blockage. Since the transient flow was generated by the valve opening, hydrants already installed in the field can be used to generate transient flow. Therefore, the proposed method is easy to apply to the field lines and economically feasible compared with previous researches using oscillation transient flow or water hammer.

5.3 Fault detection & diagnosis with Bayesian network

This study proposed the BBN-based fault diagnosis method to figure out accurate root causes of various process failures. The proposed scheme had two parts, fault detection system and fault diagnosis system. Traditional multivariate statistics, PCA-based T-squared method, was used to detect process failures. When the fault was detected, BBN-based fault diagnosis method figure out the root causes of the failures based on the Bayes' rule. To verified the fault diagnosis accuracy, three case studies were conducted.

The first case study was the CSTR process. BBN of CSTR process was built with eight measurements and learning data. Five different fault scenarios were defined to generate 30 fault state data respectively. With the 30 sets of normal and fault data, PCA-based contribution plot and BBN-based fault diagnosis was applied. As a result, fault diagnosis rate of the proposed method was better than the traditional method for both of single and double fault scenarios. Most of fault di-

agnosis rate of the BBN-based method was accurate enough to figure out the root causes of the process failures except fault $2\&4$ and $3\&5$ cases. Figure 4.9 shows the BBN-based fault diagnosis result of the fault $2\&4$ and $3\&5$ cases. For fault $2\&4$ case, root cause of fault 2 was diagnosed clearly. However, root cause of fault 4 cannot be diagnosed because root cause of fault 2 influence both of temperature of CSTR 1 and level of CSTR 1. Therefore, the RCP of fault 1 was larger than the fault 4. Moreover, fault 4 was leakage in CSTR 3 which changed temperature slowly compared with the effect of fault 2. Therefore, the influence of fault 4 was under-estimated at the fault detection time. In addition, for fault $3\&5$ case, the root cause of fault 3 was diagnosed correctly. However, because the fault 5 only influence the temperature of CSTR 4, the RCP value of fault 5 was under-estimated when the two faults influence neighbouring variables in BBN. Even if the diagnosis rate of BBN-based method is lower value, the one root was successfully diagnosed in the most of simulations.

The next case study was about the WGC process described in Aspen HYSYS model with real plant data. Compared with the first case study, the nonlinearity was increased since WGC model include thermal dynamics equations. In this reason, PCA-based method failed to diagnose the root causes of the faults in every fault scenarios. However, BBN-based method can diagnose the root causes successfully except fault 2 and $1\&2$ cases. Figure 4.17 shows the diagnosis result of BBN-based method when the fault 2 occurred and Figure 4.18 shows the result when the fault $1\&2$ occurred. The root cause of fault 2 cannot be diagnosed for the both cases. Similarly the CSTR process case study, the root cause related to the variable which is located at

the end of the BBN structure was hard to diagnose accurately. Because the fault which only influence the end-point variable is usually depended on the data of little number of variables, diagnosis of the root cause is easy to be failed. In addition, WGC process has feedback loop which is not related to the measurements. The causality of the feedback loop was not learned with the learning data because the learning data was not described the characteristic of the process perfectly.

The last case study was the batch process monitoring. The Penicillin batch process is highly nonlinear and has feedback loop which is related to the on-line measurements. Therefore, to describe the feedback loop, DBN was used to describe the Penicillin batch process. The time-lag l was determined as 3 which draw the largest threshold of Granger causality. However, l depends on the feedback effect of the process. Therefore, if the controller gain was changed or different control logic was applied, time-lag l of DBN should be updated with current process data. Fortunately, because the proposed method is free of limitation of the huge computation time of NP-hard problem, updated value of l will be easy to calculate with new process data.

The fault diagnosis rate is worse than the previous case studies because of the batch process nonlinearity. However, the fault diagnosis performance of BBN-based method was outstanding compared with the PCA-based method. In addition, BBN-based method diagnosed the root causes of the most of double faults cases although the PCA-based method was failed. Figure 4.25 shows the unsuccessful

diagnosis result of BBN-method. For Fault $3\&4$ case, the root cause of fault 3 can be diagnosed. Because the fault 4 influence the acid flow rate with time delay, the DO_2 was also influenced by the fault 4. Therefore, the RCP of the fault 4 was under-estimated and RCP of fault 2 was over-estimated. Similarly for fault $5\&7$ case, the RCP of fault 2 was over-estimated because of the feedback loop. For fault $4\&5$ case, since both of the faults were related to the feedback loop, RCP of the root causes were under-estimated. Moreover, the fault 6 influence the temperature directly. Therefore, the root cause of fault 5 can be misunderstood with fault 6 under double fault case.

