# Performance Analysis and Design for Abnormality Detection Systems

by

Ying Xiong

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

Doctor of Philosophy

in

Control Systems

Department of Electrical and Computer Engineering

University of Alberta

# Abstract

The performance-orientated abnormality detection design problem has drawn more attention nowadays. This thesis focuses on analytical performance analysis and the corresponding detector design problem for linear filters, the Kullback-Leibler divergence (KLD) and Rényi divergence.

First, sensitivity analysis and sensitivity-based design are conducted for linear alarm filters. Analytical expressions are derived to quantify the sensitivity of a linear alarm filter with unknown data distributions, based on which a new design scheme is formulated to minimize the weighted sum of detection errors subject to upper bounds on the system sensitivities. The second work is on the KLD based detection for independent and identically distributed (i.i.d.) data under generalized Gaussian distributions with shape parameters greater than 1. The false alarm rate (FAR) is analytically obtained and two detection algorithms with constant and adaptive thresholds are proposed. The third work studies the Rényi divergence based detection for i.i.d. multivariate Gaussian data, where the divergence order is between 0 and 1. The off-set and scaling faults are considered under the abnormal condition. The FAR and missed alarm rate (MAR) are derived analytically, based on which a detection scheme is proposed with an adaptive divergence order.

Intensive case studies with both simulated and experimental data are conducted to verify the analytical results and to show improvement of the proposed detection schemes over existing ones.

# Preface

The ideas in Chapters 2 and 3 were from discussions with Dr. Chen and Dr. Jing, respectively. The ideas in Chapter 4 were my original ones. The algorithm formulation and case studies were developed by myself, as well as the introduction in Chapter 1. Dr. Chen offered insightful suggestions on the results and helped the manuscript edits. Dr. Jing provided great support on the technical derivations and contributed in the manuscript organization.

- Chapter 2 has been published as: Ying Xiong, Yindi Jing, Tongwen Chen, Sensitivity analysis and sensitivity-based design for linear alarm filters, *Control Engineering Practice*, 71: 29-39, 2018. Part of the work has been published as: Ying Xiong, Yindi Jing, Tongwen Chen, Performance sensitivity analysis of linear alarm filters, *2017 American Control Conference (ACC)*, Seattle, USA, pages 4424-4429, 2017.

- Chapter 3 has been published as: Ying Xiong, Yindi Jing, Tongwen Chen, Abnormality detection based on the Kullback-Leibler divergence for generalized Gaussian data, *Control Engineering Practice*, 85: 257-270, 2019.

- Chapter 4 has been submitted for publication as: Ying Xiong, Yindi Jing, Tongwen Chen, Abnormality detection using the Rényi divergence with multi-variate Gaussian data, *IEEE Transactions on Signal Processing*, 2019. Part of the work has been accepted as: Ying Xiong, Yindi Jing, Tongwen Chen, Abnormality detection with the Rényi divergence for univariate Gaussian data, *2019 IEEE Pacific Rim Conference on Communications, Computers and Signal Processing*, Victoria, Canada, 2019.

*To my beloved teachers, parents, husband and family members.*

# Acknowledgements

This thesis would not have been possible without the help and support from people around me.

Above all, I would like to thank my two supervisors, Dr. Tongwen Chen and Dr. Yindi Jing for their time and efforts in guiding and supporting me throughout my entire PhD program. Dr. Tongwen Chen offered me this precious opportunity to pursue my career in the control community, and allowed me full freedom to explore new research topics. When I was confused at the beginning, he taught me to start with simple cases and make them solid, which has encouraged me to concentrate on small research ideas first and make continuous efforts to excel. Meanwhile, I would like to thank Dr. Yindi Jing, who is always efficient, inspiring and energetic. It was Dr. Yindi Jing who patiently and kindly held my hands and walked me through the long learning curve on how to formulate problems, organize articles, write manuscripts, respond to reviewers and prepare for presentations. The most significant thing I have learned from her was how to think and interpret problems critically and to always value understanding problems other than getting decent results. Without the persistent guidance and help from my two supervisors, this dissertation would not have been possible.

Second, I would like to thank my previous supervisor, Dr. Fang Fang, with the School of Control and Computer Engineering in the North China Electric Power University, for his guidance, support and encouragement during my MSc program. It was Dr. Fang Fang who encouraged me to pursue PhD degree abroad and always dream big.

Further, I would like to thank the current and former group members in Dr. Chen and Dr. Jing's groups for the happy memories they brought to

me. I would like to thank M. Shahzad Afzal, David Li, Shiqi Lai, Yuzhe Li, Jiarao Huang, Wenkai Hu, Ning He, W. Al-Dabbagh, Cheng Zhang, Samira Rahimian, Jun Shang, Hao Yu, Boyuan Zhou, M. Hossein Roohi, Jing Zhou, Junyi Yang, Zhe Guan, Lily Wang, Iman Amini, M. Rezwan Parvez, Mani H. Dhullipalla and Haniyeh S. Alinezhad for their kindness and encouragement.

Last but not the least, I would like to thank my family members, particularly my parents, parents-in-law and husband, for their unselfish love and support. I cannot be more grateful to have my parents and parents-in-law in my life. Besides, I would like to express my deepest gratitude to my husband Qi Zeng for his consistent trust, understanding, encouragement and company ever since we first met.

# Contents

# List of Tables

# List of Figures

# List of Acronyms

| | |
|---|---|
| ARL | Average Run Length |
| CUSUM | Cumulative Sum |
| DCS | Distributed Control System |
| DDoS | Distributed Denial of Service |
| EWMA | Exponentially Weighted Moving Average |
| MA | Moving Average |
| FAR | False Alarm Rate |
| GGD | Generalized Gaussian Distribution |
| HMI | Human Machine Interface |
| i.i.d. | independent and identically distributed |
| KLD | Kullback-Leibler Divergence |
| LW | Linear Weighted |
| MA | Moving Average |
| MAD | Mean Absolute Deviation |
| MAR | Missed Alarm Rate |
| PCA | Principal Component Analysis |
| PDF | Probability Density Function |
| SCADA | Supervisory Control and Data Acquisition |
| SGD | Signed Directed Graph |
| SPE | Square Prediction Error |
| SVD | Singular Value Decomposition |

# Chapter 1

# Introduction

With increasingly high requirements on safety, reliability and economy of industrial systems, the design of effective abnormality detection mechanism has become a heated topic during last decades. In this chapter, the fundamental background and challenges on the abnormality detection system design are introduced in the first section. Then, the states of art on current advances in abnormality detection systems are demonstrated in the second section, followed by the contributions of this thesis in the last section.

## 1.1 Motivation and Background

Along with the booming development in the distributed control systems (DCS) and computer based supervisory control and data acquisition (SCADA) systems, nowadays industrial plants are highly integrated. For such complex and integrated industrial plants, it becomes rather challenging to maintain in terms of reliability and operational safety. Even though the performance of advanced control schemes has also been greatly improved, abnormal situations are still almost inevitable due to a variety of causes, such as parameter changes, disconnection problems, disturbance variations, and actuator and/or sensor problems [50]. In order to detect those abnormal situations in the early stage, the abnormality detection system design has drawn huge attention from both industry and academia.

### 1.1.1 Abnormality Detection Systems

As in [4], an abnormality detection system deals with data collected from both device and control levels through a SCADA and/or a DCS; meanwhile, it communicates with the manufacturing execution system level with the help of human machine interface (HMI), where operators monitor the process measurements and alarm states. It plays a crucial rule in supporting and guiding operators to take remedial moves that prevent fatal accidents at the preliminary stage.

The diagram of a typical abnormality detection system is depicted in Figure 1.1. In general, it contains an off-line training stage, where historical data is analyzed, and an on-line detection stage, where current measurements are processed. In the off-line training stage, abnormal features are extracted from historical data to distinguish abnormal situations from normal ones. In other words, the selected abnormal features are expected to appear only when the abnormality occurs, excluding the cases of mode transitions and normal operations. The candidate abnormal features can be obtained with the help of process knowledge, data analysis or direct observations. In order to reveal the selected key abnormal features in observed data, the detection scheme including the alarm generation rules is designed; then, parameters involved in the detection scheme are tuned. Specifically, the level-crossing mechanism with a threshold is usually utilized as the alarm generation rule in both industry and academia. The parameters involved in the detection schemes can be either numerically determined based on simulations or theoretically optimized via performance analysis as per [128]. In the on-line detection stage, current measurements are processed with the detection scheme obtained from the off-line training stage; then, alarms are generated by applying the alarm generation rules into the processed signals.

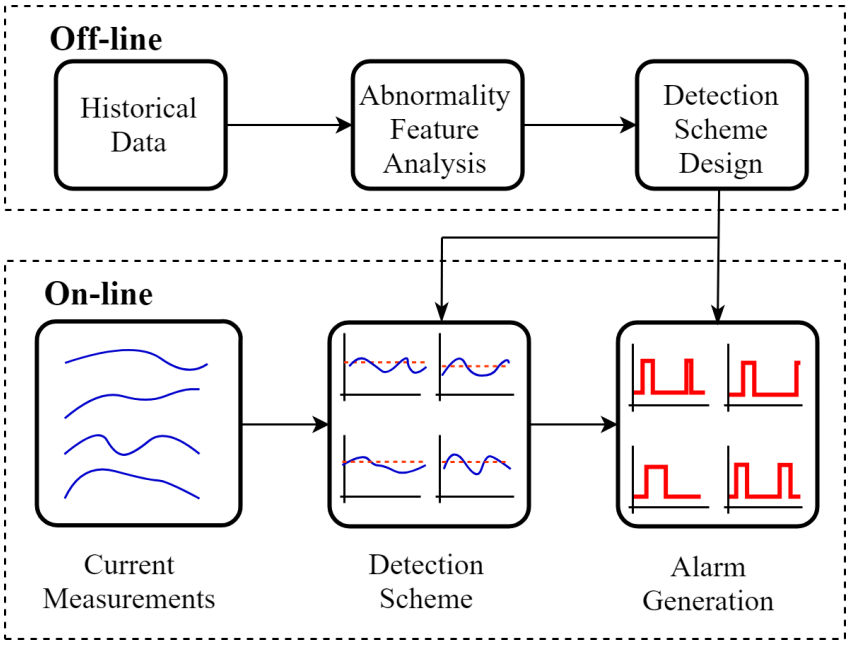



**Figure 1.1:** Typical abnormality detection system.

### 1.1.2 Current Issues and Challenges

Ideally, the abnormality detection system generates one alarm per abnormal situation so that operators can precisely take corresponding actions to bring the plant back to normal. However, most industrial abnormality detection systems suffer from an overwhelming number of false and nuisance alarms. This situation even gets worse for highly integrated and large-scale industrial plants with multi-operational modes. Specifically, according to [128], the average number of alarms per day should be less than 144 as per the standard benchmark; however, the actual statistics were 1200, 1500 and 2000 in oil-gas, petrochemical and power industries, respectively. As a crucial tool that operators highly rely on, high false and nuisance alarms distract operators' attention and lower the confidence level of operators on the detection system. By contrast, an abnormality detection system with low false and nuisance alarms spares more precious time for operators to tackle with important and vital alarms that induce severe events. In order to reduce the number of false and nuisance alarms on a plant level, it is necessary and significant to sys-

tematically and prudentially design abnormality detection schemes for each operating unit, control loop and key variable.

Nevertheless, the design of effective abnormality detection schemes is difficult due to several challenges including the following two as per [125] and [128].

- **Analytically relating the design of detection schemes with detection performance:** This, in other words, refers to the derivation of explicit expressions on the performance measures with respect to the parameters in the industrial plant and/or detection scheme. However, this task is complicated even for basic detection schemes such as dead-band [5], delay-timer [3] and filters [116], and can be more challenging for nonlinear and/or complex detection schemes [128]. Many existing studies evaluate the detection performance with case studies, which induces biased performance evaluations and a heavy computational burden.

- **Design of robust detection scheme:** A desirable abnormality detection system can handle uncertainties and noise [125]. Considering the fact that uncertainties and noise are inevitable in practice owing to computational errors, assumptions on models and measurement noise, it is desirable to guarantee that abnormality detection systems still acquire good detection ability and low false alarms even under the conditions with uncertainties and noise. However, it is rather difficult to quantify the robustness analytically and formulate the corresponding design problem due to unknown uncertainty resources and complex structures of detection schemes.

Targeting at the two challenges above, this thesis investigates analytical performance and sensitivity evaluations and the corresponding performance based design problems.

## 1.2 Literature Survey

Based on fruitful results on the abnormality detection topic, many surveys were written to review, summarize and classify core techniques. Some of the surveys aimed at categorizing the methods with general applications [50, 128, 125], while others focused on specific objectives such as the benchmark Tennessee Eastman process [138], waste water treatment plants [94], pipelines [93] and photovoltaic systems [97]. The method classifications are not unique, and can be found in several surveys such as [125], [50], [8] and [69]. In this thesis, abnormality detection methods are introduced with two categories, namely, model-based and data-based methods, since they cover most of the methods in the literature according to [125]. Accordingly, the taxonomy of detection methods is depicted in Figure 1.2. More details about each method are illustrated in the next two subsections, and the work related to performance analysis for those methods is reviewed in the third subsection.

### 1.2.1 Model-Based Methods

Model-based methods utilize mathematical models or system behavior descriptions for abnormality detection. Depending on the type of knowledge involved, the model-based methods can be further divided into quantitative model-based methods and qualitative model-based ones [125].

For the quantitative model-based methods, models can be obtained based on physical principles or system identification techniques. In those methods, the analytical redundancy is calculated with the help of the quantitative model to generate residuals for abnormality detection purposes [125]. According to the type of models considered in the detection scheme, it can be further classified into deterministic model-based methods, stochastic model-based methods, methods for discrete-events and hybrid systems, and methods for networked and distributed systems [50]. Specifically, in the deterministic model-based methods, observers [37, 46, 83] play an crucial role since 1995

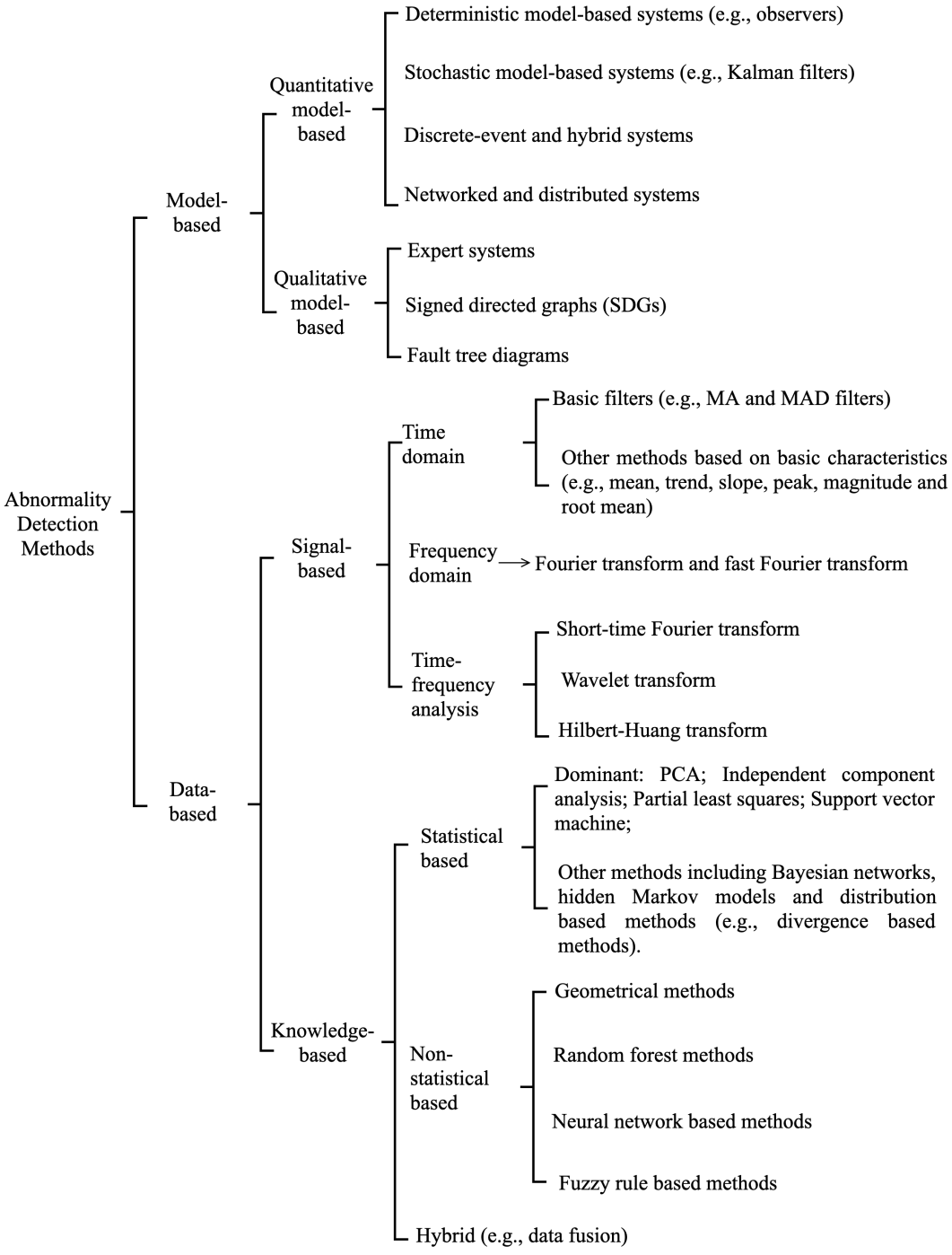



**Figure 1.2:** Taxonomy of abnormality detection methods.

[135]; while in the stochastic model-based methods, the Kalman filter based methods [25, 130, 17, 66] and model parameter estimation methods [67, 68] are frequently discussed in the literature. Comparing to the conventional continuous deterministic or stochastic models where variables are changing smoothly, the discrete-events or hybrid systems consider sampled data and/or events. It has been shown that event-based systems can improve the robustness of the abnormality detection system under large scale disturbances [90] and relieve computational burdens [6]. The development of methods for networked and distributed systems is stimulated by booming development in complex industrial systems, for which the goal is to find a fault signal generator that is robust against transmission delays, data dropouts and limited capacity of communication channels [49].

Different from the quantitative models, the qualitative models are usually derived based on the knowledge of system structures and behavior descriptions (first principles) [47]. An illustrative example of quantitative model based methods is the expert system, which is essentially a rule-based system incorporating expertise from experienced operators and system characteristics [84, 57]. Other widely used tools in this category are signed directed graphs (SDGs) [79, 134] and fault tree diagrams [42]. According to [41], a SDG represents causal and connectivity information of a process, where graph nodes denote the process variables and directed arcs stand for the causal relations. In [41], the SDG based method was combined with the qualitative trend analysis, and applied on an atmospheric distillation tower unit to illustrate its advantages. In [112], a methodology was proposed to generate SDGs based on bond graphs, which are widely used to present engineering processes and frameworks in industry; then, the abnormality detection performance was evaluated with a multi-energy process. Different from the SDG which is generated according to connectivity information, a fault tree starts from top or undesired events and terminates with primary or root causes [71]. A fault tree based method was proposed in [71] for automobile systems where parameters of subsystems

under normal and abnormal conditions were related to failure symptoms.

In general, model-based methods highly depend on the explicit mathematical models or sound system behavior knowledge, which requires intensive efforts and time to obtain [13]. Specifically, the model identification process relies on judgments of model assumptions and operational modes, which need long-term experience and inputs from operators. Besides, the model selection process can be time-consuming. To be specific, a sound physical model is usually highly nonlinear and complex to deal with, while an over-simplified version can miss descriptions of transient behavior. Thus, experimental validations and professional evaluations, which also take intensive time and efforts, are required to achieve a trade-off between model complexity and detection performance before a proper model is adopted and applied.

### 1.2.2 Data-Based Methods

Different from the model-based methods introduced above, data-based approaches can detect abnormalities with the help of historical process data, without prior process models [126]. Data-based methods can be further classified into signal-based and leaning-based ones.

Features extracted in the signal-based methods are basic statistical characteristics in the time domain (e.g., mean, trend, slope, peak, magnitude and root mean square) or spectrum-based patterns in the frequency domain [50]. For instance, the trend of slopes in the time domain was utilized in [54] to track the maximum power point in photovoltaic systems and achieved accurate tracking results as well as fast-tracking responses. Another illustrative example is the basic alarm filter including the moving average (MA) filters and mean absolute deviation (MAD) filters. In the frequency domain, spectrum-based patterns are usually obtained with the help of the Fourier transform, discrete Fourier transform or fast Fourier transform [26]. For example, in [119], the fast Fourier transform was applied on the on-line voltage, current, and vibration signals of a motor during operation to monitor the operational condition

of the motor. The proposed method was tested with on-site experiments and compared with commercial tools. Moreover, time-frequency analysis has been popularly used in abnormality detection for machinery [87] and pipelines [149]. It involves transformation tools such as the short-time Fourier transform [87], wavelet transform [148], Hilbert–Huang transform [109], Wigner-Ville distribution [117] and Choi-Williams distribution [117].