Despite of the unsuccessful result of the double fault cases, BBN-based fault diagnosis method guaranteed better performance compared with the traditional method. In addition, BBN-based method can diagnose one of the root cause successfully. After the one root cause was fixed, the process can be operated under the single fault condition which can be diagnosed well with the BBN-based method. Therefore, BBN-based method can contribute the accurate process monitoring and immediate process maintenance. Moreover, if the root causes will be analysed with Top-3 RCP for double fault cases, the average diagnosis rate increases to 0.842 which is enough value to identify the root causes of the double faults conditions. For the batch process, BBN-based method is much more effective method compared with the PCA-based method because the BBN can described the process nonlinearity and complicate correlations. Therefore, the BBN-based method is useful alternative for the batch process monitoring when the traditional PCA-based cannot diagnose the root cause of the batch process failures.

When the undefined fault occurred, RCPs of pre-defined faults may not diagnosis the root cause of faults accurately. However, RCP value can be calculated not only for the pre-defined fault, but for the measurement variables. In most of cases, observation of RCPs of measurement variables can make complicate the process monitoring. However, when the diagnosis result cannot identify the root cause under the unknown fault condition, the RCPs of measurement variables can help to find true root cause. For example, Figure 5.1 shows the result when the tank 2 level fault occurred. Because the tank 2 level fault was undefined fault, RCPs of pre-defined faults cannot identify the root cause. However, if RCP of tank 2 level which is a measurement variable is considered, the root cause of unknown fault can be figure out.

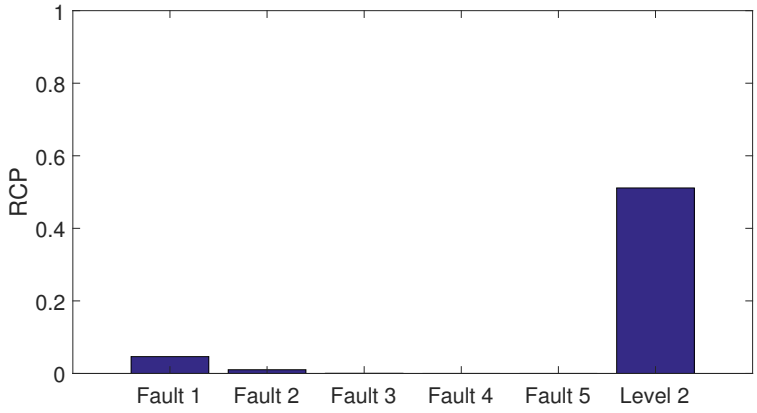

Figure 5.1: Fault diagnosis result for undefined fault which is tank 2 level fault

5.4 Summary & Suggested future works

This thesis studied to improve the performance of the traditional FDD approaches. First method was the combination of model-based approach and data-driven approach. Base on the Bayesian inference, pharmaceutical model parameter can be estimated robustly with a small number of experiment data. This can help to make robust model which prevent frequent miss alarm of the model-based FDD approach. The second study was the data-driven approach with machine learning technique. SVM structure from the frequency domain data can classify the blocked pipe and its size automatically. With the proposed water pipe network diagnosis method, the blocked pipe and blockage size can be detected without additional equipment settlement or damage to the pipe network. The last topic was the combination of the knowledge-based approach and data-driven approach. Knowledge-based model can help to defined the fault condition with prior probability and reduce the model mismatch of the data-based model. With proposed BBN-based scheme, fault diagnosis performance of single and double faults from nonlinear process was improved. Moreover, the proposed structure learning method can help to reduce calculation time when the leaning data cannot describe the characteristic of the process perfectly and apply the experience and information of the process from the expert engineers.

There are several suggestions for further research based on the FDD approaches in this thesis to improve the process monitoring and control methods.

1. Define RCP threshold value to classify the root causes of the process faults.
2. Development plant-wide multi-layers BBN learning method using multivariate statistics.
3. Estimation and prediction of off-line measurements and model predictive control using BBN model.

T-squared contribution plot shows the priority of the variables related to the current faulty state. Similarly, the root causes from the proposed method was determined by the descending order of the RCP value. Since the value of RCP is the probability, root cause threshold of RCP value can help to classify the root causes of current state if the threshold can be determined form a certain distribution.

In addition, there exist limitation of the number of process variables to learn BBN because the structure learning of BBN is NP-hard problem. Therefore, if the number of measurements is huge, plant-wide BBN, it is impossible to construct BBN with all process variables. If the plant can be divided into small size unit process, the BBN of unit process can be built and the state of the unit process can be represented with multivariate statistics. In this way, the BBNs of unit processes with representative value can be described upper-layer BBN. Finally the multi-layers plant-wide BBN can be constructed by reducing the calculation time.

In this thesis, only on-line measurements were using to build BBN. However, if the BBN between on-line measurements and off-line measurements can be constructed and reliability of BBN is sat-

isfied, the batch product, usually off-line measurements, can be estimated and predicted. The difficult part of batch process model predictive control is development of accurate model of the batch process. If the BBN between on-line measurements and off-measurements is used as a process model, construction of model predictive control system for the batch process is possible.