The learning-based methods are referred to those considering large size of historical data as per [50]. It can be further divided into statistical-analysis based methods, non-statistical based methods and hybrid methods. Specifically, under the statistical analysis framework, the principal component analysis (PCA), independent component analysis, partial least squares and support vector machine based methods dominate the abnormality detection area. Detailed introduction, comments and comparisons of those methods can be found in [138] and [50]. Besides, several other methods were studied in recent publications, such as regression based methods (e.g., auto-associative kernel regression [23]), correlation analysis based methods (e.g., canonical correlation analysis [74], correlation direction analysis [142], and mutual information analysis [72]), learning machines based methods (e.g., pick replace machine [64]), hidden Markov model based methods [137], Bayesian network based methods [92, 52, 10], sparse representation [115], swarm decomposition [91], symmetrical component analysis [114], and distribution analysis (e.g., Gaussian mixture model [140], entropy and divergence based methods [81, 59], and likelihood ratio based methods [45]). In the non-statistical based methods, there are geometrical methods [129], random forest methods [31], fuzzy rule based methods [99] and neural network based methods (e.g., artificial neural networks [150], convolutional neural networks [73], deep belief networks [141], and probabilistic neural networks [136]). The hybrid methods include those combining at least two methods from statistical and/or non-statistical based categories above, such as the data fusion based method studied in [60].

However, the detection performance is rarely studied analytically for most

of the above methods, and the detection threshold is usually tuned numerically, which can be computationally expensive and ineffective. This is mainly owing to the complicated detection procedure, which induces difficulties in relating detection statistics with the statistical properties of measurements. Fortunately, the detection performance of the basic filter based methods and divergence based methods can be analyzed based on the statistical properties of measured data. Particularly, the development of the two methods are reviewed in details below.

**Basic Filter Based Methods**

Belonging to the category of univariate signal-based methods, basic filters are widely used as the abnormality detection tools with a statistical framework.

Processing signals in the form of linear or nonlinear filters, control charts are frequently utilized in process monitoring, quality control and abnormality detection. The first control chart in the literature is the Shewhart control chart, which was first proposed by Walter A. Shewhart in 1920s [14]. For the Shewhart control chart, the control limit for measurements is usually set to be two or three times of the standard deviation. The other two popular control charts are the cumulative sum (CUSUM) control chart [35] and the exponentially weighted moving average (EWMA) control chart [102]. Compared with the Shewhart control chart, the latter two control charts are faster and more effective in detecting small shifts [113]. The CUSUM chart was first proposed by Page in 1954 [95] in the form of a nonlinear filter using the log-likelihood ratio between the two hypotheses for the abnormal and normal cases. Later, the EWMA chart, which is also called geometric moving average filter, was proposed by Roberts in 1959 [101] with an infinite impulse response filter structure [102]. The CUSUM chart treats past signals equally, while the EWMA chart wights the current data and the past average values differently. In the literature, studies about the CUSUM and EWMA control charts are mainly on investigating their applications in some certain distribu-

tions [63], dealing with correlations in data [106], studying adaptive detection schemes based on fault parameters [75] and combining the two control charts with other advanced methods such as wavelet representations [88] and Kalman filters [36].

Other than the three classic control charts, there are several univariate filters proposed in recent work. Specifically, a general optimal design framework for linear and quadratic alarm filters were proposed in [33] without any distribution knowledge. Later, based on the general optimal design framework in [33], MA filters were analytically proved to be optimal among linear filters under log-concave and symmetric data distributions [34]. In [12], the MAD filter was used to detect leakage in a pipeline; but the threshold tuning scheme was not mentioned. In [118], the rank order filter was designed to balance false alarms and detection delays without constraints on distributions. Then, as a special case of the rank order filter, the median filter was adopted to deal with the chattering alarms in [116].

**Divergence Based Methods**

While the aforementioned basic filters deal with changes in simple statistics such as the mean, median and variance, the divergence based methods focus on statistical changes in distributions. In order to capture the distribution change, the divergence from the current distribution to the known distribution under the normal condition can be adopted as the detection statistic.

One illustrative example is the Kullback -Leibler divergence (KLD). Closely related with the Shannon entropy, the KLD was firstly proposed in 1951 [80] and has been widely used in information theory and statistics for decades. It was recently applied for the abnormality detection purpose in power systems [82], glass melt plants [144] and bearing systems [38].

In the KLD-based detection scheme, the kernel density estimation can be utilized for the KLD calculation and its distribution estimation [44]. With the kernel density estimation, the distribution in a moving window and that under the normal condition are obtained by calculating the sum of wighted

kernel functions for each sample. Then, the KLD of the two numerical distributions is calculated via its definition in [80], and the distribution of the KLD can be obtained numerically by using a large number of data windows and another kernel density estimation of the KLD values. This tedious process causes a heavy computational burden even with the simplifications for the density ratio estimation proposed in [144, 59] and with the Monte Carlo simulation for the KLD calculation used in [139]. Compared to the numerical results using the kernel based method, tractable analytical results can relieve computational load and simulation imprecisions. Thus, some papers studied the KLD derivation and its distribution in an explicit form. In [144], the KLD derivation under the normal condition with multivariate identically Gaussian distributed data was obtained analytically as well as its asymptotic probability density function (PDF). Also, the KLD-based method was proved to be more sensitive to incipient faults than the Hotelling's T-squared statistic. In [29], the distribution of the symmetrical KLD was derived based on the Gaussian assumption and applied in detecting slight anomalous behavior in electrical traction systems. As for non-Gaussian distributions, there has been very limited work. In [38], the KLD under the Gamma distribution was derived and used for the fault detection, estimation and isolation of incipient single sensor failures. In [39], the KLD was calculated between two centralized generalized Gaussian distributions (GGDs). But the analytical distribution of the KLD was not obtained in either [38] or [39].

By adding the divergence order to the KLD, the Rényi divergence originally proposed by Rényi in 1961 [100] offers one more degree of freedom. It covers or relates with a wide range of divergences including the KLD, the squared Hollinger distance, $\chi^2$-divergence and the total variation distance [62]. It has wide applications including image registration [61], channel coding [105], unmanned aerial vehicle path planning [123] and moving object tracking [24]. In the abnormality detection area, it is mainly applied in distributed denial of service (DDoS) attack detections and bearing damage assessments. Precisely,

in [18], the Hartley entropy, Shannon entropy, Rényi entropy, generalized entropy, KLD and Rényi divergence were applied and compared in detecting both low-rate and high-rate DDoS attacks in a univariate scheme. In [15] and [16], the Rényi divergence was proposed to detect different types of DDoS attacks and flush events; then, the performance of the proposed method was evaluated with simulated case studies. The divergence order was selected in [15] based on the numerical FAR. As per [16], the Rényi divergence based method outperformed the KLD based one. In [22], the Rényi divergence of vibrational signals generated by Gaussian process models was used for bearing monitoring and life-time estimation. Later, in [111], the Rényi divergence was combined with the ensemble empirical model decomposition to detect the degradation of bearing using vibration data. In both [22] and [111], the divergence order was selected as a constant 0.5. As per the literature above, the relation between the divergence order and the detection performance has not been studied analytically, let alone the optimization problem formulation regarding to the divergence order.

### 1.2.3 Performance Measures

According to [33], the FAR and MAR are usually regarded as two of the most important metrics for alarm systems. The two metrics are closely related to other two practical performance measures, namely the 'in-control' average run length ($ARL_0$) and the 'out-of-control' average run length ($ARL_1$) [131]. Specifically, $ARL_0$ indicates the average number of observations taken before the process goes out of the control limit. According to [32], it is directly related with the FAR. As for the $ARL_1$, it measures the average number of observations taken before the process goes back in the control limit. It has been shown that the $ARL_1$ has a straightforward relation with the average detection delay [32], which can be expressed as a function of the MAR [2]. Additionally, the FAR and MAR are popularly used as performance measures for a variety of abnormality detection schemes (e.g., the linear filters [32],

time-deadbands [5], delay-timers [145] and point-targeted adaptive detectors [86]) and as key performance indicators in order to compare different detection schemes [146].

In terms of system robustness, it is usually quantified by sensitivity measures. In the literature, sensitivity measures can be generally classified into two types: local sensitivity measures and global sensitivity measures [21]. Global sensitivity measures are usually obtained based on statistical properties (such as variance ratio [21]) based on assigned known density distributions [104], which are unnecessary for local sensitivity measures [21]. Besides, the two types of measures have different applications. Local sensitivity measures are widely used in evaluating output variations due to slight changes in inputs, while global ones deal with input variations over their entire ranges of interest.

With the help of proper sensitivity measures, a sensitivity-based design problem can be formulated to meet diverse engineering needs. For example, in mechanical engineering, derivative-based local sensitivity measures are used to optimize the feasibility of a welded beam in [56]. In control engineering, the sensitivity of a control system to disturbances is usually reflected by a global sensitivity measure, namely, the sensitivity transfer function. For example, the weighted $\mathcal{H}_\infty$-norm of the sensitivity function was minimized to achieve optimal robustness with respect to control design in [143], and the 2-norm of the sensitivity function quantified the robustness of the assigned pole to parameter perturbations in a dynamical system in [1]. However, there are rarely papers on the sensitivity analysis for abnormality detection systems.

## 1.3 Thesis Contributions

As is mentioned in the previous section, there is limited work on the analytical performance analysis for abnormality detection systems, let alone the formulation of corresponding design problems. In the existing studies on the analytical performance analysis, there is no result on how to quantify the robustness of detection systems, and only a few articles on the analytical per-

formance analysis for non-linear and complex detection schemes, such as the divergence based methods. In this thesis, analytical evaluations of several performance measures (including sensitivity, FAR and MAR) are conducted for linear alarm filters, KLD-based detection and Rényi-based detection. Further, based on the obtained results on performance, new detection schemes are proposed where parameters are optimized with respect to the performance requirement or adjusted with respect to the data dynamics to achieve improved performance.

The major contributions in this thesis are summarized by chapters below:

In Chapter 2, the sensitivity of detection errors to changes in the trip point and to uncertainties in collected data is studied for linear alarm filters. Precisely, models are proposed to assess the sensitivity of system detection errors to changes in the trip point and to uncertainties in collected data, based on a derivative-based local sensitivity measure. Analytical expressions are derived to quantitatively evaluate the sensitivity of a general linear alarm filter with unknown data distributions. Then, a new sensitivity-based linear filter design method is formulated to minimize a weighted sum of the detection errors subject to upper bounds on the system sensitivities.

In Chapter 3, assuming that the observed data set under the normal condition contains samples of a GGD with the shape parameter larger than 1, an abnormality detection based on the KLD is studied without assumptions on abnormal conditions. The KLD between the estimated GGD and the one under the normal condition is used as the test statistic. An analytical expression for the KLD is derived when the number of samples is large. A constant threshold detection method is proposed where the unconditional PDF of the KLD under the normal condition is used to find the detection threshold by limiting the FAR. Besides, an adaptive threshold method are proposed whose detection threshold is adaptively adjusted based on the PDFs of the KLD conditioned on the estimated mean and variance, respectively.

In Chapter 4, with the assumption that observed data samples under the

normal condition are independent and identically distributed (i.i.d.) with a multivariate Gaussian distribution, the abnormality detection is studied based on the Rényi divergence with the divergence order between 0 and 1. The Rényi divergence between the estimated Gaussian distribution from the observed data and that under the normal condition is proposed as the test statistic. Both off-set faults and scaling faults are considered under the abnormal condition. The FAR and MAR are analytically derived as functions of the divergence order when the number of samples is large; then, an adaptive detection algorithm is proposed where the divergence order is optimized according to observed data considering both the FAR and MAR.

With the results of this thesis, the sensitivity of the detection performance over uncertainties can be analytically quantified and calculated for linear alarm filters. Besides, the distributions of the KLD and Rényi divergence can be calculated, which relieves computational burdens and improves efficiency compared with the conventional kernel-based or simulation-based evaluation. Further, with the analytical performance measures, the parameters can be optimized, which leads to improved detection performance compared with results in the existing literature.

# Chapter 2

# Sensitivity Analysis and Sensitivity-Based Design for Linear Alarm Filters*

RESEARCH topics on alarm triggering mechanism improvement, especially the advanced filter design with a level-crossing alarm generation mechanism, are still among the most promising ones to reduce nuisance and false alarms. Moreover, linear alarm filters are of particular interest since they are not only easy to implement but also effective in false alarm reduction. In this chapter, a local sensitivity measure is adopted to quantify the robustness of linear alarm filters, and a sensitivity based design scheme is proposed to achieve trade-offs between the sensitivity and detection errors.

## 2.1 Overview

The majority of existing results for linear alarm filters still rely on the assumption that the trip point and the PDFs of data under normal and abnormal conditions are known precisely. However, in reality, uncertainties and

*A version of this chapter has been published as: Ying Xiong, Yindi Jing, Tongwen Chen, Sensitivity analysis and sensitivity-based design for linear alarm filters, *Control Engineering Practice*, 71: 29-39, 2018. Part of this chapter has been published as: Ying Xiong, Yindi Jing, Tongwen Chen, Performance sensitivity analysis of linear alarm filters, *2017 American Control Conference (ACC)*, Seattle, USA, pages 4424-4429, 2017.

changes are inevitable in both the trip point and the PDF estimation. Specifically, computational errors exist in the trip point setting. For example, in the linear filter design scheme proposed in [34], the solution was found with iterative numerical methods [33], which offered approximated solutions instead of the exact optimal one due to limited precisions and processing power. As for the uncertainties in the PDFs, strictly speaking, the exact PDFs of process signals are unknown. The PDFs are actually estimated based on the collected data and pre-assumed probability models. But measurement noise in data collection is ubiquitous. Uncertainties in both the trip point and PDFs will propagate into the filter design and alarm performance evaluation. Thus, to understand how uncertainties affect the performance of an alarm system, it is important to quantitatively measure the sensitivity of the system behavior with respect to the trip point change and the PDF variations.

In this chapter, a local sensitivity measure is chosen instead of a global one to quantify the sensitivity. Because results obtained with global sensitivity measures highly rely on pre-assumed and limited types of distribution models; without known distributions, the calculation of global sensitivity measures is mathematically troublesome, if not impossible. However, distributions of the collected data are usually unknown and have relatively large diversity for different processes. Moreover, among existing local sensitivity measures, the derivative-based elasticity measure [21, 77] is adopted and customized for linear alarm filters. Specifically, the inputs of the elasticity are chosen as the trip point variations and the PDF offsets caused by noise in collected data. Then, the KLD, a popular distance measure between two probability distributions in information theory and statistics [80], is applied to measure the PDF offsets between the ideal noise-free data and the collected noisy data. The MAR and FAR are chosen as the outputs. Without any knowledge on the distributions of collected data, sensitivities of detection errors to the trip point and KLD are calculated based on the Gaussian kernel estimation method.

This chapter is organized as follows. Backgrounds on alarm monitoring

with linear filters and sensitivity analysis models are illustrated in Section 2.2. Then in Section 2.3, the proposed sensitivity models and the detailed calculation procedures are presented, together with the sensitivity-based linear filter design scheme. In Section 2.4, simulation results on data sets from Gaussian distributions and from an industrial plant are presented and studied. Section 2.5 summaries this chapter.

## 2.2 Linear Alarm Filters and Sensitivity Models

### 2.2.1 Linear Alarm Filters

Let $\boldsymbol{x} = [x(1), x(2), \dots]$ be a vector of discrete-time signals. A filter is essentially an operator that processes $\boldsymbol{x}$ to produce another vector discrete-time signals $\boldsymbol{y} = [y(1), y(2), \dots]$, written as $\boldsymbol{y} = F(\boldsymbol{x})$. In the alarm monitoring area, linear filters show significant advantages over non-linear filters [33] in easy application and simple computation. Therefore, linear filters are considered in this chapter, with the following form:

$$y(k) = \sum_{i=0}^{N-1} \theta_i x(k-i), \tag{2.1}$$

where $N$ is the length of the linear filter and $\theta_i$'s are the filter coefficients. Without loss of generality, the summation of all the filter coefficients is set to be 1, i.e., $\sum_{i=0}^{N-1} \theta_i = 1$. With the typical level crossing alarm generation mechanism [70], the filtered signal $y(k)$ is compared with the trip point $y_{tp}$ to trigger an alarm, i.e., an alarm will be raised if $y(k) > y_{tp}$.

In this chapter, collected data under the normal and abnormal modes are denoted as $\boldsymbol{x}_n = [a_1, a_2, \dots, a_{l_n}]$ with length $l_n$ and $\boldsymbol{x}_{ab} = [b_1, b_2, \dots, b_{l_{ab}}]$ with length $l_{ab}$ respectively. It is assumed that $a_i$'s are i.i.d., $b_j$'s are i.i.d., and $a_i$'s are independent to $b_j$'s. Let $f_{X,n}(\cdot)$ and $f_{X,ab}(\cdot)$ be the PDFs of $a_i$ and $b_j$ respectively. Also, let $\boldsymbol{y}_n = F(\boldsymbol{x}_n)$ and $\boldsymbol{y}_{ab} = F(\boldsymbol{x}_{ab})$ denote the filtered signals, and $f_{Y,n}(\cdot)$ and $f_{Y,ab}(\cdot)$ represent the common PDF of each element in $\boldsymbol{y}_n$ and $\boldsymbol{y}_{ab}$ respectively.

The FAR (MAR) of the alarm system is defined as the probability that an alarm is raised (missed) when the system is under the normal (abnormal) mode. The goal in conventional linear filter design is to find the optimal combination of filter coefficients and the trip point with the lowest weighted sum of FAR and MAR. With the trip point $y_{tp}$, the cost function can be calculated as:

$$\begin{aligned} \mathrm{J_{conv}} &= c_1 \mathrm{FAR} + c_2 \mathrm{MAR} \\ &= c_1 \int_{y_{tp}}^{+\infty} f_{Y,n}(y)\mathrm{d}y + c_2 \int_{-\infty}^{y_{tp}} f_{Y,ab}(y)\mathrm{d}y, \end{aligned} \tag{2.2}$$

where $c_1$ and $c_2$ are positive weights on the FAR and MAR. From (2.2), it is obvious that the FAR, the MAR and the cost function can be significantly affected by the accuracy of the data PDFs and the trip point. Uncertainties in the estimated PDFs and the trip point will lead to performance change. Thus, there exists a strong need to quantify the sensitivity to these changes.

### 2.2.2 Sensitivity Definition and Properties

In this chapter, the derivative-based elasticity, a local sensitivity measure, is adopted because it eliminates unit differences of input and output variations [20]. More rationale for this measure can be found in [21]. The measure is defined as:

$$S_O^I = \lim_{\Delta I \to 0} \frac{\Delta O}{O} \Big/ \frac{\Delta I}{I} = \frac{\mathrm{d}O}{\mathrm{d}I} \frac{I}{O}, \tag{2.3}$$

where $O$ stands for the output of the sensitivity model and $I$ represents the input with uncertainties. The ratio $I/O$ is used to scale the changes in the input and output over their current values to obtain a normalized dimensionless value. $S_O^I$ is naturally a sensitivity measure of $O$ to $I$ representing how much the output $O$ changes with the input $I$.

To complete the sensitivity modeling, it is necessary to specify inputs and outputs that are not only relevant but also capable of representing practical industrial needs. According to [21], model outputs should be variables of most interests to the decision maker. So the performance indices, the FAR and

MAR, are chosen as outputs for sensitivity analysis. As for the inputs, once the filter design is completed, candidates are the trip point and the estimated PDFs.