Bibliography

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

[2] P. M. Frank, "Analytical and qualitative model-based fault diagnosis–a survey and some new results," *European Journal of Control*, vol. 2, no. 1, pp. 6–28, 1996.

[3] D. A. Crowl and J. F. Louvar, *Chemical process safety: fundamentals with applications*. Pearson Education, 2001.

[4] J. Labovskỳ, Z. Švandová, J. Markoš, and L. Jelemenskỳ, "Mathematical model of a chemical reactor—useful tool for its safety analysis and design," *Chemical Engineering Science*, vol. 62, no. 18, pp. 4915–4919, 2007.

[5] K. Mathioudakis and A. Stamatis, "Compressor fault identification from overall performance data based on adaptive stage stacking," in *ASME 1992 International Gas Turbine and Aeroengine Congress and Exposition*, pp. V005T15A003–V005T15A003, American Society of Mechanical Engineers, 1992.

[6] A. M. Sattar, M. H. Chaudhry, and A. A. Kassem, "Partial blockage detection in pipelines by frequency response method," *Journal of Hydraulic Engineering*, vol. 134, no. 1, pp. 76–89, 2008.

[7] K. R. Lohr and J. L. Rose, "Ultrasonic guided wave and acoustic impact methods for pipe fouling detection," *Journal of Food Engineering*, vol. 56, no. 4, pp. 315–324, 2003.

[8] C. Massari, T. C. J. Yeh, B. Brunone, M. Ferrante, and S. Meniconi, "Diagnosis of pipe systems by means of a stochastic successive linear

estimator," *Water Resources Management*, vol. 27, no. 13, pp. 4637–4654, 2013.

[9] D. T. Hutchins Sr, "Method for cleaning water pipe," Jan. 12 1993. US Patent 5,178,684.

[10] K. Mueller, "Method for cleaning and coating water-conducting pipes," Sept. 3 1991. US Patent 5,045,352.

[11] D. Li and Y. Y. Haimes, "Optimal maintenance-related decision making for deteriorating water distribution systems: 2. multilevel decomposition approach," *Water Resources Research*, vol. 28, no. 4, pp. 1063–1070, 1992.

[12] K. Lansey, C. Basnet, L. Mays, and J. Woodburn, "Optimal maintenance scheduling for water distribution systems," *Civil Engineering Systems*, vol. 9, no. 3, pp. 211–226, 1992.

[13] R. G. Quimpo and U. M. Shamsi, "Reliability-based distribution system maintenance," *Journal of Water Resources Planning and Management*, vol. 117, no. 3, pp. 321–339, 1991.

[14] J. W. Kim, G. B. Choi, J. C. Suh, and J. M. Lee, "Dynamic optimization of maintenance and improvement planning for water main system: Periodic replacement approach," *Korean Journal of Chemical Engineering*, pp. 1–8, 2015.

[15] T. Kourti, P. Nomikos, and J. F. MacGregor, "Analysis, monitoring and fault diagnosis of batch processes using multiblock and multiway pls," *Journal of Process Control*, vol. 5, no. 4, pp. 277–284, 1995.

[16] M. Jia, F. Chu, F. Wang, and W. Wang, "On-line batch process monitoring using batch dynamic kernel principal component analysis," *Chemometrics and Intelligent Laboratory Systems*, vol. 101, no. 2, pp. 110–122, 2010.

[17] Y. Shu and J. Zhao, "Fault diagnosis of chemical processes using artificial immune system with vaccine transplant," *Industrial & Engineering Chemistry Research*, vol. 55, no. 12, pp. 3360–3371, 2015.

[18] 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*, vol. 27, no. 3, pp. 293–311, 2003.

[19] R. Isermann, "Model-based fault-detection and diagnosis–status and applications," *Annual Reviews in Control*, vol. 29, no. 1, pp. 71–85, 2005.

[20] C. J. Lee, G. Lee, and J. M. Lee, "A fault magnitude based strategy for effective fault classification," *Chemical Engineering Research and Design*, vol. 91, no. 3, pp. 530–541, 2013.

[21] J. V. Kresta, J. F. MacGregor, and T. E. Marlin, "Multivariate statistical monitoring of process operating performance," *The Canadian Journal of Chemical Engineering*, vol. 69, no. 1, pp. 35–47, 1991.

[22] J. F. MacGregor, H. Yu, S. G. Muñoz, and J. Flores Cerrillo, "Data-based latent variable methods for process analysis, monitoring and control," *Computers & Chemical Engineering*, vol. 29, no. 6, pp. 1217–1223, 2005.

[23] D. W. Marquardt, "An algorithm for least-squares estimation of nonlinear parameters," *Journal of the Society for Industrial and Applied Mathematics*, vol. 11, no. 2, pp. 431–441, 1963.