The sensitivity of the detection error over the trip point is straightforward to model by using the definition in (2.3):

$$S_{\mathrm{FAR}}^{y_{tp}} = \frac{\mathrm{dFAR}}{\mathrm{d}y_{tp}} \frac{y_{tp}}{\mathrm{FAR}}, \quad S_{\mathrm{MAR}}^{y_{tp}} = \frac{\mathrm{dMAR}}{\mathrm{d}y_{tp}} \frac{y_{tp}}{\mathrm{MAR}}. \tag{2.4}$$

The sensitivity over the estimated PDFs, however, needs further modeling since the PDF itself is a function not a scalar. For alarm systems, estimated PDFs are obtained from the collected data sets with random noise. Thus, the uncertainty in estimated PDFs can be represented by the distance between the PDF estimated from the noisy data and the real PDF. In this work, the KLD is applied to measure this distance. Specifically, let $f_{Y\epsilon,n}(y)$ be the estimated PDF using data without noise, and let $f_{Y,n}(y)$ represent the estimated PDF of raw data with noise. The KLD from $f_{Y\epsilon,n}(y)$ to $f_{Y,n}(y)$ is defined as:

$$D_{kl}(f_{Y,n}(y)||f_{Y\epsilon,n}(y)) = \int_{-\infty}^{+\infty} f_{Y,n}(y) \ln \frac{f_{Y,n}(y)}{f_{Y\epsilon,n}(y)} \mathrm{d}y. \tag{2.5}$$

According to the sensitivity definition in (2.3), the sensitivities of the FAR and MAR over the KLD are:

$$S_{\mathrm{FAR}}^{D_{kl}} = \lim_{\Delta D_{kl} \to 0} \frac{\Delta \mathrm{FAR}}{\Delta D_{kl}} \frac{D_{kl}}{\mathrm{FAR}}, \quad S_{\mathrm{MAR}}^{D_{kl}} = \lim_{\Delta D_{kl} \to 0} \frac{\Delta \mathrm{MAR}}{\Delta D_{kl}} \frac{D_{kl}}{\mathrm{MAR}}. \tag{2.6}$$

Next, properties of the sensitivity measures and insights they can provide are discussed. The elasticity of the FAR over the trip point, i.e., $S_{\mathrm{FAR}}^{y_{tp}}$ defined in (2.3), is used as an example.

1. Sign of the sensitivity measure: If the scaling term in the sensitivity measure (i.e., $y_{tp}$/FAR in (2.4)) is positive, the positive sign of the sensitivity measure $S_{\mathrm{FAR}}^{y_{tp}}$ means that the FAR increases with the trip point, and vise versa. Thus, the sign indicates the direction of the FAR changes with $y_{tp}$. If the scaling term is negative, the opposite holds.

2. Magnitude of the sensitivity measure: The magnitude is the percentage change ratio of the FAR over $y_{tp}$. A larger magnitude means that a more significant change occurs in the FAR with the same change in $y_{tp}$. For good robustness, a small sensitivity magnitude is desired.

3. Information from different sensitivity measures: Since different sensitivity measures reveal model behavior from different aspects, each measure deserves careful attention in terms of the sign, magnitude and trend. So, it is natural to analyze them together. A direct way is to define an overall sensitivity measure as:

$$S_{\text{overall}} = \sum\nolimits_{i,j} c_{i,j} S_i^j, \tag{2.7}$$

where $i \in \{\text{FAR}, \text{MAR}\}$ , $j \in \{y_{tp}, \text{KLD}\}$ and $c_{i,j}$'s are constants.

## 2.3 Sensitivity Analysis and Sensitivity-Based Design for Linear Alarm Filters

In this section, calculations of the sensitivity measures defined in (2.3) are provided for a generic linear alarm filter. The collected data under the normal and abnormal modes are $\boldsymbol{x}_n$ with length $l_n$ and $\boldsymbol{x}_{ab}$ with length $l_{ab}$ as introduced in Section 2.2. The coefficients of the linear filter $\boldsymbol{\theta} = [\theta_0, \theta_1, \ldots, \theta_{N-1}]$ are assumed to be known though arbitrary.

### 2.3.1 Sensitivity over Trip Point

Since the PDFs of the process signals are unknown, the Gaussian kernel function is applied to the PDF estimation using the collected data from the normal mode:

$$f_{X,n}(x) = \frac{1}{l_n} \sum\nolimits_{i=1}^{l_n} K(x - a_i), \tag{2.8}$$

where $K(x)$ is the Kernel function. In this chapter, the Kernel function is chosen as the standard Gaussian function [96]. Thus,

$$f_{X,n}(x) = \frac{1}{l_n}\frac{1}{\sqrt{2\pi}}\sum\nolimits_{i=1}^{l_n} \mathrm{e}^{-\frac{(x-a_i)^2}{2}}. \tag{2.9}$$

The characteristic function of the estimated PDF under the normal mode, denoted with $\psi_{X,n}(\cdot)$, can be calculated as:

$$\psi_{X,n}(t) = \int_{-\infty}^{+\infty} f_{X,n}(x)\mathrm{e}^{jxt}\mathrm{d}x. \tag{2.10}$$

With the help of (2.9) and properties of the characteristic function under Gaussian distribution [127], it can be derived that

$$\psi_{X,n}(t) = \frac{1}{l_n}\sum\nolimits_{k=1}^{l_n} \mathrm{e}^{-\frac{1}{2}t^2+jta_k}. \tag{2.11}$$

As $a_i$'s are i.i.d., with the linear filter in (2.1), the characteristic function of the filtered signal $\psi_{Y,n}(\cdot)$ is derived as:

$$\begin{aligned}\psi_{Y,n}(t) =& \prod\nolimits_{i=0}^{N-1} \psi_{X,n}(\theta_i t)\\ =& \left(\frac{1}{l_n}\right)^N \prod\nolimits_{i=0}^{N-1}\sum\nolimits_{k=1}^{l_n} \mathrm{e}^{-\frac{1}{2}t^2\theta_i^2+jta_k\theta_i}.\end{aligned} \tag{2.12}$$

To simplify the notation, let $\mathscr{X}_n^N = \{\boldsymbol{x} \mid \boldsymbol{x} = [x_1, \ldots, x_i, \ldots, x_N], x_i \in \boldsymbol{x}_n\}$, then (2.12) can be rewritten as:

$$\psi_{Y,n}(t) = \left(\frac{1}{l_n}\right)^N \mathrm{e}^{-\frac{1}{2}t^2\|\boldsymbol{\theta}\|^2}\sum\nolimits_{\boldsymbol{x}\in\mathscr{X}_n^N} \mathrm{e}^{jt\boldsymbol{\theta}\boldsymbol{x}^T}, \tag{2.13}$$

where $\|\cdot\|$ denotes the Euclidean norm of a vector.

The PDF of the filtered data from the normal mode can thus be calculated as follows:

$$\begin{aligned}f_{Y,n}(y) =& \frac{1}{2\pi}\int_{-\infty}^{+\infty} \mathrm{e}^{-jyt}\psi_{Y,n}(t)\mathrm{d}t\\ =& \left(\frac{1}{l_n}\right)^N \frac{1}{\sqrt{2\pi}\|\boldsymbol{\theta}\|}\sum\nolimits_{\boldsymbol{x_n}\in\mathscr{X}_n^N} \mathrm{e}^{-\frac{(y-\boldsymbol{\theta}\boldsymbol{x}^T)^2}{2\|\boldsymbol{\theta}\|^2}}.\end{aligned} \tag{2.14}$$

According to (2.2), the FAR of the linear alarm filter can be calculated to be:

$$\mathrm{FAR} = \left(\frac{1}{l_n}\right)^N \int_{y_{tp}}^{+\infty} \frac{1}{\sqrt{2\pi}\|\boldsymbol{\theta}\|}\sum\nolimits_{\boldsymbol{x}\in\mathscr{X}_n^N} \mathrm{e}^{-\frac{(y-\boldsymbol{\theta}\boldsymbol{x}^T)^2}{2\|\boldsymbol{\theta}\|^2}}\mathrm{d}y. \tag{2.15}$$

From the definition in (2.4), the sensitivity of the FAR over the trip point is eventually obtained as:

$$S_{\text{FAR}}^{y_{tp}} = -\left(\frac{1}{l_n}\right)^N \frac{y_{tp}}{\text{FAR}\sqrt{2\pi}\|\boldsymbol{\theta}\|} \sum\nolimits_{\boldsymbol{x}\in\mathscr{X}_n^N} \mathrm{e}^{-\frac{(y_{tp}-\boldsymbol{\theta}\boldsymbol{x}^T)^2}{2\|\boldsymbol{\theta}\|^2}}. \tag{2.16}$$

Via similar analysis, the sensitivity of the MAR over the trip point can be obtained as:

$$S_{\text{MAR}}^{y_{tp}} = \left(\frac{1}{l_{ab}}\right)^N \frac{y_{tp}}{\text{MAR}\sqrt{2\pi}\|\boldsymbol{\theta}\|} \sum\nolimits_{\boldsymbol{x}\in\mathscr{X}_{ab}^N} \mathrm{e}^{-\frac{(y_{tp}-\boldsymbol{\theta}\boldsymbol{x}^T)^2}{2\|\boldsymbol{\theta}\|^2}}, \tag{2.17}$$

where $\mathscr{X}_{ab}^N = \{\boldsymbol{x} \mid \boldsymbol{x} = [x_1, \ldots, x_i, \ldots, x_N], x_i \in \boldsymbol{x}_{ab}\}$.

### 2.3.2 Sensitivity over Noise Caused PDF Offsets

In this subsection, sensitivities of the performance over the PDF offsets are derived. In order to mimic the uncertainties in the estimated PDFs, the Gaussian noises contained in the collected data are denoted by $\boldsymbol{\epsilon_n} = [\epsilon_{1,n}, \cdots, \epsilon_{l_n,n}]$ and $\boldsymbol{\epsilon_{ab}} = [\epsilon_{1,ab}, \cdots, \epsilon_{l_{ab},ab}]$ respectively under the normal and abnormal conditions. For each data set, the noises from different time instants are assumed to be i.i.d. with the same distribution $\epsilon_n \sim \mathcal{N}(m, \sigma^2)$ and $\epsilon_{ab} \sim \mathcal{N}(m, \sigma^2)$. Thus, data before filtering is modeled as

$$\boldsymbol{x}_{n,f} = \boldsymbol{x}_n + \boldsymbol{\epsilon}_n;$$

$$\boldsymbol{x}_{ab,f} = \boldsymbol{x}_{ab} + \boldsymbol{\epsilon}_{ab},$$

under the normal and abnormal modes respectively. It is reasonable to assume that noise is independent of the collected data. Thus, under the normal mode, the output of the linear filter in (2.1) can be written as:

$$y_n(k) = \sum\nolimits_{i=0}^{N-1} \theta_i \left[x_n(k-i) + \epsilon_n(k-i)\right]. \tag{2.18}$$

Thus, by following similar steps in the previous subsection, the characteristic function of $y_n$ with Gaussian noise is denoted by $\psi_{Y\epsilon,n}(\cdot)$, which can be calculated as:

$$\begin{aligned}\psi_{Y\epsilon,n}(t) &= \left(\prod\nolimits_{i=0}^{N-1} \psi_{X,n}(\theta_i t)\right)\left(\prod\nolimits_{i=0}^{N-1} \psi_e(\theta_i t)\right) \\ &= \left(\frac{1}{l_n}\right)^N \sum\nolimits_{\boldsymbol{x}\in\mathscr{X}_n^N} \mathrm{e}^{-\frac{1}{2}(1+\sigma^2)t^2 + jt\left(\boldsymbol{\theta}\boldsymbol{x}^T + m\right)},\end{aligned} \tag{2.19}$$

where $\psi_e(\cdot)$ is the characteristic function of the Gaussian noise distribution. Accordingly, the PDF of the filtered data with noises is obtained as:

$$\begin{aligned} f_{Y\epsilon,n}(y) &= \frac{1}{2\pi}\int_{-\infty}^{+\infty} \mathrm{e}^{-jyt}\psi_{Y_\epsilon,n}(t)\mathrm{d}t \\ &= \left(\frac{1}{l_n}\right)^N \sum\nolimits_{\boldsymbol{x}\in\mathscr{X}_n^N} \frac{1}{\sqrt{2\pi(1+\sigma^2)}\|\boldsymbol{\theta}\|}\mathrm{e}^{-\frac{(y-\boldsymbol{\theta}\boldsymbol{x}^T-m)^2}{2(1+\sigma^2)\|\boldsymbol{\theta}\|^2}}. \end{aligned} \tag{2.20}$$

Thus, the FAR can be expressed as follows:

$$\begin{aligned} \mathrm{FAR} &= \left(\frac{1}{l_n}\right)^N \int_{y_{tp}}^{+\infty} f_{Y\epsilon,n}(y)\mathrm{d}y \\ &= \frac{1}{2} - \frac{1}{2}\left(\frac{1}{l_n}\right)^N \sum\nolimits_{\boldsymbol{x}\in\mathscr{X}_n^N} \mathrm{erf}\left(\frac{y_{tp}-\boldsymbol{\theta}\boldsymbol{x}^T-m}{\sqrt{2(1+\sigma^2)}\|\boldsymbol{\theta}\|}\right), \end{aligned} \tag{2.21}$$

where $\mathrm{erf}(\cdot)$ is the error function.

As is shown in (2.20), both the mean and standard deviation of the Gaussian noise offset the PDF of the filtered data. One method to quantify the offset is to further introduce the mean and variance as two inputs to capture the sensitivity of the FAR over the uncertainty in the PDF estimation. In this way, the sensitivities over the mean and variance of the noise are considered separately [132]. In this chapter, a different method is taken with the help of the KLD to model the PDF estimation offset caused by both the mean and standard deviation of the noise jointly.

From (2.6), the sensitivity of the FAR over the KLD can be expanded as follows according to the total differential theorem:

$$S_{\mathrm{FAR}}^{D_{kl}} = \lim_{\substack{\mathrm{d}m\to 0,\mathrm{d}\sigma\to 0 \\ \mathrm{d}m=\lambda\mathrm{d}\sigma}} \frac{\frac{\partial \mathrm{FAR}}{\partial m}\mathrm{d}m + \frac{\partial \mathrm{FAR}}{\partial \sigma}\mathrm{d}\sigma}{\frac{\partial D_{kl}}{\partial m}\mathrm{d}m + \frac{\partial D_{kl}}{\partial \sigma}\mathrm{d}\sigma}\frac{D_{kl}}{\mathrm{FAR}}, \tag{2.22}$$

where the condition $\mathrm{d}m = \lambda\mathrm{d}\sigma$ is to specify a linear relationship between $\mathrm{d}m$ and $\mathrm{d}\sigma$. This condition is necessary to guarantee that the limit is well defined. The coefficient $\lambda$ can be seen as a weighting factor to quantify the step size ratio of the mean over that of the standard deviation. A larger $\lambda$ means that a heavier weight is put on the standard deviation change than the mean change. In applications, $\lambda$ should be tuned larger in situations where the system is

more vulnerable to the variance of noise than to the mean. In this chapter, $\lambda = 1$ is chosen for simplicity, i.e., the mean and standard deviation changes are treated equally.

$$\frac{\partial \mathrm{FAR}}{\partial m} = \left(\frac{1}{l_n}\right)^N \frac{1}{\sqrt{2\pi(1+\sigma^2)}\|\boldsymbol{\theta}\|} \sum\nolimits_{\boldsymbol{x}\in\mathscr{X}_n^N} \mathrm{e}^{-\frac{(y_{tp}-\boldsymbol{\theta}\boldsymbol{x}^T-m)^2}{2(1+\sigma^2)\|\boldsymbol{\theta}\|^2}}, \tag{2.23}$$

$$\begin{aligned}\frac{\partial \mathrm{FAR}}{\partial \sigma} &= \left(\frac{1}{l_n}\right)^N \frac{\sigma}{(1+\sigma^2)} \frac{1}{\sqrt{2\pi(1+\sigma^2)}\|\boldsymbol{\theta}\|} \\ &\quad \times \sum\nolimits_{\boldsymbol{x}\in\mathscr{X}_n^N} (y_{tp}-\boldsymbol{\theta}\boldsymbol{x}^T-m)\mathrm{e}^{-\frac{(y_{tp}-\boldsymbol{\theta}\boldsymbol{x}^T-m)^2}{2(1+\sigma^2)\|\boldsymbol{\theta}\|^2}},\end{aligned} \tag{2.24}$$

$$\begin{aligned}\frac{\partial D_{kl}}{\partial m} &= \int_{-\infty}^{+\infty} \frac{f_{Y_n}(y)}{f_{Y_{\epsilon,n}}(y)} \left(\frac{1}{l_n}\right)^N \frac{1}{\sqrt{2\pi(1+\sigma^2)}\|\boldsymbol{\theta}\|} \\ &\quad \times \sum\nolimits_{\boldsymbol{x}\in\mathscr{X}_n^N} \frac{(y_{tp}-\boldsymbol{\theta}\boldsymbol{x}^T-m)}{(1+\sigma^2)\|\theta\|^2} \mathrm{e}^{-\frac{(y_{tp}-\boldsymbol{\theta}\boldsymbol{x}^T-m)^2}{2(1+\sigma^2)\|\boldsymbol{\theta}\|^2}} \mathrm{d}y,\end{aligned} \tag{2.25}$$

$$\begin{aligned}\frac{\partial D_{kl}}{\partial \sigma} &= \int_{-\infty}^{+\infty} \left(\frac{1}{l_n}\right)^N \frac{\sigma}{\sqrt{2\pi(1+\sigma^2)}(1+\sigma^2)\|\boldsymbol{\theta}\|} \frac{f_{Y_n}(y)}{f_{Y_{\epsilon,n}}(y)} \Bigg[\frac{1}{(1+\sigma^2)\|\boldsymbol{\theta}\|^2} \\ &\quad \times \sum_{\boldsymbol{x}\in\mathscr{X}_n^N} (y_{tp}-\boldsymbol{\theta}\boldsymbol{x}^T-m)^2 \mathrm{e}^{-\frac{(y_{tp}-\boldsymbol{\theta}\boldsymbol{x}^T-m)^2}{2(1+\sigma^2)\|\boldsymbol{\theta}\|^2}} - \sum_{\boldsymbol{x}\in\mathscr{X}_n^N} \mathrm{e}^{-\frac{(y_{tp}-\boldsymbol{\theta}\boldsymbol{x}^T-m)^2}{2(1+\sigma^2)\|\boldsymbol{\theta}\|^2}}\Bigg] \mathrm{d}y.\end{aligned} \tag{2.26}$$

Via tedious but straightforward calculations, each term in (2.22) can be obtained in (2.23) to (2.26) using (2.5) and (2.21). The details are omitted due to page limit.

Similarly, the sensitivity of the MAR over the KLD can be calculated as follows:

$$S_{MAR}^{D_{kl}} = \lim_{\substack{\mathrm{d}m\to 0, \mathrm{d}\sigma\to 0 \\ \mathrm{d}m=\lambda \mathrm{d}\sigma}} \frac{\frac{\partial \mathrm{MAR}}{\partial m}\mathrm{d}m + \frac{\partial \mathrm{MAR}}{\partial \sigma}\mathrm{d}\sigma}{\frac{\partial D_{kl}}{\partial m}\mathrm{d}m + \frac{\partial D_{kl}}{\partial \sigma}\mathrm{d}\sigma} \frac{D_{kl}}{\mathrm{MAR}}, \tag{2.27}$$

where the partial derivative terms are shown in (2.28) to (2.31).

$$\frac{\partial \mathrm{MAR}}{\partial m} = -\left(\frac{1}{l_{ab}}\right)^N \frac{1}{\sqrt{2\pi(1+\sigma^2)}\|\boldsymbol{\theta}\|} \sum\nolimits_{\boldsymbol{x}\in\mathcal{X}_{ab}^N} (y_{tp} - \boldsymbol{\theta}\boldsymbol{x}^T - m)\mathrm{e}^{-\frac{(y_{tp}-\boldsymbol{\theta}\boldsymbol{x}^T-m)^2}{2(1+\sigma^2)\|\boldsymbol{\theta}\|^2}}, \tag{2.28}$$

$$\frac{\partial \mathrm{MAR}}{\partial \sigma} = \left(\frac{1}{l_{ab}}\right)^N \frac{\sigma}{(1+\sigma^2)} \frac{1}{\sqrt{2\pi(1+\sigma^2)}\|\boldsymbol{\theta}\|} \sum\nolimits_{\boldsymbol{x}\in\mathcal{X}_{ab}^N} (y_{tp} - \boldsymbol{\theta}\boldsymbol{x}^T - m)\mathrm{e}^{-\frac{(y_{tp}-\boldsymbol{\theta}\boldsymbol{x}^T-m)^2}{2(1+\sigma^2)\|\boldsymbol{\theta}\|^2}},$$

$$\frac{\partial D_{kl}}{\partial m} = \int_{-\infty}^{+\infty} \frac{f_{Y_{ab}}(y)}{f_{Y_{\epsilon,ab}}(y)} \left(\frac{1}{l_{ab}}\right)^N \frac{1}{\sqrt{2\pi(1+\sigma^2)}\|\boldsymbol{\theta}\|} \sum\nolimits_{\boldsymbol{x}\in\mathcal{X}_{ab}^N} \frac{(y_{tp} - \boldsymbol{\theta}\boldsymbol{x}^T - m)}{(1+\sigma^2)\|\theta\|^2}\mathrm{e}^{-\frac{(y_{tp}-\boldsymbol{\theta}\boldsymbol{x}^T-m)^2}{2(1+\sigma^2)\|\boldsymbol{\theta}\|^2}}\mathrm{d}y, \tag{2.29}$$

$$\frac{\partial D_{kl}}{\partial \sigma} = \int_{-\infty}^{+\infty} \left(\frac{1}{l_n}\right)^N \frac{\sigma}{\sqrt{2\pi(1+\sigma^2)}(1+\sigma^2)\|\boldsymbol{\theta}\|} \frac{f_{Y_{ab}}(y)}{f_{Y_{\epsilon,ab}}(y)} \tag{2.30}$$

$$\times \left[\frac{1}{(1+\sigma^2)\|\boldsymbol{\theta}\|^2} \sum_{\boldsymbol{x}\in\mathcal{X}_{ab}^N} (y_{tp} - \boldsymbol{\theta}\boldsymbol{x}^T - m)^2 \mathrm{e}^{-\frac{(y_{tp}-\boldsymbol{\theta}\boldsymbol{x}^T-m)^2}{2(1+\sigma^2)\|\boldsymbol{\theta}\|^2}} - \sum_{\boldsymbol{x}\in\mathcal{X}_{ab}^N} \mathrm{e}^{-\frac{(y_{tp}-\boldsymbol{\theta}\boldsymbol{x}^T-m)^2}{2(1+\sigma^2)\|\boldsymbol{\theta}\|^2}}\right]\mathrm{d}y. \tag{2.31}$$

### 2.3.3 Sensitivity-Based Linear Filter Design

Linear alarm filter design should consider not only the detection errors, but also the tolerance to system uncertainties, namely, the sensitivities to both the trip point changes and noise. In this chapter, a new linear filter design is proposed to achieve the least detection errors given sensitivity requirements. Thus, the weighted sum of performance errors shown in (2.2) is taken as the objective function, while the sensitivity measures are used as the constraints in the filter design problem.

There are two constraints involved. One is on the weighted sum of sensitivity measures of the FAR and MAR over the trip point. Since the trip point is one of the optimization variables, the formulation of this constraint is straightforward. Another is on the weighted sum of the sensitivities of the FAR and MAR over the KLD. But sensitivity measures with respect to the KLD need

to be further modeled, because the mean and standard deviation of noise are system parameters that can take any value in a continuous range. In this chapter, the average sensitivities over certain mean and variance ranges are used. Assume that the mean and standard deviation of noise have uniformly distributions respectively under the normal mode, i.e., $m_n \sim U(m_{n_1}, m_{n_2})$ and $\sigma_n \sim U(\sigma_{n_1}, \sigma_{n_2})$. The average sensitivity of the FAR over the KLD can be obtained as:

$$\bar{S}_{FAR}^{D_{kl}} = \frac{1}{(m_{n_2} - m_{n_1})(\sigma_{n_2} - \sigma_{n_1})} \int_{m_{n_1}}^{m_{n_2}} \int_{\sigma_{n_1}}^{\sigma_{n_2}} S_{FAR}^{D_{kl}}(m, \sigma) \mathrm{d}m \mathrm{d}\sigma. \tag{2.32}$$

Similarly, by denoting the uniform distributions for the noise mean and the standard deviation in the abnormal mode as $m_{ab} \sim U(m_{ab_1}, m_{ab_2})$ and $\sigma_{ab} \sim U(\sigma_{ab_1}, \sigma_{ab_2})$, the average sensitivity of the MAR over the KLD can be expressed as:

$$\bar{S}_{MAR}^{D_{kl}} = \frac{1}{(m_{ab_2} - m_{ab_1})(\sigma_{ab_2} - \sigma_{ab_1})} \int_{m_{ab_1}}^{m_{ab_2}} \int_{\sigma_{ab_1}}^{\sigma_{ab_2}} S_{FAR}^{D_{kl}}(m, \sigma) \mathrm{d}m \mathrm{d}\sigma. \tag{2.33}$$

In practice, the ranges of the mean and variance can be defined from the knowledge of the disturbance or noise.

Thus, the new sensitivity-based filter design problem is formulated as:

$$\underset{\theta_0, \cdots, \theta_{N-1}, y_{tp}}{\operatorname{argmin}} \quad c_1 \mathrm{FAR} + c_2 \mathrm{MAR} \tag{2.34}$$

$$\text{suject to} \quad c_1 |S_{FAR}^{y_{tp}}| + c_2 |S_{MAR}^{y_{tp}}| \leq U_3, \tag{2.35}$$

$$c_1 |\bar{S}_{FAR}^{D_{kl}}| + c_2 |\bar{S}_{MAR}^{D_{kl}}| \leq U_4, \tag{2.36}$$

where $U_3$ and $U_4$ are upper bound values of the sensitivities constraints over the trip point and over the KLD, respectively. They are set as:

$$U_3 = c_3 \min_{\theta_0, \cdots, \theta_{N-1}, y_{tp}} \left( c_1 |S_{FAR}^{y_{tp}}| + c_2 |S_{MAR}^{y_{tp}}| \right),$$

$$U_4 = c_4 \min_{\theta_0, \cdots, \theta_{N-1}, y_{tp}} \left( c_1 |\bar{S}_{FAR}^{D_{kl}}| + c_2 |\bar{S}_{MAR}^{D_{kl}}| \right),$$

where $c_3$ and $c_4$ are coefficients to adjust the constraints. In this model, the upper bounds depend on the minimum weighted sum of the sensitivity

values. Notice that $c_3 \geq 1$ and $c_4 \geq 1$ are necessary for the feasibility of the optimization.

In (2.34), the $c_i$'s are weighting coefficients. They should be set and adjusted to meet practical engineering needs. Specifically, the weight ratio $c_1/c_2$ should be larger if the cost for false alarms is higher than that for miss alarms. Since the FAR and MAR are scaled by $c_1$ ad $c_2$, their corresponding sensitivities are added up with the same weights $c_1$ and $c_2$. Besides, the coefficients $c_3$ and $c_4$ are directly related to the sensitivity requirements. Smaller coefficients $c_3$ and $c_4$ indicate stronger sensitivity requirement on the corresponding uncertainties.

However, the ratio $c_3/c_4$ does not have straightforwardly physical meaning unlike $c_1/c_2$, since the sensitivity over the trip point can take a totally different scale with that over the KLD. For example, when coefficients $c_3$ and $c_4$ are set to be equal, it is likely that one of the sensitivity constraint is not effective anymore, which means that an arbitrary linear filter can meet this sensitivity requirement; while another constraint is still too strong to allow any feasible solution. In order to quantify strength levels of two sensitivity constraints and to reach an appropriate balance between them, firstly, the ranges of $c_3$ and $c_4$ should be calculated. Define

$$c_3^{\max} = \frac{\max\limits_{\theta_0,\cdots,\theta_{N-1},y_{tp}} (c_1 S_{FAR}^{y_{tp}} + c_2 S_{MAR}^{y_{tp}})}{\min\limits_{\theta_0,\cdots,\theta_{N-1},y_{tp}} (c_1 S_{FAR}^{y_{tp}} + c_2 S_{MAR}^{y_{tp}})},$$

$$c_4^{\max} = \frac{\max\limits_{\theta_0,\cdots,\theta_{N-1},y_{tp}} (c_1 \bar{S}_{FAR}^{D_{kl}} + c_2 \bar{S}_{MAR}^{D_{kl}})}{\min\limits_{\theta_0,\cdots,\theta_{N-1},y_{tp}} (c_1 \bar{S}_{FAR}^{D_{kl}} + c_2 \bar{S}_{MAR}^{D_{kl}})},$$

which are the maximum values of $c_3$ and $c_4$ for the constraints to be functional. Thus, the ranges for $c_3$ and $c_4$ are $c_3 \in [1, c_3^{\max}]$ and $c_4 \in [1, c_4^{\max}]$ respectively. Secondly, different sensitivity constraint levels can be described by the corresponding percentages over the entire coefficient range denoted by $\mu_3$ and $\mu_4$. Then, the two coefficients should be set as $c_3 = 1 + \mu_3(c_3^{\max} - 1)$ and $c_4 = 1 + \mu_4(c_4^{\max} - 1)$. The sensitivity levels over the trip point and the KLD can be adjusted by $\mu_3$ and $\mu_4$. If the sensitivity requirement over the

trip point is higher than that over the KLD, $\mu_3$ should be set to be smaller than $\mu_4$, and vise versa. When the sensitivity requirement over the trip point is treated as equal to that over the KLD, then $\mu_3$ should be set to be equal to $\mu_4$. Accordingly, given percentage values of $\mu_3$ and $\mu_4$, $U_3$ and $U_4$ can be further calculated as:

$$U_3 = [1 + \mu_3(c_3^{\max} - 1)] \min_{\theta_0,\cdots,\theta_{N-1},y_{tp}} \left( c_1 |S_{FAR}^{y_{tp}}| + c_2 |S_{MAR}^{y_{tp}}| \right), \tag{2.37}$$

$$U_4 = [1 + \mu_4(c_4^{\max} - 1)] \min_{\theta_0,\cdots,\theta_{N-1},y_{tp}} \left( c_1 |\bar{S}_{FAR}^{D_{kl}}| + c_2 |\bar{S}_{MAR}^{D_{kl}}| \right). \tag{2.38}$$

To solve the constrained optimization problem in (2.34), a grid search method shown in Algorithm 1 is used. In the grid generation, $N-1$ dimensions of the hypercube $[-1, 1]^{N-1}$ are considered for the filter coefficients, since the summation of the filter coefficients is normalized to be 1. Let $s$ be the step size of the grid search in each coefficient dimension. There are $p^{N-1}$ candidate vectors for the filter coefficients, where $p = 2/s$. Another dimension of the gird is the trip point. Let its range be $[y_1, y_2]$ and the step size be $y_s$. There are $(y_2 - y_1)/y_s$ candidate values for the trip point. Firstly, for every possible filter vector, the characteristic functions and the PDFs for both normal and abnormal data sets are calculated. Then, for every possible trip point value, the four sensitivity measures are calculated. In this step, sensitivities of the FAR and MAR over the KLD are calculated separately. Precisely, the mean sensitivity of the FAR over the KLD is calculated with (2.32) and that of the MAR is calculated with (2.33). Following that, two pairs of weighted summations are compared with the corresponding sensitivity requirements in (2.35) and (2.36). If both requirements are satisfied, the weighted detection error performance is calculated, the filter coefficient vector is saved, together with its corresponding feasible trip point range. For every filter vector, the trip point within the feasible range that with the minimum weighted sum of the detection errors and the corresponding optimal trip point are found and saved. By searching over all saved filter coefficients, one with the minimum weighted summation of detection errors is discovered.

**Algorithm 1** Sensitivity-based filter design

**Input**: Data collected under normal and abnormal condition: $\boldsymbol{x}_n, \boldsymbol{x}_{ab}$

**Output**: Solution of the optimization problem (2.34) ($\boldsymbol{\theta}^*$ and $y_{tp}^*$)

**Initialization**: Generate $p^{N-1}$ candidate vectors $\boldsymbol{\theta}$ with $\theta_i \in [-1, 1]$;
$i = 1$; $\leftarrow$ Save the number of feasible filters.

**for** $j \leftarrow 1$ **to** $p^{N-1}$ **do**
  For the $j$ th candidate of the filter coefficient vector $\boldsymbol{\theta}_j$, calculate $f_{Y_n}(y)$ and $f_{Y\epsilon,n}(y)$ by (2.14) and (2.19);
  Similarly, obtain $f_{Y_{ab}}(y)$ and $f_{Y\epsilon,ab}(y)$;
  **for** $k \leftarrow 1$ **to** $(y_2 - y_1)/y_s$ **do**
    Set $y_{tp}$ to be the $k$ th candidate of the trip point;;
    Calculate $S_{FAR}^{y_{tp}}$ and $\bar{S}_{FAR}^{D_{kl}}$ by (2.16) and (2.32);
    Calculate $S_{MAR}^{y_{tp}}$ and $\bar{S}_{MAR}^{D_{kl}}$ by (2.17) and (2.33);
    **if** *(2.35) and (2.36) hold* **then**
      $\boldsymbol{\Theta}(i,:) = \boldsymbol{\theta}_j$;
      $\boldsymbol{Y}_{tp}(i,k) = y_{tp}$;
      $J(i,k) = c_1\text{FAR} + c_2\text{MAR}$ ;
      $[J_{min}(i), m] = \min(J(i,:))$; $\leftarrow$ Save the minimum cost function for each filter.
      $M(i) = m$; $\leftarrow$ Save the optimal trip point for each filter.
      $i = i + 1$; **close**;
  **end**
**end**
$[J_{min}, n] = \min(J_{min}(:))$; $\leftarrow$ Find the minimum cost function in all feasible filters;

**Return** $\boldsymbol{\theta}^* = \boldsymbol{\Theta}(n,:)$ and $y_{tp}^* = \boldsymbol{Y}_{tp}(n, M(n))$.

## 2.4 Case Study

In this section, simulation results based on data sets from Gaussian distribution and from an industrial plant are shown to verify the analytical results and perform the proposed filter design procedures in Section 2.3. Two widely used linear filters, the MA filter and the linear weighted (LW) filter [120], are considered and compared with the proposed optimal filter. The MA filter has equal coefficients, shown as:

$$y(k) = \frac{1}{N}\sum\nolimits_{i=0}^{N-1} x(k-i); \tag{2.39}$$

while the LW filter has ascending coefficients, expressed as:

$$y(k) = \frac{1}{N^2} \sum_{i=0}^{N-1} (i+1)x(k-i). \tag{2.40}$$

The LW filter reduces both the mean and variance of data, while the MA filter only changes the variance and keeps the mean unchanged. The variance reduction effect of the LW filter is stronger than the MA filter. In this sense, the LW filter can achieve lower FAR and MAR than the MA filter for some asymmetric distributions, making the LW filter a competition to the MA filter.

Besides, there is no constraints in terms of the filter length; however, longer length induces greater alarm delay, and how to analytically achieve the trade-off between the detection errors and the alarm delay is out of the scope of this chapter. Similar to the order selection for delay-timers in [3], a simple way is to choose the least order of the MA filter that fits the detection error requirements based on industrial needs. In the simulation, the detection error requirements are set as FAR $\leq 10\%$ and MAR $\leq 10\%$ for both Gaussian and industrial data sets. With the help of the receiver operating characteristic (ROC) curve (details can be referred to [3]), the order of $N = 5$ has been chosen and considered to verify the effectiveness of the proposed method. Also, the sensitivity measures with different orders ($N = 4, N = 5$ and $N = 6$) are evaluated and compared.

### 2.4.1 Simulation Results on Sensitivity Measures

In order to check the effectiveness of the sensitivity analysis in Section 2.3, Gaussian distributions are considered firstly. The data sets under the normal and abnormal modes are generated from Gaussian distributions, where $x_n \sim \mathcal{N}(6, 9)$ and $x_{ab} \sim \mathcal{N}(10, 16)$. The size of each set is 1000, i.e., $l_n = l_{ab} = 1000$.

#### Sensitivity over Trip Point

In Fig. 2.1, both proposed and theoretical sensitivity results over the trip point using the MA filter are shown. The proposed results are obtained from (2.16) and (2.17), where the distribution knowledge is unknown. For

the theoretical curves, the PDFs of the data are assumed to be known and denoted with $x_n \sim \mathcal{N}(m_n, \sigma_n^2)$ and $x_{ab} \sim \mathcal{N}(m_{ab}, \sigma_{ab}^2)$. Via straightforward calculations according to (2.4), the sensitivity of the FAR and MAR over the trip point can be calculated as

$$S_{\text{FAR}}^{y_{tp}} = -\frac{y_{tp}}{\text{FAR}\sqrt{2\pi}\sigma_n\|\boldsymbol{\theta}\|}\text{e}^{-\frac{(y_{tp}-m_n)^2}{2\|\boldsymbol{\theta}\|^2\sigma_n^2}},$$

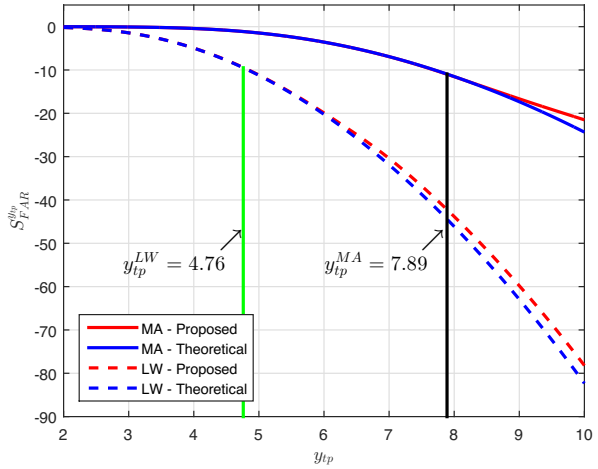
$$S_{\text{MAR}}^{y_{tp}} = \frac{y_{tp}}{\text{MAR}\sqrt{2\pi}\sigma_{ab}\|\boldsymbol{\theta}\|}\text{e}^{-\frac{(y_{tp}-m_{ab})^2}{2\|\boldsymbol{\theta}\|^2\sigma_{ab}^2}}.$$

Firstly, it can be observed that the derived sensitivity formulas tightly match the theoretical ones for most of the parameter range. The small differences in Fig. 2.1 are caused by errors in the PDF estimation with the Gaussian kernel method. The gap can be further reduced by careful tuning of the bandwidth in the kernel function (2.8) [58]. This verifies the effectiveness of the proposed sensitivity analysis under unknown distribution cases.

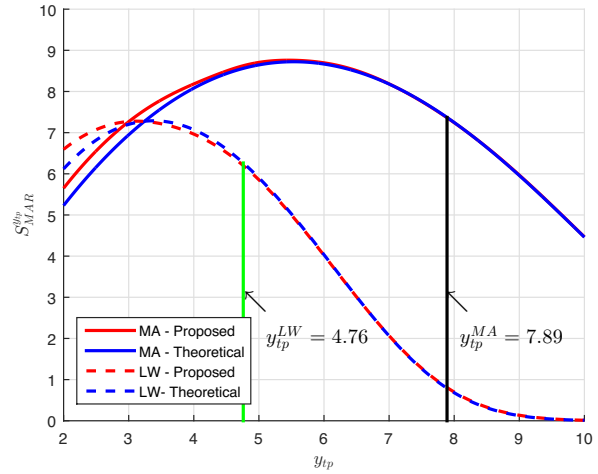
Secondly, properties of sensitivity measures in Section 2.2 can be used to understand sensitivities of the FAR and MAR. According to Fig. 2.1a, the sensitivity of the FAR has a downward trend with a negative sign. This means that the FAR decreases when the trip point grows. And the increasing magnitude means that the FAR gets more sensitive as the trip point rises. Similar analysis is applicable for the MAR in Fig. 2.1b. The magnitudes of the sensitivities show that the MAR is less sensitive than the FAR with respect to the trip point change.

In addition, the sensitivities of the two filters are compared. When the objective function in (2.2) has equal weights, i.e., $c_1 = c_2$, the optimal trip points for the MA filter and the LW filter are $y_{tp}^{\text{MA}} = 7.89$ and $y_{tp}^{\text{LW}} = 4.76$ respectively. The sensitivity values of the FAR and MAR for the two filters at their optimal trip points are listed in Table 2.1. The derived results in (2.16) and (2.17) are used for the calculations. It can be seen that the LW filter is less sensitive to the trip point variation.

Moreover, the trend of sensitivity changing with different filter orders can be observed from Fig. 2.2. Specifically, in Figs. 2.2a and 2.2b, the weighted




**(a)**




**(b)**

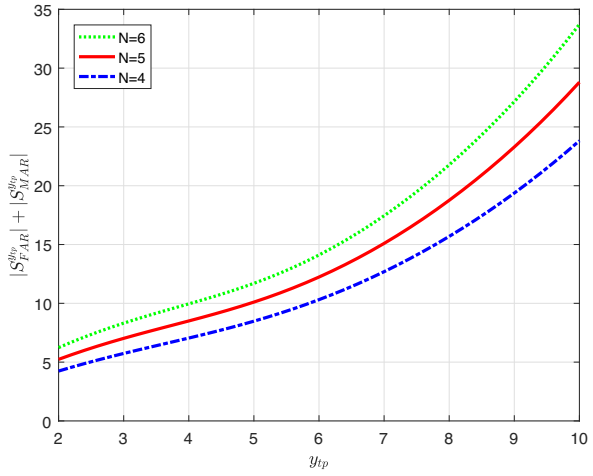
**Figure 2.1:** Sensitivity over the trip point for MA and LW filters: (a) FAR; (b) MAR.