[24] H. Yoshida, T. Iwami, H. Yuzawa, and M. Suzuki, "Typical faults of air conditioning systems and fault detection by arx model and extended kalman filter," tech. rep., American Society of Heating, Refrigerating and Air-Conditioning Engineers, Inc., Atlanta, GA (United States), 1996.

[25] I. Hwang, S. Kim, Y. Kim, and C. E. Seah, "A survey of fault detection, isolation, and reconfiguration methods," *IEEE Transactions on Control Systems Technology*, vol. 18, no. 3, pp. 636–653, 2010.

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

[27] P. Gehring, P. Watanabe, and G. Blaut, "Risk assessment of environmental carcinogens utilizing pharmacokinetic parameters," *Annals of the New York Academy of Sciences*, vol. 329, no. 1, pp. 137–152, 1979.

[28] H. J. K. Kang, M. G. Wientjes, and J. L. S. Au, "Physiologically based pharmacokinetic models of 2′, 3′-dideoxyinosine," *Pharmaceutical Research*, vol. 14, no. 3, pp. 337–344, 1997.

[29] J. Lindsey, W. Byrom, J. Wang, P. Jarvis, and B. Jones, "Generalized nonlinear models for pharmacokinetic data," *Biometrics*, vol. 56, no. 1, pp. 81–88, 2000.

[30] M. Andersen, r. Clewell, HJ, M. Gargas, F. Smith, and R. Reitz, "Physiologically based pharmacokinetics and the risk assessment process for methylene chloride," *Toxicology and Applied Pharmacology*, vol. 87, no. 2, pp. 185–205, 1987.

[31] P. Zhao, L. Zhang, J. Grillo, Q. Liu, J. Bullock, Y. Moon, P. Song, S. Brar, R. Madabushi, T. Wu, *et al.*, "Applications of physiologically based pharmacokinetic (pbpk) modeling and simulation during regulatory review," *Clinical Pharmacology & Therapeutics*, vol. 89, no. 2, pp. 259–267, 2011.

[32] L. Lasagna and H. K. Beecher, "The optimal dose of morphine," *Journal of the American Medical Association*, vol. 156, no. 3, pp. 230–234, 1954.

[33] S. Davis, J. Hardy, and J. Fara, "Transit of pharmaceutical dosage forms through the small intestine.," *Gut*, vol. 27, no. 8, pp. 886–892, 1986.

[34] J. B. Dressman, G. L. Amidon, C. Reppas, and V. P. Shah, "Dissolution testing as a prognostic tool for oral drug absorption: immediate release dosage forms," *Pharmaceutical Research*, vol. 15, no. 1, pp. 11–22, 1998.

[35] D. L. Kroetz, S. W. Yee, and K. M. Giacomini, "The pharmacogenomics of membrane transporters project: research at the interface of genomics and transporter pharmacology," *Clinical Pharmacology & Therapeutics*, vol. 87, no. 1, pp. 109–116, 2010.

[36] R. P. Brown, M. D. Delp, S. L. Lindstedt, L. R. Rhomberg, and R. P. Beliles, "Physiological parameter values for physiologically based pharmacokinetic models," *Toxicology and Industrial Health*, vol. 13, no. 4, pp. 407–484, 1997.

[37] M. C. Coleman and D. E. Block, "Bayesian parameter estimation with informative priors for nonlinear systems," *AIChE Journal*, vol. 52, no. 2, pp. 651–667, 2006.

[38] A. Gelman, F. Bois, and J. Jiang, "Physiological pharmacokinetic analysis using population modeling and informative prior distributions," *Journal of the American Statistical Association*, vol. 91, no. 436, pp. 1400–1412, 1996.

[39] S. M. Kay, *Fundamentals of statistical signal processing*. Prentice Hall PTR, 1993.

[40] P. J. Harrison and C. F. Stevens, "Bayesian forecasting," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 205–247, 1976.

[41] G. Marrelec, H. Benali, P. Ciuciu, M. Pélégrini Issac, and J. B. Poline, "Robust bayesian estimation of the hemodynamic response function

in event-related bold fmri using basic physiological information," *Human Brain Mapping*, vol. 19, no. 1, pp. 1–17, 2003.

[42] J. M. Lainez, G. Blau, L. Mockus, S. Orcun, and G. V. Reklaitis, "Pharmacokinetic based design of individualized dosage regimens using a bayesian approach," *Industrial & Engineering Chemistry Research*, vol. 50, no. 9, pp. 5114–5130, 2011.

[43] S. Jang and R. Gopaluni, "Parameter estimation in nonlinear chemical and biological processes with unmeasured variables from small data sets," *Chemical Engineering Science*, vol. 66, no. 12, pp. 2774–2787, 2011.

[44] P. K. Mohapatra, M. H. Chaudhry, A. Kassem, and J. Moloo, "Detection of partial blockages in a branched piping system by the frequency response method," *Journal of Fluids Engineering*, vol. 128, no. 5, pp. 1106–1114, 2006.