**Table 2.1:** Sensitivity values at the optimal trip points for MA and LW filters under Gaussian distributions.

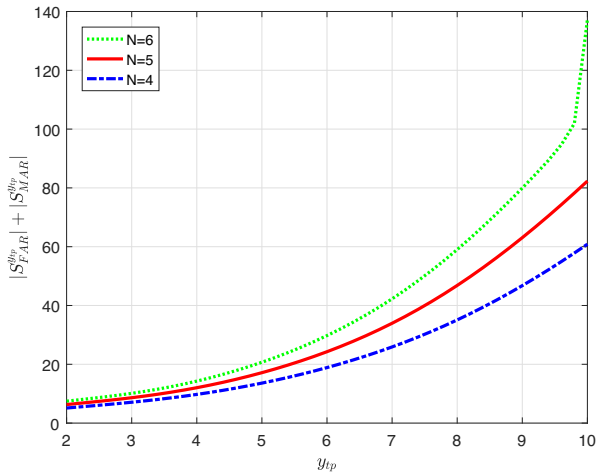
| Filter | $S_{\text{FAR}}^{y_{tp}}$ | $S_{\text{MAR}}^{y_{tp}}$ | $\lvert S_{\text{FAR}}^{y_{tp}}\rvert + \lvert S_{\text{MAR}}^{y_{tp}}\rvert$ |
|---|---|---|---|
| MA Filter$\lvert y_{tp} = 7.89$ | -10.95 | 7.20 | 18.15 |
| LW Filter$\lvert y_{tp} = 4.76$ | -9.48 | 5.55 | 15.03 |

summation of the sensitivity magnitude of the FAR and MAR shown in (2.35) is calculated with $c_1 = c_2 = 1$ for the MA and LW filters, respectively. Both

figures show that the magnitude of the sensitivity measures become larger with higher filter orders. This means both the MA and LW filters tend to be more sensitive to the trip point change when the order increases.




**(a)**

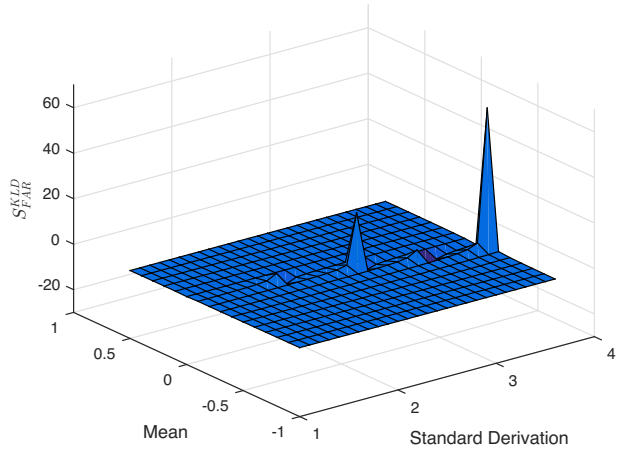



**(b)**

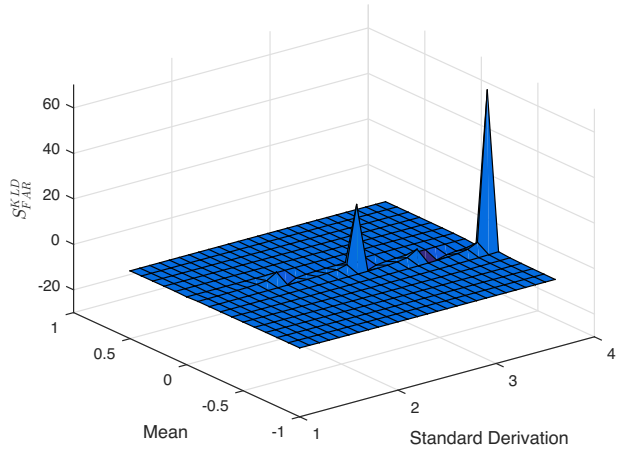
**Figure 2.2:** Sensitivity over the trip point for MA and LW filters with different orders: (a) MA; (b) LW.

### Sensitivity over Noise

In this subsection, the sensitivity over noise is studied, where both the proposed analysis and the theoretical analysis are used. In the simulations, the ranges of the mean and standard deviation of the noise are chosen as




**(a)**




**(b)**

**Figure 2.3:** Sensitivity of the FAR over the KLD for the MA filter: (a) proposed results without PDF information; (b) theoretical results with PDF information.

$m \in [-0.2\bar{m}, 0.2\bar{m}]$ and $\sigma \in [0.8\bar{\sigma}, 1.2\bar{\sigma}]$, where $\bar{m}$ and $\bar{\sigma}$ represent the mean and standard deviation of the data set respectively. Thus, the noise mean and standard deviation are within 20% of the counterparts of data. The step sizes for the mean and standard deviation are set as $\mathrm{d}m = \mathrm{d}\sigma = 0.1$ to calculate the sensitivity values over the KLD.

Considering Gaussian noise $\epsilon_n \sim \mathcal{N}(m, \sigma^2)$, the filtered data is also Gaussian distributed with $y_{\epsilon,n} \sim \mathcal{N}(m + m_n, \sigma^2\|\boldsymbol{\theta}\| + \sigma_n^2\|\boldsymbol{\theta}\|)$. With the help of the error function, the partial differentiations of the FAR over the noise mean and

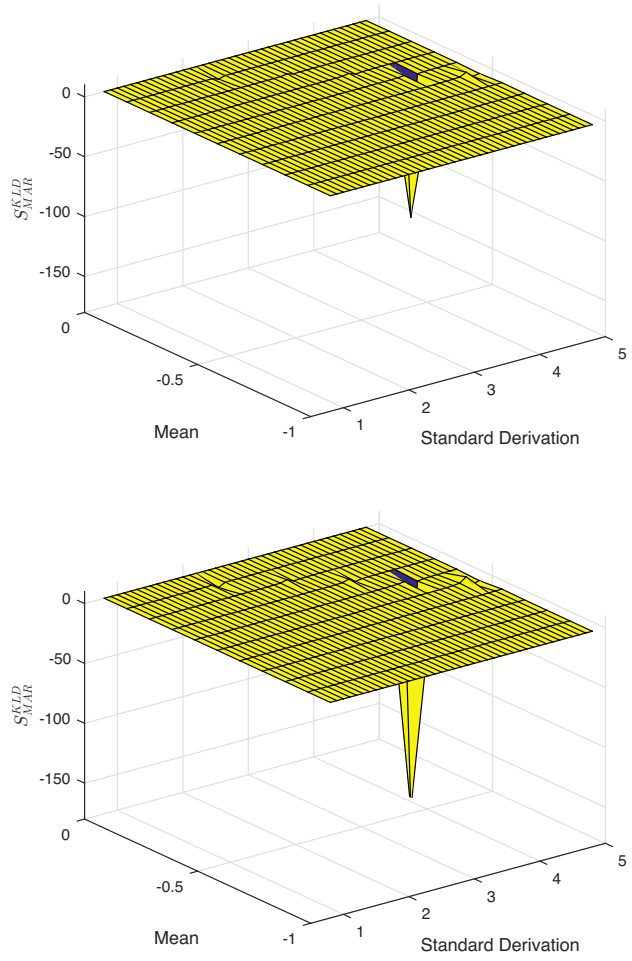



**Figure 2.4:** Sensitivity of the MAR over the KLD for the MA filter: (a) proposed results without PDF information; (b) theoretical results with PDF information.

variance in (2.22) are obtained as

$$\frac{\partial \text{FAR}}{\partial m} = \frac{1}{\sqrt{2\pi(\sigma_n^2+\sigma^2)}\|\boldsymbol{\theta}\|} \mathrm{e}^{-\frac{(y_{tp}-m_n-m)^2}{2\|\boldsymbol{\theta}\|^2(\sigma_n^2+\sigma^2)}},$$
$$\frac{\partial \text{FAR}}{\partial \sigma} = \frac{\sigma}{(\sigma_n^2+\sigma^2)} \frac{(y_{tp}-m_n-m)}{\sqrt{2\pi(\sigma_n^2+\sigma^2)}} \mathrm{e}^{-\frac{(y_{tp}-m_n-m)^2}{2\|\boldsymbol{\theta}\|^2(\sigma_n^2+\sigma^2)}}.$$

Besides, according to [39], the closed form expression of the KLD with the filtered Gaussian data are

$$D_{KL}(f_{Y,n}(y)||f_{Y\epsilon,n}(y)) = \frac{1}{2}\ln\frac{\sigma_n^2+\sigma^2}{\sigma_n^2} + \frac{\|\boldsymbol{\theta}\|^2\sigma_n^2+m^2}{2\|\boldsymbol{\theta}\|^2(\sigma_n^2+\sigma^2)}.$$

Then it is straightforward to get the partial differentiations of the KLD over the noise mean and standard deviation as

$$\frac{\partial D_{kl}}{\partial m} = \frac{m}{\|\boldsymbol{\theta}\|^2(\sigma_n^2+\sigma^2)},$$
$$\frac{\partial D_{kl}}{\partial \sigma} = \frac{\sigma}{\sigma_n^2+\sigma^2} - \frac{\sigma(\|\boldsymbol{\theta}\|^2\sigma_n^2+m^2)}{\|\boldsymbol{\theta}\|^2(\sigma_n^2+\sigma^2)^2}.$$

With the same filter and trip point, the results with respect to the MAR in (2.27) can be calculated by similar steps. For the MA filter at the trip point $y_{tp}^{\mathrm{MA}} = 7.89$, the sensitivities of the FAR and MAR over the KLD are shown in Figs. 2.3 and 2.4 respectively. The proposed results are shown in Figs. 2.3a and 2.4a; while he theoretical results are depicted in Figs. 2.3b and 2.4b.

Additionally, when the trip point ranges from 2 to 10 with the step size 0.01, the average sensitivity of the FAR over the KLD calculated with (2.32) using both the MA and the LW filters are plotted in Fig. 2.5. Similarly, the average sensitivity of the MAR over the KLD calculated with (2.33) are depicted in Fig. 2.6.

Firstly, from Figs. 2.3 and 2.4, it can be observed that the sensitivities derived with the proposed method can precisely catch the trend of the theoretical one. Specifically, peaks of both methods appear synchronously. The differences in the magnitudes of the peaks are caused by the limited size of the data set. It has been tested that the differences decrease with larger data size.

Secondly, the effectiveness of the proposed method is verified by the average sensitivity curves shown in Figs. 2.5 and 2.6. It is obvious that for both the MA and the LW filters, the proposed sensitivity results with respect to the FAR and MAR can tightly match the ones derived with the PDF information. The differences are also caused by the limited size of the data set.

Additionally, from Figs. 2.5 and 2.6, it can be concluded that the average sensitivity values of the LW filter over the KLD are smaller than those of the MA filter in most of the entire trip point range in terms of the sensitivity of the FAR over the KLD. As for the sensitivity of the MAR, the magnitude of

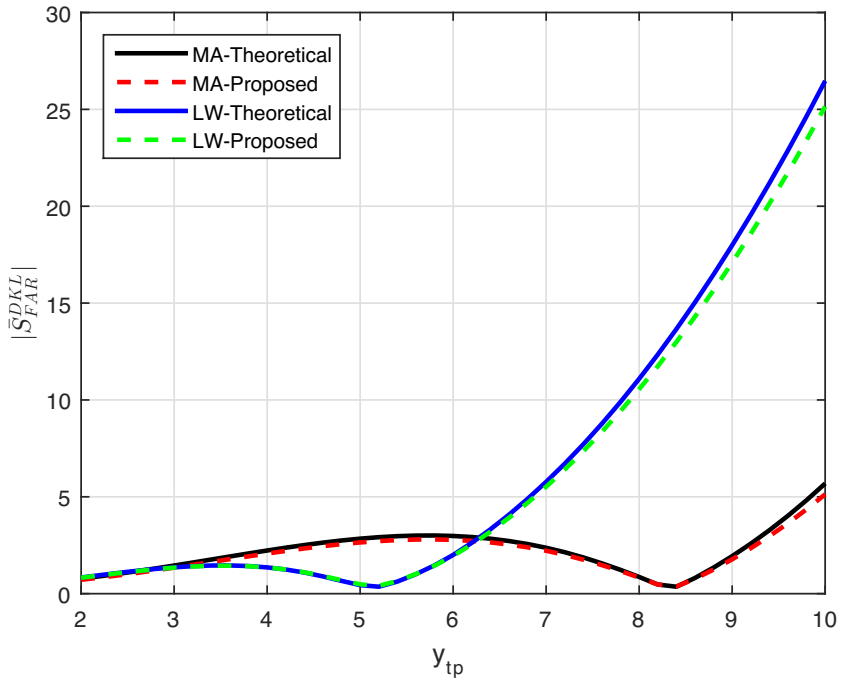



**Figure 2.5:** The average sensitivity of the FAR over the KLD for the MA and LW filters.

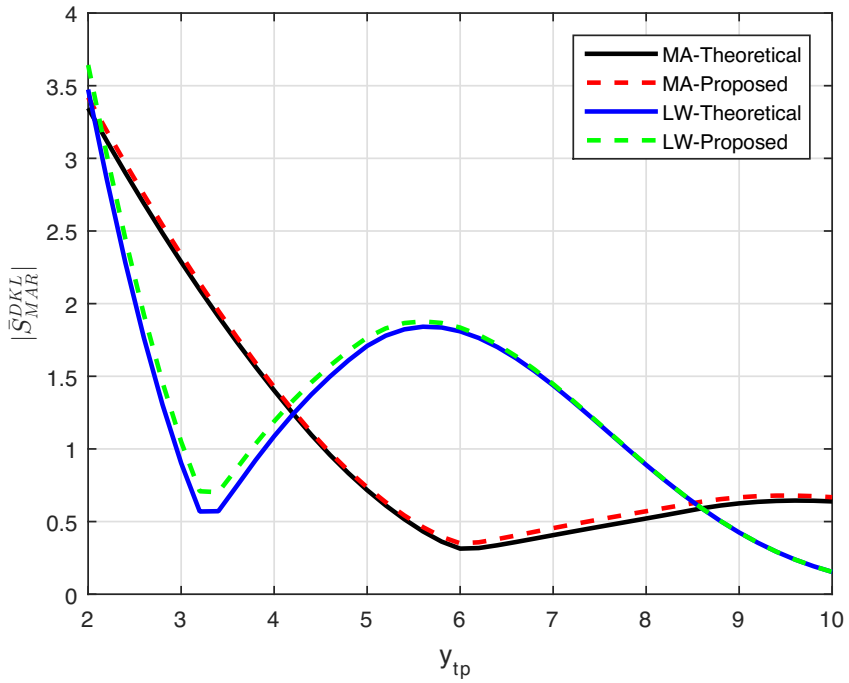



**Figure 2.6:** The average sensitivity of the MAR over the KLD for the MA filter.

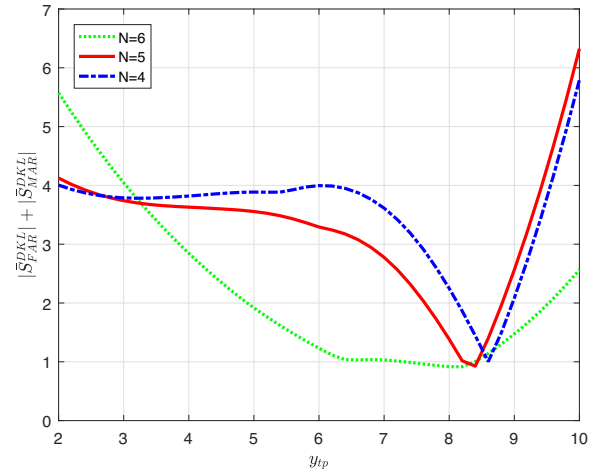
the LW filter is greater than that of the MA filter for the trip point ranging from 2 to 7, and it becomes smaller than that of the MA filter between 7 to 10. This reveals that the LW filter is less sensitive in terms of the FAR but more sensitive in the MAR case.

Meanwhile, the sensitivity trends of the MA and LW filters with different orders are shown Fig. 2.7 by setting $c_1 = c_2 = 1$ in (2.36). It can be seen from Fig. 2.7a that higher order MA filters gets less sensitive to noise in the collected data, while the LW filter becomes more sensitive to noise according to Fig. 2.7b. One explanation for this phenomenon is that the noise reduction effect of the MA filters gets more powerful when the order increases.

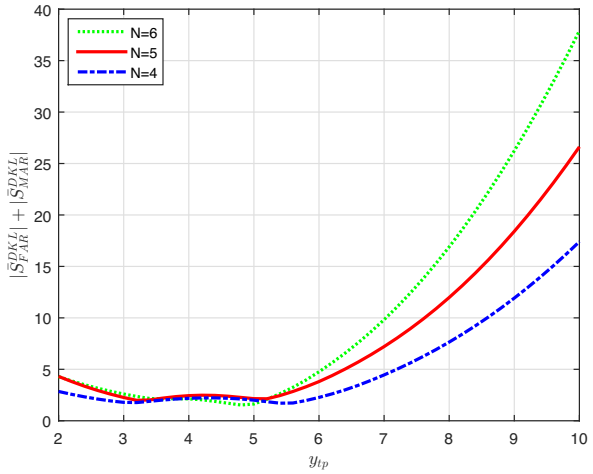
### 2.4.2 Sensitivity-Based Filter Design under Gaussian Distributions

In this subsection, the sensitivity-based design procedure introduced in Section 2.3 is applied for the same Gaussian data. In the simulation, the step size in each dimension of the filter coefficient is 0.1, i.e., $s = 0.1$. Other parameters ($\mathrm{d}m$, $\mathrm{d}\sigma$ and the ranges of the noise mean and standard deviation) are set to be the same with those in the previous subsection. The sensitivity measures of the proposed filter, the MA filter and the LW filter are listed in Tables 2.2 and 2.3, respectively. Also, the sensitivity over the trip point is treated equally to that over the KLD, i.e., $\mu_3 = \mu_4$. Three constraint levels are simulated, which are denoted with different sensitivity constraint levels: $\mu_3 = \mu_4 = 5\%$, $\mu_3 = \mu_4 = 10\%$ and $\mu_3 = \mu_4 = 15\%$. The corresponding upper-bounds of the sensitivities in each constraint (2.35) and (2.36) are listed in the first row in Tables 2.2 and 2.3, denoted with $U_3$ and $U_4$ and calculated with (2.37) and (2.38), respectively. Besides, taking the case when $\mu_3 = \mu_4 = 15\%$ as an example, the sensitivities of the proposed filter, the MA and the LW filters over the trip point are shown in Fig. 2.8; and the sensitivity results over the KLD using the three filters are plotted in Fig. 2.9.

Firstly, from Table 2.2, the trade-off between detection errors and sensitivity values can be observed. On one hand, the MA filter outperforms




**(a)**




**(b)**

**Figure 2.7:** Sensitivity over the KLD for MA and LW filters with different orders: (a) MA; (b) LW.

the other two linear filters in terms of the detection error, but its sensitivity is larger than others. On the other hand, when the constraint level changes from $\mu_3 = \mu_4 = 5\%$ to $\mu_3 = \mu_4 = 15\%$, the summation of detection errors decreases, while the sensitivity of the proposed filter becomes larger. Also, it has been confirmed by simulations that when the constraint level is $\mu_3 = \mu_4 = 20\%$, i.e., $U_3 = 16.46$ and $U_4 = 2.50$, the MA filter becomes the sensitivity-based optimal filter.