[45] J. P. Vítkovskỳ, A. R. Simpson, and M. F. Lambert, "Leak detection and calibration using transients and genetic algorithms," *Journal of Water Resources Planning and Management*, vol. 126, no. 4, pp. 262–265, 2000.

[46] M. Stephens, M. Lambert, A. Simpson, J. Nixon, and J. Vitkovsky, "Water pipeline condition assessment using transient response analysis," in *New Zealand Water and Wastes Association. Conference (47th: 2005: Auckland, New Zealand)*, 2005.

[47] H. F. Duan, P. J. Lee, A. Kashima, J. Lu, M. Ghidaoui, and Y. K. Tung, "Extended blockage detection in pipes using the system frequency response: analytical analysis and experimental verification," *Journal of Hydraulic Engineering*, vol. 139, no. 7, pp. 763–771, 2013.

[48] V. Venkatasubramanian, R. Rengaswamy, S. N. Kavuri, and K. Yin, "A review of process fault detection and diagnosis: Part iii: Pro-

cess history based methods," *Computers & Chemical Engineering*, vol. 27, no. 3, pp. 327–346, 2003.

[49] H. Vedam and V. Venkatasubramanian, "Pca-sdg based process monitoring and fault diagnosis," *Control Engineering Practice*, vol. 7, no. 7, pp. 903–917, 1999.

[50] L. H. Chiang and R. D. Braatz, "Process monitoring using causal map and multivariate statistics: fault detection and identification," *Chemometrics and Intelligent Laboratory Systems*, vol. 65, no. 2, pp. 159–178, 2003.

[51] J. A. B. Geymayr and N. F. F. Ebecken, "Fault-tree analysis: a knowledge-engineering approach," *IEEE Transactions on Reliability*, vol. 44, no. 1, pp. 37–45, 1995.

[52] W. E. Vesely, F. F. Goldberg, N. H. Roberts, and D. F. Haasl, "Fault tree handbook," tech. rep., Nuclear Regulatory Commission Washington dc, 1981.

[53] S. J. Qin, "Survey on data-driven industrial process monitoring and diagnosis," *Annual Reviews in Control*, vol. 36, no. 2, pp. 220–234, 2012.

[54] D. Dong and T. J. McAvoy, "Nonlinear principal component analysis—based on principal curves and neural networks," *Computers & Chemical Engineering*, vol. 20, no. 1, pp. 65–78, 1996.

[55] S. W. Choi, C. Lee, J. M. Lee, J. H. Park, and I. B. Lee, "Fault detection and identification of nonlinear processes based on kernel pca," *Chemometrics and Intelligent Laboratory Systems*, vol. 75, no. 1, pp. 55–67, 2005.

[56] V. Venkatasubramanian, R. Vaidyanathan, and Y. Yamamoto, "Process fault detection and diagnosis using neural networks—i. steady-state processes," *Computers & Chemical Engineering*, vol. 14, no. 7, pp. 699–712, 1990.

[57] Y. Maki and K. A. Loparo, "A neural-network approach to fault detection and diagnosis in industrial processes," *IEEE Transactions on Control Systems Technology*, vol. 5, no. 6, pp. 529–541, 1997.

[58] C. Cortes and V. Vapnik, "Support vector machine," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[59] Y. B. Dibike, S. Velickov, D. Solomatine, and M. B. Abbott, "Model induction with support vector machines: introduction and applications," *Journal of Computing in Civil Engineering*, vol. 15, no. 3, pp. 208–216, 2001.

[60] B. S. Yang, W. W. Hwang, M. H. Ko, and S. J. Lee, "Cavitation detection of butterfly valve using support vector machines," *Journal of Sound and Vibration*, vol. 287, no. 1, pp. 25–43, 2005.

[61] S. R. Mounce, R. B. Mounce, and J. B. Boxall, "Novelty detection for time series data analysis in water distribution systems using support vector machines," *Journal of Hydroinformatics*, vol. 13, no. 4, pp. 672–686, 2011.

[62] B. W. Karney and D. McInnis, "Transient analysis of water distribution systems," *Journal (American Water Works Association)*, pp. 62–70, 1990.

[63] B. S. Jung, B. W. Karney, P. F. Boulos, and D. J. Wood, "The need for comprehensive transient analysis of distribution systems," *Journal (American Water Works Association)*, pp. 112–123, 2007.

[64] S. J. Lee, G. Lee, J. C. Suh, and J. M. Lee, "Online burst detection and location of water distribution systems and its practical applications," *Journal of Water Resources Planning and Management*, vol. 142, no. 1, p. 04015033, 2015.

[65] X. J. Wang, M. F. Lambert, and A. R. Simpson, "Detection and location of a partial blockage in a pipeline using damping of fluid

transients," *Journal of Water Resources Planning and Management*, vol. 131, no. 3, pp. 244–249, 2005.

[66] B. Brunone, M. Ferrante, and S. Meniconi, "Discussion of "detection of partial blockage in single pipelines" by pk mohapatra, mh chaudhry, aa kassem, and j. moloo," *Journal of Hydraulic Engineering*, vol. 134, no. 6, pp. 872–874, 2008.

[67] P. J. Lee, H. F. Duan, M. Ghidaoui, and B. Karney, "Frequency domain analysis of pipe fluid transient behaviour," *Journal of Hydraulic Research*, vol. 51, no. 6, pp. 609–622, 2013.

[68] J. Tuck, P. J. Lee, M. Davidson, and M. S. Ghidaoui, "Analysis of transient signals in simple pipeline systems with an extended blockage," *Journal of Hydraulic Research*, vol. 51, no. 6, pp. 623–633, 2013.

[69] M. Adewumi, D. Lysak, and S. Ibraheem, "Blockage detection method and associated system," Aug. 9 2001. US Patent App. 09/925,860.

[70] H. F. Duan, P. J. Lee, M. S. Ghidaoui, and Y. K. Tung, "Extended blockage detection in pipelines by using the system frequency response analysis," *Journal of Water Resources Planning and Management*, vol. 138, no. 1, pp. 55–62, 2011.

[71] A. Ligęza and J. Kościelny, "A new approach to multiple fault diagnosis: A combination of diagnostic matrices, graphs, algebraic and rule-based models. the case of two-layer models," *International Journal of Applied Mathematics and Computer Science*, vol. 18, no. 4, pp. 465–476, 2008.

[72] R. Reiter, "A theory of diagnosis from first principles," *Artificial Intelligence*, vol. 32, no. 1, pp. 57–95, 1987.

[73] F. V. Jensen, *An introduction to Bayesian networks*, vol. 210. UCL Press London, 1996.

[74] S. Dey and J. Stori, “A bayesian network approach to root cause diagnosis of process variations,” *International Journal of Machine Tools and Manufacture*, vol. 45, no. 1, pp. 75–91, 2005.

[75] Y. Huang, R. McMurran, G. Dhadyalla, and R. P. Jones, “Probability based vehicle fault diagnosis: Bayesian network method,” *Journal of Intelligent Manufacturing*, vol. 19, no. 3, pp. 301–311, 2008.

[76] S. Y. Kim, S. Imoto, and S. Miyano, “Inferring gene networks from time series microarray data using dynamic bayesian networks,” *Briefings in Bioinformatics*, vol. 4, no. 3, pp. 228–235, 2003.

[77] M. Zou and S. D. Conzen, “A new dynamic bayesian network (dbn) approach for identifying gene regulatory networks from time course microarray data,” *Bioinformatics*, vol. 21, no. 1, pp. 71–79, 2004.

[78] X. Hao and S. Cai Xin, “Artificial immune network classification algorithm for fault diagnosis of power transformer,” *IEEE Transactions on Power Delivery*, vol. 22, no. 2, pp. 930–935, 2007.

[79] S. C. Chapra and R. P. Canale, *Numerical methods for engineers*, vol. 2. McGraw-Hill New York, 1988.

[80] S. K. Sengijpta, “Fundamentals of statistical signal processing: Estimation theory,” 1995.

[81] S. E. Fienberg *et al.*, “When did bayesian inference become" bayesian"?,” *Bayesian Analysis*, vol. 1, no. 1, pp. 1–40, 2006.

[82] P. L. Bonate and J. L. Steimer, *Pharmacokinetic-pharmacodynamic modeling and simulation*. Springer, 2011.

[83] A. Dokoumetzidis and P. Macheras, “A century of dissolution research: from noyes and whitney to the biopharmaceutics classification system,” *International Journal of Pharmaceutics*, vol. 321, no. 1, pp. 1–11, 2006.

[84] J. F. MacGregor and T. Kourti, "Statistical process control of multivariate processes," *Control Engineering Practice*, vol. 3, no. 3, pp. 403–414, 1995.

[85] C. W. Hsu, C. C. Chang, C. J. Lin, *et al.*, "A practical guide to support vector classification," 2003.

[86] B. Schölkopf, K. K. Sung, C. J. Burges, F. Girosi, P. Niyogi, T. Poggio, and V. Vapnik, "Comparing support vector machines with gaussian kernels to radial basis function classifiers," *Signal Processing, IEEE Transactions on*, vol. 45, no. 11, pp. 2758–2765, 1997.

[87] P. Spirtes, C. N. Glymour, and R. Scheines, *Causation, prediction, and search*. MIT Press, 2000.

[88] Z. Li, P. Li, A. Krishnan, and J. Liu, "Large-scale dynamic gene regulatory network inference combining differential equation models with local dynamic bayesian network analysis," *Bioinformatics*, vol. 27, no. 19, pp. 2686–2691, 2011.

[89] C. W. Granger, "Investigating causal relations by econometric models and cross-spectral methods," *Econometrica: Journal of the Econometric Society*, pp. 424–438, 1969.