Secondly, how the sensitivity constraints work in the filter design is studied

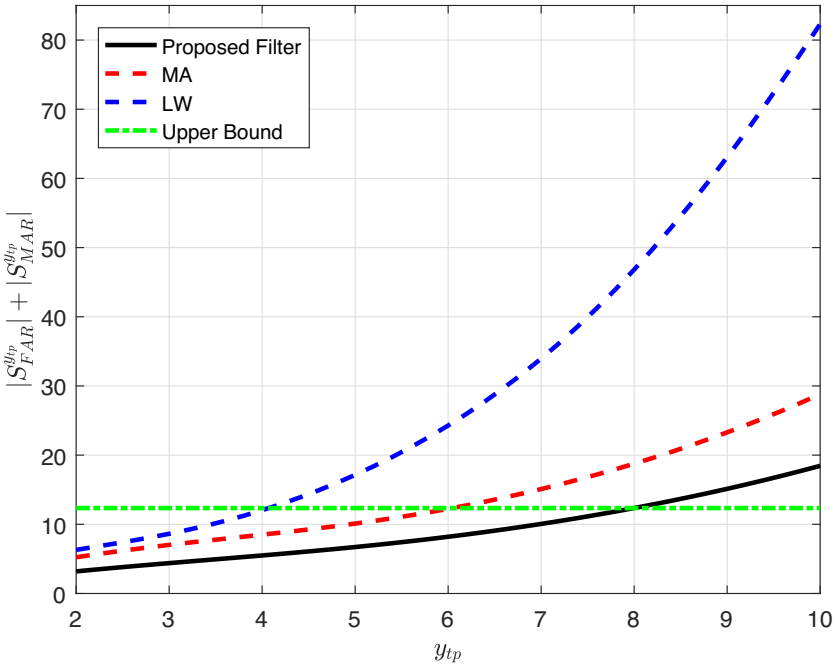



**Figure 2.8:** Three filters comparison in terms of the average sensitivity over the trip point with Gaussian data.

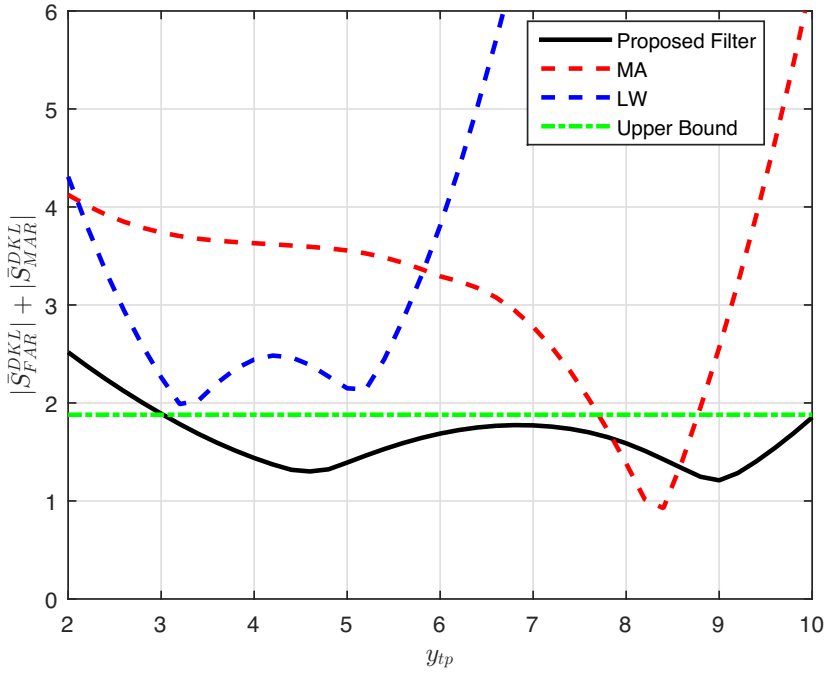



**Figure 2.9:** Three filters comparison in terms of the average sensitivity over the KLD with Gaussian data.

**Table 2.2:** Sensitivity values with different constraint levels with Gaussian data.

| Constraint Levels $[\mu_3, \mu_4]$ | [5%, 5%] | [10%, 10%] | [15%, 15%] |
|---|---|---|---|
| Constraints $[U_3, U_4]$ | $[4.43, 0.69]$ | $[8.27, 1.28]$ | $[12.34, 1.88]$ |
| Proposed $\boldsymbol{\theta}$ | $[-0.7, 0.1, 0.4, 0.8, 0.4]$ | $[-0.1, 0.3, 0.4, 0.5, -0.1]$ | $[-0.1, 0.2, 0.3, 0.4, 0.2]$ |
| $y_{tp}$ | 8.78 | 8.21 | 7.97 |
| $FAR + MAR$ | 0.67 | 0.42 | 0.32 |
| $\lvert S_{FAR}^{y_{tp}}\rvert + \lvert S_{MAR}^{y_{tp}}\rvert$ | 4.36 | 8.21 | 12.23 |
| $\lvert \bar{S}_{FAR}^{KLD}\rvert + \lvert \bar{S}_{MAR}^{KLD}\rvert$ | 0.64 | 1.13 | 1.56 |

**Table 2.3:** Sensitivity values with MA and LW filters with Gaussian data.

| | MA | LW |
|---|---|---|
| $\boldsymbol{\theta}$ | $[0.20, 0.20, 0.20, 0.20, 0.20]$ | $[0.04, 0.08, 0.12, 0.16, 0.20]$ |
| $y_{tp}$ | 7.89 | 4.76 |
| $FAR + MAR$ | 0.19 | 0.24 |
| $\lvert S_{FAR}^{y_{tp}}\rvert + \lvert S_{MAR}^{y_{tp}}\rvert$ | 17.96 | 15.97 |
| $\lvert \bar{S}_{FAR}^{KLD}\rvert + \lvert \bar{S}_{MAR}^{KLD}\rvert$ | 1.73 | 2.27 |

when the sensitivity level is $\mu_3 = \mu_4 = 15\%$, i.e., $U_3 = 12.34, U_4 = 1.88$. The constraint on the trip point is considered first in Fig. 2.8, where the summation of sensitivity magnitudes over the FAR and MAR is drawn with different trip points. The constraint is labeled with dotted green line for reference. It can be seen that because of the constraint, the feasible trip point ranges of the proposed filter, the LW filter and the MA filter are truncated to be $[2, 4.02]$, $[2, 5.93]$ and $[2, 7.91]$ respectively. Similarly, the second sensitivity constraint about the sensitivity over the KLD in (2.34) is shown in Fig. 2.9. It can be seen that the proposed filter and the MA filter can satisfy the constraint with the feasible ranges for the trip point denoted as $[3.02, 9.98]$ and $[7.69, 8.81]$, respectively; while the LW filter can not meet the second sensitivity constraint, thus it is infeasible. Then, combining the two constraints, there is no intersections for neither of the LW and MA filters. Finally, it can be concluded that the proposed filter has a better balance on the detection error and sensitivity than the other two linear filters.

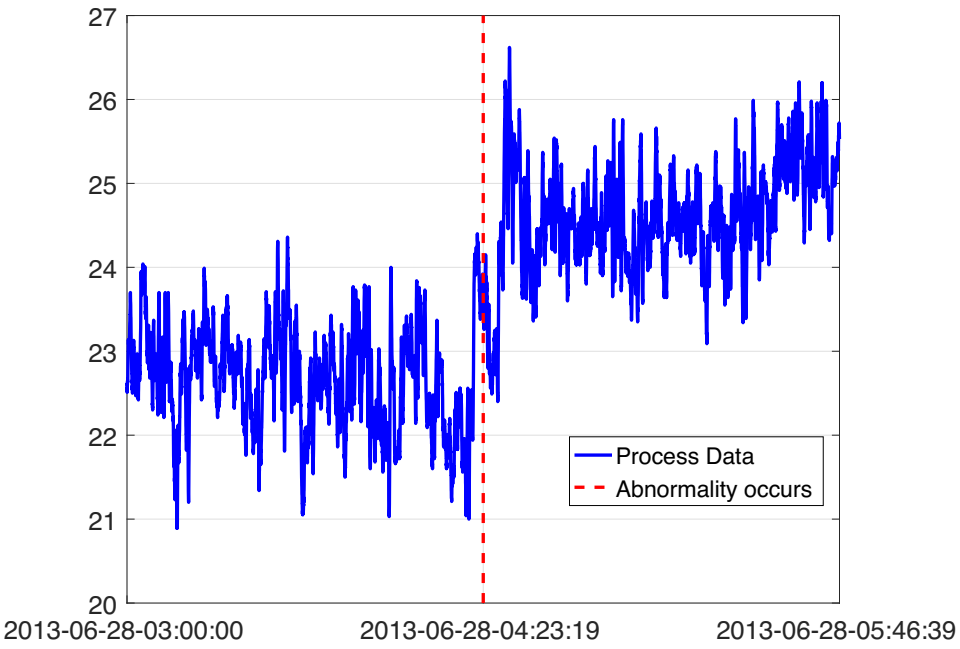



**Figure 2.10:** Industrial data.

### 2.4.3 Sensitivity-Based Filter Design with Industrial Data

In this section, industrial data collected from a Chinese power plant in 2013 is used for the sensitivity analysis and design. The data plot is shown in Fig. 2.10, where the data points are sampled every 0.1 seconds and the data sizes are $l_n = l_{ab} = 1000$. Same as the Gaussian case, in the simulation, the step size of the trip point is set to be 0.01, and the step size of the linear filter is 0.1. In the calculation of the sensitivity over the KLD, the range of the noise mean under the normal and abnormal modes are $m_n \in [-0.45, 0.45]$ and $m_{ab} \in [-0.49, 0.49]$ with the step size 0.01; and the ranges of the noise standard deviation are $\sigma_n \in [0.52, 0.78]$ and $\sigma_{ab} \in [0.53, 0.79]$ with a step size 0.01. By following the sensitivity-based filter design procedure in Section III, the sensitivity values with different constraint levels $\mu_3 = \mu_4 = 5\%$, $\mu_3 = \mu_4 = 10\%$ and $\mu_3 = \mu_4 = 20\%$ are simulated. The corresponding upper bounds are listed in the first row of Tables 2.4 and 2.5. By taking the third case as an example, the effects of the constraint with respect to the sensitivities over

the trip point on the proposed filter, the MA and the LW filters are shown in Fig. 2.11, and those with respect to the sensitivity over the KLD are shown in Fig.2.12.

**Table 2.4:** Sensitivity values with different constraints levels with the industrial data.

| Constraint Levels $[\mu_3, \mu_4]$ | [5%, 5%] | [10%, 10%] | [20%, 20%] |
|---|---|---|---|
| Constraint $[U_3, U_4]$ | $[25.60, 5.33]$ | $[34.70, 10.25]$ | $[52.91, 22.34]$ |
| Proposed $\boldsymbol{\theta}$ | $[-1, 0.3, 1.7]$ | $[0.4, 0.5, 0.7$ | $[0.1, 0.2, 0.4,$ |
| | $1.0, -1.6]$ | $0.5, -0.1]$ | $1, -0.7]$ |
| $y_{tp}$ | 23.59 | 23.60 | 23.61 |
| $FAR + MAR$ | 0.30 | 0.26 | 0.23 |
| $\lvert S_{FAR}^{y_{tp}} \rvert + \lvert S_{MAR}^{y_{tp}} \rvert$ | 23.81 | 33.46 | 51.14 |
| $\lvert \bar{S}_{FAR}^{KLD} \rvert + \lvert \bar{S}_{MAR}^{KLD} \rvert$ | 4.79 | 8.68 | 15.61 |

**Table 2.5:** Sensitivity values with MA and LW filters with the industrial data.

| | MA | LW |
|---|---|---|
| $\boldsymbol{\theta}$ | $[0.20, 0.20, 0.20, 0.20, 0.20]$ | $[0.04, 0.08, 0.120.16, 0.20]$ |
| $y_{tp}$ | 23.61 | 14.23 |
| $FAR + MAR$ | 0.18 | 0.41 |
| $\lvert S_{FAR}^{y_{tp}} \rvert + \lvert S_{MAR}^{y_{tp}} \rvert$ | 60.72 | 32.86 |
| $\lvert \bar{S}_{FAR}^{KLD} \rvert + \lvert \bar{S}_{MAR}^{KLD} \rvert$ | 28.67 | 2.49 |

Similar to the observations in the Gaussian case, the trade-offs between the performance error and the sensitivity measure also exist with industrial data. Specifically, the sensitivity over the trip point and over the KLD both increase with a lower sensitivity constraint, while the weighted sum of the detection errors experiences a downward trend. Still, the MA filter is more sensitive to both the trip point and the KLD changes than the proposed filter, even though the MA filter has lower detection errors.

Also, from Fig. 2.11, it can be noticed that neither the MA filter nor the LW filter can meet the sensitivity requirement with respect to the trip point. Thus, the proposed filter still outperforms the other two filters in terms of the sensitivity taking into account both detection errors and sensitivity.

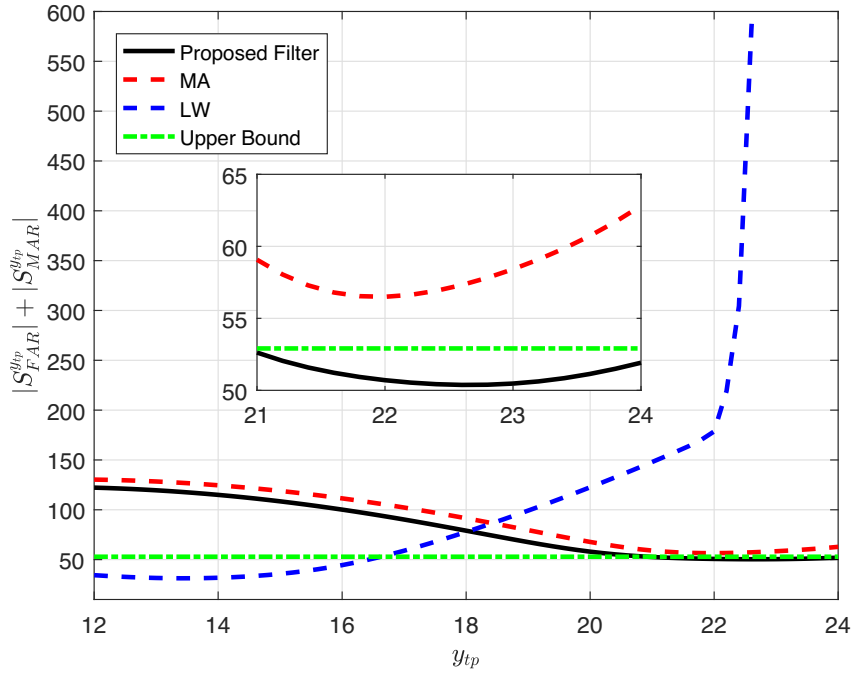



**Figure 2.11:** Three filters comparison in terms of the average sensitivity over the trip point with industrial data.

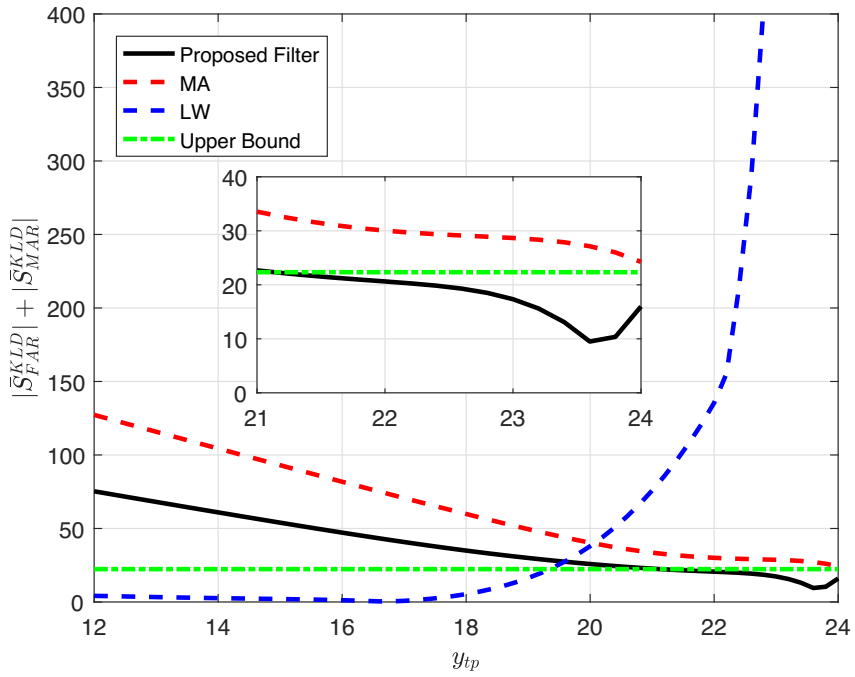



**Figure 2.12:** Three filters comparison in terms of the average sensitivity over the KLD with industrial data.

## 2.5 Summary

In this chapter, both sensitivity analysis and sensitivity-based design are studied for linear alarm filters. To quantify the sensitivity, a local sensitivity measure, namely, elasticity, is adopted and customized. In the sensitivity model, uncertainties in the trip point and in the KLD between the PDF of the noise-free data and that of the noisy data are modeled as the inputs; while the FAR and MAR are modeled as the outputs. By relating the definition of the elasticity, sensitivity measures are defined to quantify the performance change caused by variations in the inputs. With the help of the Gaussian Kernel based method, analytical results of sensitivity values are derived, which do not require knowledge on data distributions. Based on these sensitivity measures, a linear filter design problem is formulated in a constrained minimization structure, where sensitivity requirements are incorporated as constraints and the summation of the FAR and MAR is the objective function. The grid search is used to find the optimal sensitivity-based linear filter.

Simulation results with the proposed filter and the MA and LW filters are conducted for the Gaussian and industrial data respectively. By comparing proposed results derived without any distribution information and the theoretical ones obtained with the known PDFs, the proposed performance sensitivity analysis is verified to be effective. Furthermore, trade-offs between sensitivities and performance errors are observed. The proposed linear filter is shown to outperform the MA and LW filters when the sensitivity is considered.

# Chapter 3

# Abnormality Detection Based on the Kullback-Leibler Divergence for Generalized Gaussian Data*

---

WHILE the basic filters deal with changes in simple statistics such as the mean, median and variance, the divergence based methods focus on statistical changes in distributions. In this chapter, the KLD is adopted as the detection statistic for i.i.d. data from the non-centralized GGDs. The focus is on the analytical calculation of the KLD and its distribution under the normal condition, based on which the FAR can be estimated and the threshold can be tuned.

---

## 3.1 Overview

In this chapter, we investigate the KLD based abnormality detection for the non-centralized GGD, where the data samples are assumed to be i.i.d. under the normal condition. The non-centralized GGD is studied instead of the centralized one, since it is more general and is especially helpful for incipient fault detection in the abnormality detection systems. Taking into account the

*A version of this chapter has been published as: Ying Xiong, Yindi Jing, Tongwen Chen, Abnormality detection based on the Kullback-Leibler divergence for generalized Gaussian data, *Control Engineering Practice*, 85: 257-270, 2019.

mean change, the non-centralized GGD can magnify the KLD value under the abnormal condition; while the centralized GGD cannot relate the mean change with the KLD. Thus, for incipient faults with a constant bias caused by sensor failures [144], the non-centralized GGD can capture this fault signature; for faults with both mean and variance change, the non-centralized GGD can further amplify the KLD value than the centralized one. However, there was no result on the KLD calculation and its distribution with the non-centralized GGDs due to two main challenges. First, because of the mean change, there is no straightforward analytical formula for the KLD calculation. Further, the complicated KLD expression makes the PDF derivation even more difficult.

In particular, non-centralized GGDs with the shape parameter (denoted as $\beta$) greater than 1 are considered in this chapter. The reasons of focusing on this $\beta$ range are three-fold. First, the mathematical analysis is more tractable for this case. Second, existing parameter estimations for the case of $\beta \leq 1$ are inaccurate and unstable even for a large number of data samples [124]. Lastly, while KLD calculations for the case of $\beta \leq 1$ can still be conducted with more involved manipulations, the result takes a fundamentally different format from the case of $\beta > 1$. Thus, we consider the case of $\beta > 1$ only and leave the other case for future work.

The remainder of this chapter is organized as follows. Section 3.2 contains the detection problem formulation, the KLD expression, derivations of the joint PDF and conditional PDFs of the KLD, and threshold selection schemes. The algorithms for the constant and adaptive threshold methods are summarized and discussed in Section 3.3. Then, Section 3.4 shows the case study results. Section 3.5 concludes this work.