[90] G. F. Cooper, "The computational complexity of probabilistic inference using bayesian belief networks," *Artificial Intelligence*, vol. 42, no. 2-3, pp. 393–405, 1990.

[91] P. Dagum and M. Luby, "Approximating probabilistic inference in bayesian belief networks is np-hard," *Artificial Intelligence*, vol. 60, no. 1, pp. 141–153, 1993.

[92] J. B. Houston and K. E. Kenworthy, "In vitro-in vivo scaling of cyp kinetic data not consistent with the classical michaelis-menten model," *Drug Metabolism and Disposition*, vol. 28, no. 3, pp. 246–254, 2000.

[93] M. Rowland, "Protein binding and drug clearance," *Clinical Pharmacokinetics*, vol. 9, no. 1, pp. 10–17, 1984.

[94] J. H. Sung, A. Dhiman, and M. L. Shuler, "A combined pharmacokinetic–pharmacodynamic (pk–pd) model for tumor growth in the rat with uft administration," *Journal of Pharmaceutical Sciences*, vol. 98, no. 5, pp. 1885–1904, 2009.

[95] H. Duan, P. Lee, M. Ghidaoui, and J. Tuck, "Transient wave-blockage interaction and extended blockage detection in elastic water pipelines," *Journal of Fluids and Structures*, vol. 46, pp. 2–16, 2014.

[96] J. H. Fan, "I-type counterflow absorber," Oct. 13 1998. US Patent 5,819,802.

[97] T. Wago, P. Gambier, J. L. Pessin, and R. Luharuka, "Externally assisted valve for a positive displacement pump," Feb. 5 2013. US Patent 8,366,408.

[98] A. Souza, S. Cruz, and J. Pereira, "Leak detection in pipelines through spectral analysis of pressure signals," *Brazilian Journal of Chemical Engineering*, vol. 17, no. 4-7, pp. 557–564, 2000.

[99] P. Alexander, B. Anson, K. Evans, I. McEwan, N. Ryan, T. Stebbings, C. Lindsey Curran, P. Griffiths, M. McKay, *et al.*, "Using platelet technology tm to locate and seal leaks in long subsea umbilical lines," in *Subsea Controls and Data Acquisition 2006: Controlling the Future Subsea*, Society of Underwater Technology, 2006.

[100] D. Covas and H. Ramos, "Case studies of leak detection and location in water pipe systems by inverse transient analysis," *Journal of Water Resources Planning and Management*, vol. 136, no. 2, pp. 248–257, 2010.

[101] J. W. Kim, G. B. Choi, J. C. Suh, and J. M. Lee, "Dynamic optimization of maintenance and improvement planning for water main

system: Periodic replacement approach," *Korean Journal of Chemical Engineering*, vol. 33, no. 1, pp. 25–32, 2016.

[102] D. Covas, H. Ramos, and A. B. De Almeida, "Standing wave difference method for leak detection in pipeline systems," *Journal of Hydraulic Engineering*, vol. 131, no. 12, pp. 1106–1116, 2005.

[103] P. J. Lee, J. P. Vítkovskỳ, M. F. Lambert, A. R. Simpson, and J. A. Liggett, "Frequency domain analysis for detecting pipeline leaks," *Journal of Hydraulic Engineering*, vol. 131, no. 7, pp. 596–604, 2005.

[104] S. H. Kim, A. Zecchin, and L. Choi, "Diagnosis of a pipeline system for transient flow in low reynolds number with impedance method," *Journal of Hydraulic Engineering*, vol. 140, no. 12, p. 04014063, 2014.

[105] G. Palshikar *et al.*, "Simple algorithms for peak detection in time-series," in *Proc. 1st Int. Conf. Advanced Data Analysis, Business Analytics and Intelligence*, 2009.

[106] S. Dash, R. Rengaswamy, and V. Venkatasubramanian, "Fuzzy-logic based trend classification for fault diagnosis of chemical processes," *Computers & Chemical Engineering*, vol. 27, no. 3, pp. 347–362, 2003.

[107] S. Yoon and J. F. MacGregor, "Principal-component analysis of multiscale data for process monitoring and fault diagnosis," *AIChE Journal*, vol. 50, no. 11, pp. 2891–2903, 2004.

[108] S. Goldrick, A. Ştefan, D. Lovett, G. Montague, and B. Lennox, "The development of an industrial-scale fed-batch fermentation simulation," *Journal of Biotechnology*, vol. 193, pp. 70–82, 2015.

[109] V. Pavlovic, J. M. Rehg, T. J. Cham, and K. P. Murphy, "A dynamic bayesian network approach to figure tracking using learned dynamic

models," in *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, vol. 1, pp. 94–101, IEEE, 1999.