## 3.2 Abnormality Detection Problem

A series of observed data of size $N$ is used for detection and is denoted by $\{x_i\}_{i=1}^{N}$ or the $N$-dimensional vector $\boldsymbol{x}$. It is assumed that under the normal condition, $x_i$'s are i.i.d. samples from a generalized Gaussian PDF denoted

as GGD$(m, \sigma^2, \beta)$, where $m, \sigma^2, \beta$ are known parameters. The distributions of data samples under abnormal situations are different to GGD$(m, \sigma^2, \beta)$. Thus, the binary hypothesis test can be described as to decide whether the observed data samples in $\boldsymbol{x}$ follow the known PDF, i.e.,

$$\begin{cases} H_0 : x_i\text{'s follow distribution GGD}(m, \sigma^2, \beta), \\ H_1 : x_i\text{'s do not follow distribution GGD}(m, \sigma^2, \beta). \end{cases} \tag{3.1}$$

### 3.2.1 Background on Generalized Gaussian Distribution and Its Parameter Estimation

The analytical form of the generalized Gaussian PDF can be expressed as the following [39]:

$$\text{GGD}(m, \sigma^2, \beta) = \frac{\beta}{2\sigma c(\beta)\Gamma(1/\beta)} e^{-\left(\frac{|x-m|}{\sigma c(\beta)}\right)^{\beta}}, \tag{3.2}$$

where $\Gamma(\cdot)$ is the Gamma function defined as $\Gamma(z) \triangleq \int_0^{+\infty} e^{-t} t^{z-1}\, \mathrm{d}t, z > 0$ and $c(\beta) \triangleq [\Gamma(1/\beta)/\Gamma(3/\beta)]^{1/2}$. The three parameters $m, \sigma^2$ and $\beta$ are the mean, the variance, and the shape parameter of the GGD, respectively. The mean determines the center of a distribution, and the variance represents the width of the PDF peak [39]. The shape parameter $\beta$ inversely quantifies the decreasing rate of the peak. When $\beta = 1$ or $\beta = 2$, the GGD becomes Laplacian or Gaussian distribution, respectively. The PDFs of the GGD with respect to different shape parameters when $m = 0$ and $\sigma = 1$ are shown in Figure 3.1. It can be seen that the top gets flatter and the tail is lighter with greater $\beta$. When one random variable $x$ follows a GGD, it is denoted as $x \sim \text{GGD}(x|m, \sigma^2, \beta)$ for the rest of this chapter to help the presentation.

Given a set of i.i.d. data samples $\{x_i\}_{i=1}^N$ with $x_i \sim \text{GGD}(x|m, \sigma^2, \beta)$, the three parameters $m, \sigma^2, \beta$ can be estimated from $x_i$'s. Several schemes have been proposed in existing literature including the maximum likelihood estimation, moment-based estimation, entropy matching, and globally convergent method [124]. In this chapter, moment-based estimation [7] is used to estimate $m$ and $\sigma^2$ in consideration of its low computational complexity and tractability in the KLD distribution analysis.

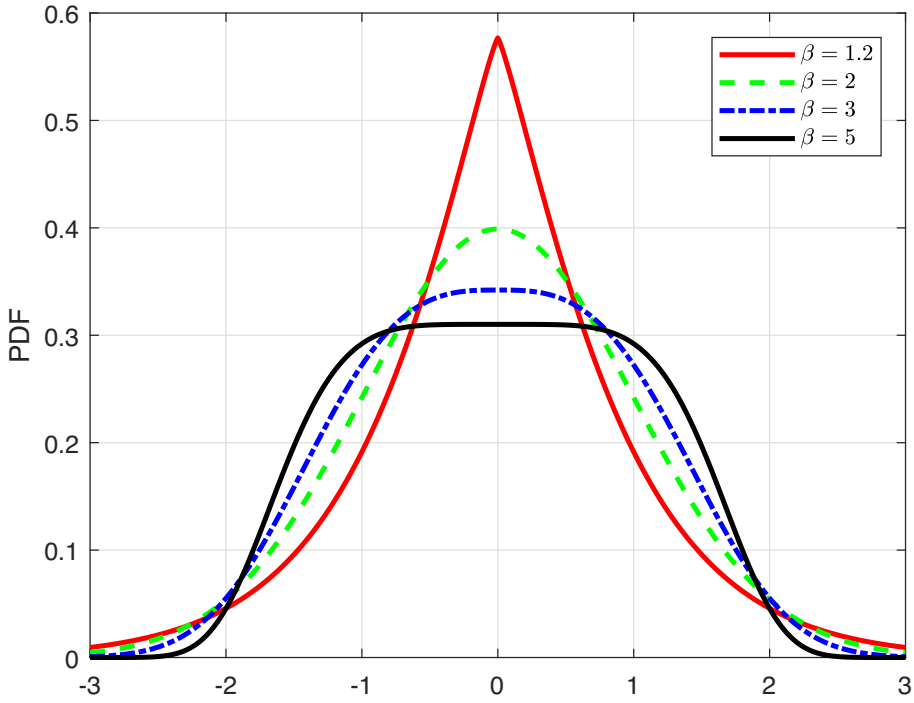



**Figure 3.1:** PDFs of the GGD model with different shape parameters.

The mean and variance can be estimated as

$$\hat{m} = \frac{1}{N}\sum_{i=1}^{N} x_i, \tag{3.3}$$

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{m})^2 = \frac{1}{N}\sum_{i=1}^{N} x_i^2 - \left(\frac{1}{N}\sum_{i=1}^{N} x_i\right)^2. \tag{3.4}$$

The estimation of $\beta$ is more difficult and the maximum likelihood approach [39] is used. The likelihood function of the sample set can be derived as

$$L(\boldsymbol{x}|m, \sigma^2, \beta) = \ln \prod_{i=1}^{N} \mathrm{GGD}(x_i|m, \sigma^2, \beta), \tag{3.5}$$

where $\ln(\cdot)$ is the natural-logarithm function. The estimation of the shape

parameter $\hat{\beta}$ satisfies the following equation:

$$\left.\frac{\partial L(\boldsymbol{x}|m,\sigma^2,\beta)}{\partial\beta}\right|_{\beta=\hat{\beta}} = \frac{\psi(1/\hat{\beta})}{\hat{\beta}} - \frac{\sum_{i=1}^{N}|x_i-m|^{\hat{\beta}}\ln|x_i-m|}{\sum_{i=1}^{N}|x_i-m|^{\hat{\beta}}} + \frac{\ln\left(\frac{\hat{\beta}}{N}\sum_{i=1}^{N}|x_i-m|^{\hat{\beta}}\right)}{\hat{\beta}} + 1 = 0, \tag{3.6}$$

where $\Psi(\cdot)$ is the digamma function [39]. For given values of $m$ and $\sigma^2$, it has been proved that (3.6) has a unique real root [124]. The equation can be solved numerically using iteration methods such as the Newton-Raphson method, where the estimates $\hat{m}$ and $\hat{\sigma}^2$ in (3.3) and (3.4) are applied.

### 3.2.2 KLD-Based Detection Scheme

As the hypothesis problem in (3.1) is to test whether the data set follows i.i.d. GGD$(m,\sigma^2,\beta)$, a natural way is to fit the data as a GGD denoted with GGD$(\hat{m},\hat{\sigma}^2,\hat{\beta})$ and check if GGD$(\hat{m},\hat{\sigma}^2,\hat{\beta})$ is close enough to the known distribution GGD$(m,\sigma^2,\beta)$. To quantitatively measure the distance from one PDF to another, the KLD is a natural choice. The unsymmetrical KLD from GGD$(m,\sigma^2,\beta)$ to GGD$(\hat{m},\hat{\sigma}^2,\hat{\beta})$ is defined as [82]

$$\begin{aligned} &D_{kl}\left(\text{GGD}(\hat{m},\hat{\sigma}^2,\hat{\beta})||\text{GGD}(m,\sigma^2,\beta)\right) \\ &\triangleq \int_{-\infty}^{+\infty}\text{GGD}(\hat{m},\hat{\sigma}^2,\hat{\beta})\ln\frac{\text{GGD}(\hat{m},\hat{\sigma}^2,\hat{\beta})}{\text{GGD}(m,\sigma^2,\beta)}\,\mathrm{d}x. \end{aligned} \tag{3.7}$$

When there is no confusion, we use the short notation $D_{kl}$ to help the presentation of $D_{kl}\left(\text{GGD}(\hat{m},\hat{\sigma}^2,\hat{\beta})||\text{GGD}(m,\sigma^2,\beta)\right)$. The KLD-based detection rule for the hypothesis test in (3.1) is as follows:

$$\begin{cases} \text{Decide on } H_0 \text{ when } D_{kl} \le D_{th}, \\ \text{Decide on } H_1 \text{ when } D_{kl} > D_{th}, \end{cases} \tag{3.8}$$

where $D_{th}$ is the threshold for detection.

The KLD is used as the test statistic in (3.8). Theorem 3.2.1 gives an analytical expression for the KLD value, which is required for the KLD distribution analysis in the next subsection.

**Theorem 3.2.1.** The KLD from the generalized Gaussian PDF with parameters $(\hat{m}, \hat{\sigma}^2, \hat{\beta})$ to the one with parameters $(m, \sigma^2, \beta)$ has the following asymptotic behaviour when $|\hat{m} - m| \ll 1$,

$$\begin{aligned} D_{kl} &= \ln \frac{\hat{\beta}\sigma\Gamma(1/\beta)\sqrt{\Gamma(1/\beta)\Gamma(3/\hat{\beta})}}{\beta\hat{\sigma}\Gamma(1/\hat{\beta})\sqrt{\Gamma(1/\hat{\beta})\Gamma(3/\beta)}} - \frac{1}{\hat{\beta}} \\ &+ \frac{\beta(\beta-1)}{2\sigma^2}\left(\frac{\hat{\sigma}\sqrt{\Gamma(1/\hat{\beta})\Gamma(3/\beta)}}{\sigma\sqrt{\Gamma(1/\beta)\Gamma(3/\hat{\beta})}}\right)^{\beta-2} \frac{\Gamma(3/\beta)\Gamma(\beta/\hat{\beta}-1/\hat{\beta})}{\Gamma(1/\hat{\beta})\Gamma(1/\beta)}(\hat{m}-m)^2 \\ &+ \left(\frac{\hat{\sigma}\sqrt{\Gamma(1/\hat{\beta})\Gamma(3/\beta)}}{\sigma\sqrt{\Gamma(1/\beta)\Gamma(3/\hat{\beta})}}\right)^{\beta} \frac{\Gamma(\beta/\hat{\beta}+1/\hat{\beta})}{\Gamma(1/\hat{\beta})} + o[(\hat{m}-m)^2], \end{aligned} \tag{3.9}$$

where $f(x) = o(g(x))$ means $\lim_{x\to 0} f(x)/g(x) = 0$.

*Proof.* By substituting the GGD expression in Eq. (3.2) into the KLD definition in Eq. (3.7), we have

$$\begin{aligned} &D_{kl}\left(\mathrm{GGD}(\hat{m}, \hat{\sigma}^2, \hat{\beta}) || GGD(m, \sigma, \beta)\right) \\ &= \int_{-\infty}^{+\infty} \frac{\hat{\beta}}{2\hat{\sigma}c(\hat{\beta})\Gamma(1/\hat{\beta})} e^{-\left(\frac{|x|}{\hat{\sigma}c(\hat{\beta})}\right)^{\hat{\beta}}} \left[\ln \frac{\hat{\beta}\sigma c(\beta)\Gamma(1/\beta)}{\hat{\beta}\hat{\sigma}c(\hat{\beta})\Gamma(1/\hat{\beta})} - \left(\frac{|x|}{\hat{\sigma}c(\hat{\beta})}\right)^{\hat{\beta}} \right. \\ &\qquad \left. + \left(\frac{|x+\hat{m}-m|}{\sigma c(\beta)}\right)^{\beta}\right] \mathrm{d}x \\ &= \ln \frac{\hat{\beta}\sigma c(\beta)\Gamma(1/\beta)}{\beta\hat{\sigma}c(\hat{\beta})\Gamma(1/\hat{\beta})} - \frac{1}{\hat{\beta}} + T, \end{aligned} \tag{3.10}$$

where we have defined $T$ to represent the last term of the KLD as follows:

$$T \triangleq \int_{-\infty}^{+\infty} \frac{\hat{\beta}}{2\hat{\sigma}c(\hat{\beta})\Gamma(1/\hat{\beta})} e^{-\left(\frac{|x|}{\hat{\sigma}c(\hat{\beta})}\right)^{\hat{\beta}}} \left(\frac{|x+\hat{m}-m|}{\sigma c(\beta)}\right)^{\beta} \mathrm{d}x. \tag{3.11}$$

By denoting $a \triangleq \hat{m} - m$, we have

$$|x + \hat{m} - m|^{\beta} = |x + a|^{\beta}, \tag{3.12}$$

which can be further expanded with the help of Taylor series under the con-

vergence range of $|x| \geq |a|$ as follows:

$$
\begin{aligned}
|x+\hat{m}-m|^{\beta} &= \left|1+\frac{a}{x}\right|^{\beta}|x|^{\beta} \\
&= |x|^{\beta}\left[1+\beta\frac{a}{x}+\frac{1}{2}\beta(\beta-1)\left(\frac{a}{x}\right)^{2}+o\left(\frac{a^2}{x^2}\right)\right]. \qquad (3.13)
\end{aligned}
$$

Thus, the integration in Eq. (3.11) can be calculated as

$$
\begin{aligned}
T \overset{(a)}{=}\ & 2\int_{|a|}^{\infty}\frac{\hat{\beta}}{2\hat{\sigma}c(\hat{\beta})\Gamma(1/\hat{\beta})}e^{-\left(\frac{|x|}{\hat{\sigma}c(\hat{\beta})}\right)^{\hat{\beta}}}\left(\frac{|x|}{\sigma c(\beta)}\right)^{\beta}\left[1+\frac{1}{2}\beta(\beta-1)\frac{a^2}{x^2}\right. \\
&\left.+\sum_{n=2}^{\infty}\left[\binom{\beta}{2n}\frac{a^{2n}}{x^{2n}}\right]\right]\mathrm{d}x \\
&+\int_{-|a|}^{|a|}\frac{\hat{\beta}}{2\hat{\sigma}c(\hat{\beta})\Gamma(1/\hat{\beta})}e^{-\left(\frac{|x|}{\hat{\sigma}c(\hat{\beta})}\right)^{\hat{\beta}}}\left(\frac{1}{\sigma c(\beta)}\right)^{\beta}|x+a|^{\beta}\,\mathrm{d}x \\
\overset{(b)}{=}\ & \underbrace{\int_{0}^{\infty}\frac{\hat{\beta}}{\hat{\sigma}c(\hat{\beta})\Gamma(1/\hat{\beta})}e^{-\left(\frac{|x|}{\hat{\sigma}c(\hat{\beta})}\right)^{\hat{\beta}}}\left(\frac{|x|}{\sigma c(\beta)}\right)^{\beta}dx}_{T_1} \\
&+a^2\underbrace{\int_{0}^{\infty}\frac{\hat{\beta}\beta(\beta-1)}{2\hat{\sigma}c(\hat{\beta})\Gamma(1/\hat{\beta})}e^{-\left(\frac{|x|}{\hat{\sigma}c(\hat{\beta})}\right)^{\hat{\beta}}}\frac{|x|^{\beta-2}}{[\sigma c(\beta)]^{\beta}}dx}_{T_2} \\
&-\underbrace{\int_{0}^{|a|}\frac{\hat{\beta}}{\hat{\sigma}c(\hat{\beta})\Gamma(1/\hat{\beta})}e^{-\left(\frac{|x|}{\hat{\sigma}c(\hat{\beta})}\right)^{\hat{\beta}}}\left(\frac{|x|}{\sigma c(\beta)}\right)^{\beta}dx}_{T_3} \\
&-a^2\underbrace{\int_{0}^{|a|}\frac{\hat{\beta}\beta(\beta-1)}{2\hat{\sigma}c(\hat{\beta})\Gamma(1/\hat{\beta})}e^{-\left(\frac{|x|}{\hat{\sigma}c(\hat{\beta})}\right)^{\hat{\beta}}}\frac{|x|^{\beta-2}}{[\sigma c(\beta)]^{\beta}}dx}_{T_4} \\
&+\sum_{n=2}^{\infty}\binom{\beta}{2n}\underbrace{\int_{|a|}^{\infty}\frac{\hat{\beta}}{\hat{\sigma}c(\hat{\beta})\Gamma(1/\hat{\beta})}e^{-\left(\frac{|x|}{\hat{\sigma}c(\hat{\beta})}\right)^{\hat{\beta}}}\frac{a^{2n}}{x^{2n}}\left[\frac{x}{\sigma c(\beta)}\right]^{\beta}\mathrm{d}x}_{T_{5n}} \\
&+\underbrace{\int_{-|a|}^{|a|}\frac{\hat{\beta}}{2\hat{\sigma}c(\hat{\beta})\Gamma(1/\hat{\beta})}e^{-\left(\frac{|x|}{\hat{\sigma}c(\hat{\beta})}\right)^{\hat{\beta}}}\left(\frac{1}{\sigma c(\beta)}\right)^{\beta}|x+a|^{\beta}\,\mathrm{d}x}_{T_6}, \qquad (3.14)
\end{aligned}
$$

where $(a)$ is obtained by the Taylor expansion in (3.13) and noticing that the integrands for the $x^{2n+1}$ terms are odd functions and the integrands of the $x^{2n}$

terms are even functions; $(b)$ is obtained by using $\int_{|a|}^{\infty} f(x)dx = \int_0^{\infty} f(x)dx - \int_0^{|a|} f(x)dx$.

Via the following integration formula in [55]

$$\int_0^{\infty} \frac{e^{-\beta x^n}}{x^m}\,\mathrm{d}x = \frac{\Gamma\left(\frac{1-m}{n}\right)}{n\beta^{\frac{1-m}{n}}}, \quad [Re\ \beta > 0, \mathrm{Re}\ n > 0, \mathrm{Re}\ m > -1],$$

it is straightforward to obtain the following:

$$T_1 = \left(\frac{\hat{\sigma}c(\hat{\beta})}{\sigma c(\beta)}\right)^{\beta} \frac{\Gamma(1/\hat{\beta} + \beta/\hat{\beta})}{\Gamma(1/\hat{\beta})};$$
$$T_2 = \frac{\beta(\beta-1)}{2[\sigma c(\beta)]^2}\left(\frac{\hat{\sigma}c(\hat{\beta})}{\sigma c(\beta)}\right)^{\beta-2} \frac{\Gamma(\beta/\hat{\beta} - 1/\hat{\beta})}{\Gamma(1/\hat{\beta})}.$$

For the rest of the terms of $T$, the following lemma can be proved.

**Lemma 3.2.2.** When $\beta > 1$ and $a \ll 1$, we have $T_3, T_4, T_{5n}, T_6 = o(a^2)$.

*Proof.* For $T_3$, since $e^{-\left(\frac{|x|}{\hat{\sigma}c(\hat{\beta})}\right)^{\hat{\beta}}} \leq 1$ and $\int_0^{|a|} |x|^{\beta}\,\mathrm{d}x \leq \frac{|a|^{\beta+1}}{\beta+1}$ when $\beta > 1$,

$$T_3 \leq \frac{\hat{\beta}}{(\beta+1)\hat{\sigma}c(\hat{\beta})\Gamma(1/\hat{\beta})} \frac{|a|^{\beta+1}}{[\sigma c(\beta)]^{\beta}} = o(a^2). \tag{3.15}$$

Similarly, we have

$$T_4 \leq a^2 \frac{\hat{\beta}\beta(\beta-1)}{2(\beta-1)\hat{\sigma}c(\hat{\beta})\Gamma(1/\hat{\beta})} \frac{|a|^{\beta-1}}{[\sigma c(\beta)]^{\beta}} = o(a^2).$$

For $T_{5n}$, via integration by parts and by using $\int_{|a|}^{\infty} f(x)\,\mathrm{d}x = \int_0^{\infty} f(x)\,\mathrm{d}x - \int_0^{|a|} f(x)\,\mathrm{d}x$, we have

$$\begin{aligned}
T_{5n} = & \\
& - \frac{\hat{\beta}}{\hat{\sigma}c(\hat{\beta})\Gamma(1/\hat{\beta})[\sigma c(\beta)]^{\beta}} e^{-\left(\frac{|a|}{\hat{\sigma}c(\hat{\beta})}\right)^{\hat{\beta}}} \sum_{i=0}^{B_n-1} \prod_{j=0}^{i} \frac{|a|^{\beta+i\hat{\beta}+1}\hat{\beta}^j}{(\beta + j\hat{\beta} - 2n + 1)[\hat{\sigma}c(\hat{\beta})]^{j\hat{\beta}}} \\
& + \frac{\Gamma(\beta/\hat{\beta} + B_n - 2n/\hat{\beta} + 1/\hat{\beta})}{[\hat{\sigma}c(\hat{\beta})]^{2n}\Gamma(1/\hat{\beta})} \left[\frac{\hat{\sigma}c(\hat{\beta})}{\sigma c(\beta)}\right]^{\beta} \prod_{i=0}^{B_n-1} \frac{\hat{\beta}^i |a|^{2n}}{(\hat{\beta} + i\hat{\beta} - 2n + 1)} \\
& - \frac{\hat{\beta}\int_0^{|a|} |x|^{\beta+B_n\hat{\beta}-2n} e^{-\left(\frac{x}{\hat{\sigma}c(\hat{\beta})}\right)^{\hat{\beta}}}\,\mathrm{d}x}{\hat{\sigma}c(\hat{\beta})\Gamma(1/\hat{\beta})[\sigma c(\beta)]^{\beta}} \prod_{i=0}^{B_n-1} \frac{\hat{\beta}^i |a|^{2n}}{(\beta + i\hat{\beta} - 2n + 1)[\hat{\sigma}c(\hat{\beta})]^{i\hat{\beta}}},
\end{aligned} \tag{3.16}$$

where $B_n$ is the smallest integer number that makes $\beta + B_n\hat{\beta} > 2n-1$, i.e., $B_n = \arg\min_{B_n \in \mathbb{N}} \beta + B_n\hat{\beta} > 2n-1$. In other words, $B_n = \lfloor (2n-1-\beta)/\hat{\beta} \rfloor$. We denote the three terms of $T_{5n}$ as $T_{5n.1}$, $T_{5n.2}$ and $T_{5n.3}$ respectively. It is also obvious to see that $T_{5n.1}, T_{5n.2} = o(a^2)$. Considering $e^{-\left(\frac{|x|}{\hat{\sigma}c(\hat{\beta})}\right)^{\hat{\beta}}} \leq 1$, we have that $T_{5n.3}$ is bounded as

$$\begin{aligned} T_{5n.3} &\leq \frac{\hat{\beta}}{\hat{\sigma}c(\hat{\beta})\Gamma(1/\hat{\beta})[\sigma c(\beta)]^{\beta}} \prod_{i=0}^{B_n-1} \frac{\hat{\beta}^i |a|^{\beta+B_n\hat{\beta}+1}}{(\beta+i\hat{\beta}-2n+1)[\hat{\sigma}c(\hat{\beta})]^{i\hat{\beta}}} \\ &= o(a^2). \end{aligned} \tag{3.17}$$

For $T_6$, since $e^{-\left(\frac{|x|}{\hat{\sigma}c(\hat{\beta})}\right)^{\hat{\beta}}} \leq 1$ and $|x+a|^{\beta} \leq 2^{\beta}|a|^{\beta}$,

$$T_6 \leq \frac{\hat{\beta}}{(\beta+1)\hat{\sigma}c(\hat{\beta})\Gamma(1/\hat{\beta})} \left(\frac{1}{\sigma c(\beta)}\right)^{\beta} |2a|^{\beta+1} = o(a^2). \tag{3.18}$$

□

By using the results in Lemma 3.1.1 and Eq. (3.14), Theorem 3.2.1 is proved. □

The expression in (3.9) specifies the constant term and the quadratic term of the KLD with respect to the mean-bias $\hat{m} - m$. The linear term disappears since the corresponding coefficient is zero. When $N \to \infty$, via the law of large numbers, $\hat{m} \to m$ almost surely. Thus, when the window size is large enough, we have $|\hat{m} - m| \ll 1$. In this case, the term $o[(\hat{m}-m)^2]$ deminishes to zero, thus the KLD can be approximated by the sum of the constant term and the quadratic term in (3.9).