초 록

이상 감지 및 진단은 최적 운전 상태 유지와 공정 안전을 위해 실제 화학 사업에서 매우 중요한 문제로 인식 되고 있다. 이상 감지 및 진단 기법은 크게 세 가지로 분류할 수 있는데, 모델 기반 접근법, 지식 기반 접근법과 데이터 기반 접근법이 그들이다. 최근 데이터 수집 및 저장 기술의 발전으로 데이터 수집 빈도가 향상 되었고 이에 따른 신호 처리 기법도 발전되었다. 따라서 데이터 기반 접근법은 과거 이상 감지 및 진단 기법들이 가지는 한계를 극복할 수 있다고 기대되고 있다.

이상 감지 및 진단 기법의 성능 향상을 위해 세 가지 방법을 제시하였다. 첫 번째는 모델 기반 접근법과 데이터 기반 접근법이 융합된 기법이다. 만약 모델의 변수가 올바른 값을 가지지 않는다면 모델 기반 접근법을 통한 이상 진단 및 감지 결과는 잘못될 수밖에 없다. 제약 공정과 같은 데이터를 모으기 힘든 공정의 경우, 정확한 변수 추정이 어려워 제한된 데이터 개수로 강건한 모델을 구축할 수 있는 기법이 요구된다. 따라서 베이지안 추정 기법을 도입해 생리학적 약동학 모델의 변수를 추정하기로 하였다. 제안된 새로운 추정 기법을 통해 보다 강건한 변수 추정이 가능하였고, 생리학적 약동학 모델에 약물 용해 모델을 융합하여 보다 정확한 약물 전달 모델을 구축할 수 있었다. 이 모델을 통해 모델 기반 이상 감지 및 진단 기법의 성능 향상을 기대할 수 있다.

만약 충분한 데이터 수집이 가능하다면, 이상 상태와 정상 상태 데이터는 둘 간의 차이점을 찾아서 구분할 수 있다. 이러한 차이점을 표현하기 위해 기계 학습 기법의 일종인 서포트벡터머신 기법을

도입하여 상수관망의 전이 분석 데이터를 처리하였다. 시간 위상의 전이 분석 데이터를 주파수 위상의 데이터로 변환하면 관망 내부의 폐색 정도에 따라 구분할 수 있다. 실험용 상수 관망을 통해 정상 데이터와 낮은 정도의 폐색 존재 시, 중간 정도의 폐색 존재 시, 심각한 폐색 존재 시 데이터를 수집하였고 이 데이터들을 통해 각각의 서포트벡터머신 구조를 구축하였다. 이러한 구조를 통해 관폐색의 존재 여부와 그 정도를 자동으로 진단할 수 있는 상수관망 전이 분석 기법을 제안하였다.

만약 수집된 데이터가 공정의 특성을 완벽하게 표현 할 수 있다면, 매우 정확한 데이터 기반 모델을 구축할 수 있다. 하지만 보통의 경우 운전되고 있는 공정에서 완벽한 학습 데이터를 수집하는 것은 불가능하다. 따라서 사전 지식과 전문가의 직관을 바탕으로 구축된 지식 기반 모델을 통해 데이터에서 표현되지 못한 공정의 특성을 반영하여 준다면 모델 불일치 정도를 줄일 수 있을 것이다. 베이즈 신뢰망은 공정 변수 간의 연관성을 표현하는 데이터 기반 모델이다. 불완전한 데이터를 통해 정확한 베이즈 신뢰망을 구축하기 위해 지식 기반 모델인 부호유향그래프에서 가중치 행렬을 구축하고 이를 베이즈 신뢰망 구축을 위한 알고리즘에 적용하여 보다 정확한 베이즈 신뢰망 구축을 위한 방법을 제안하였다. 또한 사전에 정의된 이상원인 역시 사전 정보를 바탕으로 베이즈 신뢰망 내에 표현하였다. 제안된 방법의 성능을 검증하기 위해 세 가지 사례 연구를 수행하였고, 베이즈 신뢰망 기법은 모든 사례 연구에서 기존의 주성분 분석 기법에 비해 월등한 성능을 보여주었다. 이는 단일 이상 상황 뿐 아니라 다중 이상 상황에서도 유효한 성능을 보여주었다. 결론적으로 지식 기반 접근법과 데이터 기반 접근법을 융합한 베이즈 신뢰망 기반 이상 진단 기법은 기존의 데이터 기반 접근법의 이상 진단 성능을 향상시키는데 사용 될 수 있었다.

이러한 세 가지 제안된 방법을 통해 기존의 이상 감지 및 진단 기법의 성능을 향상시켜 실시간으로 보다 정확한 공정 모니터링을 수행할 수 있을 것으로 기대된다. 따라서 이 연구에서 제시된 기법들은 이상 발생 시 공정 유지 보수에 도움을 줄 수 있고 이를 통해 공정이 최적의 상태로 유지될 수 있을 것이다.

주요어 : 공정모니터링, 데이터 기반 접근법, 베이즈망, 다변량분석, 이상진단, 기계학습
학번 : 2014-30259