### 3.2.3 Threshold Selection

In selecting the threshold value $D_{th}$, the Neyman–Pearson approach is used to ensure that the false alarm probability is no larger than a predetermined value $\gamma$ while maximizing the probability of detection. As the probability of detection is a non-increasing function of $D_{th}$. To find the value of $D_{th}$ is equivalent to solving the following equation:

$$\mathbb{P}(D_{kl} \geq D_{th}) = \gamma. \tag{3.19}$$

In the previous subsection, we find an analytical expression for the KLD when the data size is large in (3.9). As $\hat{m}, \hat{\sigma}^2, \hat{\beta}$ are functions of the data set $\{x_i\}_{i=1}^N$ via (3.3), (3.4), and (3.6), and the data samples are i.i.d. and following GGD$(m, \sigma^2, \beta)$, the KLD value $D_{kl}$ is also a random variable. To solve (3.19), the distribution of $D_{kl}$ needs to be derived. This is a challenging task due to the complicated expression for the KLD in (3.9). In what follows, we propose two methods to find the threshold value.

**Constant Threshold**

The first method directly derives the PDF of the KLD by using (3.9), the estimation formulas in (3.3), (3.4), and (3.6), and the distribution of the data set $\{x_i\}_{i=1}^N$. However, the estimation of $\beta$ as represented in (3.6) is in an implicit form, not an explicit function of $\{x_i\}_{i=1}^N$, $\hat{m}$, and $\hat{\sigma}^2$. Thus, for tractable analysis, instead of using the estimation $\hat{\beta}$, the value of $\beta$ (which is known a-priori) is used. This is the same as assuming that there is no change in $\beta$ for the data in the moving window.

With this simplification, we need to calculate the KLD from GGD$(m, \sigma^2, \beta)$ to GGD$(\hat{m}, \hat{\sigma}^2, \beta)$, i.e., $D_{kl}(\text{GGD}(\hat{m}, \hat{\sigma}^2, \beta)||\text{GGD}(m, \sigma^2, \beta))$. For the simplicity of presentation, it is denoted as $D_{kl,\beta}$. From (9), we have

$$D_{kl,\beta} = \frac{\beta(\beta-1)}{2\sigma^2}\left(\frac{\hat{\sigma}}{\sigma}\right)^{\beta-2}\frac{\Gamma(3/\beta)\Gamma(1-1/\beta)}{\Gamma(1/\beta)^2}(\hat{m}-m)^2 - \ln\frac{\hat{\sigma}}{\sigma} + \frac{1}{\beta}\left(\frac{\hat{\sigma}}{\sigma}\right)^{\beta} - \frac{1}{\beta} + o[(\hat{m}-m)^2]. \tag{3.20}$$

Define

$$c_1 \triangleq \frac{\beta(\beta-1)\Gamma(1-1/\beta)\Gamma(3/\beta)}{\Gamma^2(1/\beta)}, \tag{3.21}$$

$$M \triangleq \sqrt{N}\frac{\hat{m}-m}{\sigma} = \frac{\sqrt{N}}{\sigma}\left(\frac{1}{N}\sum_{i=1}^{N}x_i - m\right), \tag{3.22}$$

$$S \triangleq \frac{\hat{\sigma}^2}{\sigma^2} = \frac{1}{N}\sum_{i=1}^{N}\frac{\left(x_i - \frac{1}{N}\sum_{j=1}^{N}x_j\right)^2}{\sigma^2}, \tag{3.23}$$

where $c_1$ is a deterministic coefficient while $M$ and $S$ are random variables as they are functions of $x_i$'s. By ignoring the last term in (3.20), we have the following approximation when $N$ is large ($N \gg 1$):

$$D_{kl,\beta} \approx D_{kl,\beta}^{asy} \triangleq -\frac{1}{2}\ln S + \frac{1}{\beta}S^{\frac{\beta}{2}} - \frac{1}{\beta} + \frac{c_1}{2N}M^2 S^{\frac{\beta}{2}-1}. \tag{3.24}$$

We prove the following lemma on the distributions of $M$ and $S$ for $N \to \infty$, then a closed-form approximation of the PDF of $D_{kl,\beta}$ is obtained for large $N$.

**Lemma 3.2.3.** Define

$$c_2 \triangleq \frac{\Gamma(1/\beta)\Gamma(5/\beta)}{\Gamma(3/\beta)\Gamma(3/\beta)} - 1. \tag{3.25}$$

If $x_i$'s are i.i.d. following $\mathrm{GGD}(x, \sigma^2, \beta)$, we have

$$M \xrightarrow[N\to\infty]{d} \mathcal{N}(0,1). \tag{3.26}$$

$$\sqrt{N}(S-1) \xrightarrow[N\to\infty]{d} \mathcal{N}(0,c_2). \tag{3.27}$$

$$\mathrm{Cov}\left[M, \sqrt{N}S\right] = 0, \tag{3.28}$$

where the notation $\xrightarrow[N\to\infty]{d}$ represents convergence in distribution when $N \to \infty$, $\mathcal{N}(a,b)$ denotes the Gaussian distribution with mean $a$ and variance $b$, and $\mathrm{Cov}(A,B)$ denotes the covariance of $A$ and $B$.

*Proof.* Since $x_i$'s are i.i.d. folllwing $\mathrm{GGD}(m,\sigma,\beta)$, when $N \to \infty$, from central limit theorem,

$$\sqrt{N}\left(\frac{1}{N}\sum_{i=1}^{N} x_i - m\right) \xrightarrow[N\to\infty]{d} \mathcal{N}(0,\sigma^2).$$

Thus

$$M = \frac{\sqrt{N}}{\sigma}\left(\frac{1}{N}\sum_{i=1}^{N} x_i - m\right) \xrightarrow[N\to\infty]{d} \mathcal{N}(0,1), \tag{3.29}$$

which is (4.9) in Lemma 3.2.3.

Define $y_i \triangleq (x_i - m)/\sigma$, from the definition in (3.23), we have

$$S = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - \frac{1}{N}\sum_{j=1}^{N} y_j\right)^2 = \frac{1}{N}\sum_{i=1}^{N} y_i^2 - \frac{M^2}{N}.$$

According to (3.29), it is easy to get that $M^2$ converges in distribution to a chi-square random variable with one degree of freedom, i.e., $M^2 \xrightarrow[N\to\infty]{d} \chi_1^2$. Thus, $M^2$ has bounded variance. By applying the Tchebychev's inequality to the term $M^2/N$, we have

$$\lim_{N\to\infty} \mathbb{P}\left(\frac{M^2}{N}\right) = 0. \tag{3.30}$$

So the asymptotic distribution of $S$ can be derived from the distribution of $\sum_{i=1}^N y_i^2/N$.

Via straightforward calculations, we have

$$\begin{aligned}
\mathbb{E}\left(y_i^2\right) =& \mathbb{E}\left[\frac{(x_i-m)^2}{\sigma^4}\right] = \frac{\sigma^2}{\sigma^2} = 1, \\
\mathrm{Var}\left(y_i^4\right) =& \mathbb{E}\left(y_i^4\right) - \left[\mathbb{E}\left(y_i^2\right)\right]^2 = \mathbb{E}\left[\frac{(x_i-m)^4}{\sigma^2}\right] - 1 \\
=& \frac{\Gamma(1/\beta)\Gamma(5/\beta)}{\Gamma(3/\beta)\Gamma(3/\beta)} - 1 = c_2,
\end{aligned} \tag{3.31}$$

Further, since $y_i^2$'s are i.i.d., it is easy to get $\sqrt{N}(S-1) \xrightarrow[N\to\infty]{d} \mathcal{N}(0, c_2)$ according to the central limit theorem. Thus, (4.10) is derived.

In terms of the covariance $\mathrm{Cov}(M, \sqrt{N}S)$, it can be calculated as

$$\mathrm{Cov}(M, \sqrt{N}S) = \mathbb{E}(\sqrt{N}MS) - \mathbb{E}(M)\,\mathbb{E}(\sqrt{N}S). \tag{3.32}$$

By using $y_i$'s defined above to simply the calculation, we have

$$\begin{aligned}
\mathrm{Cov}(M, \sqrt{N}S) =& \frac{1}{N}\,\mathbb{E}\left(\sum_{i=1}^N y_i \sum_{i=1}^N y_i^2\right) - \frac{1}{N^2}\,\mathbb{E}\left[\sum_{i=1}^N y_i \left(\sum_{i=1}^N y_i\right)^2\right] \\
=& -\frac{1}{N^2}\,\mathbb{E}\left[\sum_{i=1}^N y_i \left(\sum_{i=1}^N y_i^2 + 2\sum_{i=1}^N \sum_{j=i+1}^N y_i y_j\right)\right] \\
&+ \frac{1}{N}\,\mathbb{E}\left(\sum_{i=1}^N \sum_{j=1}^N y_i y_j^2\right), \\
=& \left(1 - \frac{1}{N}\right)\mathbb{E}(y_i^3).
\end{aligned} \tag{3.33}$$

Since $\mathbb{E}(y_i^3) = 0$, we have $\mathrm{Cov}(M, \sqrt{N}S) = 0$, i.e., $M$ and $\sqrt{N}S$ are uncorrelated. □

**Theorem 3.2.4.** Define $L \triangleq 2ND_{kl,\beta}^{asy}$. When $\beta > 1$ and $N \gg 1$, the PDF of $L$ can be approximated by $f_L(l)$ as defined as follows:

$$f_L(l) \triangleq \frac{1}{\sqrt{\beta c_1 c_2}} \exp\left(-\frac{c_1 + \frac{\beta}{4}c_2}{\beta c_1 c_2} l\right) I_0\left(\frac{c_1 - \frac{\beta}{4}c_2}{\beta c_1 c_2} l\right), \quad l \geq 0, \tag{3.34}$$

with $I_0(\cdot)$ representing the modified Bessel function of the first kind.

*Proof.* To help the presentation, define

$$A(S) \triangleq 2N_w \left(-\ln S^{1/2} - \frac{1}{\beta} + \frac{1}{\beta} S^{\beta/2}\right), \tag{3.35}$$

$$B(S) \triangleq \frac{\Gamma(3/\beta)}{\Gamma(1/\beta)} \frac{\beta(\beta-1)\Gamma(1-1/\beta)}{\Gamma(1/\beta)} S^{\beta/2-1}. \tag{3.36}$$

Thus, $L$ can be further expressed as

$$L = A(S) + M^2 B(S), \tag{3.37}$$

In order to obtain the distribution of $L$, the key step is to get the characteristic function, which can be expressed as

$$\begin{aligned} \psi_L(t) &= \mathbb{E}(e^{it[A(S)+M^2B(S)]}) \\ &= \mathbb{E}_S[\mathbb{E}_M(e^{it[A(S)+M^2B(S)]}|S)]. \end{aligned} \tag{3.38}$$

The correlation between $M$ and $S$ is zero; we use the following approximation:

$$\mathbb{E}_M(e^{it[A(S)+M^2B(S)]}|S) = e^{itA(S)}\, \mathbb{E}_M(e^{itM^2B(S)}),$$

which leads to

$$\psi_L(t) = \mathbb{E}_S[e^{itA(S)}\, \mathbb{E}_M(e^{itM^2B(S)})]. \tag{3.39}$$

According to the distribution of $M$ in (4.9) from Lemma 3.2.3, it is easy to conclude that $M^2 \xrightarrow[N\to\infty]{d} \chi_1^2$ for $N \gg 1$. So the characteristic function of $M^2$ can be derived as

$$\psi_{M^2}(t) \approx (1 - 2it)^{-1/2}. \tag{3.40}$$

Using the linear operation property of the characteristic function, we have

$$\mathbb{E}_M(e^{itM^2B(S)}) \approx [1 - 2itB(S)]^{-1/2}. \tag{3.41}$$

By plugging Eq. (3.41) into Eq. (3.39), we get

$$\psi_L(t) \approx \int_{-\infty}^{+\infty} e^{itA(S)}[1 - itB(S)]^{-1/2} f_S(S)\,\mathrm{d}S, \tag{3.42}$$

where $f_S(S)$ is the probability distribution function of the variable $S$.

Three steps are conducted for the calculations of (3.42). Firstly, from Lemma 3.2.3, the PDF of $S$ can be approximated in (3.43) for $N \gg 1$.

$$f_S(S) \approx \frac{1}{\sqrt{2\pi}\sqrt{c_2/N}} e^{-\frac{(S-1)^2}{2c_2/N}}. \tag{3.43}$$

Secondly, the term $[1-2itB(S)]^{-1/2}$ is expanded with the help of Taylor series around the point $S = 1$ as follows, since $S$ approaches to 1 when $N \gg 1$:

$$\begin{aligned}[1 - 2itB(S)]^{-1/2} = {} & (1 - 2itc_1)^{-1/2} \\ & + \frac{c_2}{2}(1 - 2itc_1)^{-1/2} it \left(\frac{\beta}{2} - 1\right)(S - 1) \\ & + o(S - 1).\end{aligned} \tag{3.44}$$

Thirdly, with the help of $A(S)$ in (3.35), by performing Taylor expansion around the point $S = 1$, we have that Eq. (3.35) can be expanded as

$$A(S) = N\frac{\beta}{4}(S - 1)^2 + o[(S - 1)^2]. \tag{3.45}$$

By plugging (3.43) - (3.45) into (3.41) and ignoring the lower order terms of $S - 1$, we get that $\psi_L(t)$ can be approximated as follows:

$$\begin{aligned}\psi_L(t) &\approx \frac{(1 - 2itc_1)^{-1/2}}{\sqrt{2\pi}\sqrt{c_2/N}} \int_{-\infty}^{+\infty} e^{-\left[\frac{N}{2c_2} - \frac{1}{4}it\beta N\right](x-1)^2}\,\mathrm{d}x \\ &= (1 - 2itc_1)^{-1/2}\left(1 - \frac{1}{2}it\beta c_2\right)^{-1/2}.\end{aligned} \tag{3.46}$$

Thus, according to [110] (Chapter 5, page 37), the probability density function of this variable can be derived as in (3.34).

□

Based on the derived PDF of $L$, the threshold or rejection region for the detection rule in (3.8) can be determined to meet the false alarm rate requirement. The equation of the threshold in (3.19) becomes

$$\mathbb{P}_L(L \geq 2ND_{th}) = \int_{2ND_{th}}^{+\infty} f_L(l)\,\mathrm{d}l = \gamma. \tag{3.47}$$

With the PDF result in (3.34), the threshold value can be obtained numerically with one dimensional line search methods.

### Adaptive Threshold

In the previous method, the PDF of the KLD is derived considering the randomness in $\hat{m}$ and $\hat{\sigma}^2$ jointly. As can be seen from the previous subsection that the derivations are complicated and approximations are made for tractable analysis. An alternative way to handle the complication is to consider the PDF of the KLD with respect to one random parameter while treating the other two as conditions, that is, to work on the conditional PDF. Since the estimation of the shape parameter does not have an explicit form, its distribution cannot be obtained analytically. Thus two viable choices are: 1) to use values of $\hat{\sigma}^2, \hat{\beta}$ as conditions, and 2) to use values of $\hat{m}, \hat{\beta}$ as conditions. The thresholds derived following this method depend on the values of the estimated parameters, which are obtained with (3.3), (3.4) and (3.6) using data samples in a window, thus the scheme is referred to as an adaptive threshold method.

When the values of $\hat{\sigma}^2, \hat{\beta}$ are seen as conditions, the only random variable in the KLD formula in (3.9) is $\hat{m}$ as given in (3.3). Further, we use the approximation: $f_{\hat{M}|\hat{\sigma}^2,\hat{\beta}}(m) \approx f_{\hat{M}}(m)$. In other words, the dependence of $\hat{m}$ to $\hat{\sigma}^2, \hat{\beta}$ are ignored. It is easily seen that the correlation between $\hat{m}$ and $\hat{\sigma}^2, \hat{\beta}$ diminishes as the window size $N$ increases. Thus the approximation is valid for large $N$.

To simplify the notation, we define

$$c_3 \triangleq \ln \frac{\hat{\beta}\Gamma(1/\beta)\sqrt{\Gamma(1/\beta)\Gamma(3/\hat{\beta})}}{\beta\Gamma(1/\hat{\beta})\sqrt{\Gamma(1/\hat{\beta})\Gamma(3/\beta)}}; \tag{3.48}$$

$$c_4 \triangleq \left(\frac{\sqrt{\Gamma(1/\beta)\Gamma(3/\hat{\beta})}}{\sqrt{\Gamma(1/\hat{\beta})\Gamma(3/\beta)}}\right)^{\beta} \frac{\Gamma(\beta/\hat{\beta}+1/\hat{\beta})}{\Gamma(1/\hat{\beta})}; \tag{3.49}$$

$$c_5 \triangleq \beta(\beta-1)\left(\frac{\sqrt{\Gamma(1/\beta)\Gamma(3/\hat{\beta})}}{\sqrt{\Gamma(1/\hat{\beta})\Gamma(3/\beta)}}\right)^{\beta-2} \frac{\Gamma(3/\beta)\Gamma(\beta/\hat{\beta}-1/\hat{\beta})}{\Gamma(1/\hat{\beta})\Gamma(1/\beta)}. \tag{3.50}$$

Thus $D_{kl}$ in (3.9) can be rewritten as

$$D_{kl} = c_3 + \ln\frac{\sigma}{\hat{\sigma}} + c_4\left(\frac{\hat{\sigma}}{\sigma}\right)^{\beta} + c_5\left(\frac{\hat{\sigma}}{\sigma}\right)^{\beta-2}\frac{(\hat{m}-m)^2}{2\sigma^2} - \frac{1}{\hat{\beta}}, \tag{3.51}$$

It can be shown straightforwardly that $c_4 > 0$ and $c_5 > 0$ for $\beta > 1$. In the following theorem, the conditional PDF of the KLD given $\hat{\sigma}^2, \hat{\beta}$ is derived, from which the corresponding threshold value can be found.

**Theorem 3.2.5.** Define $L_1 \triangleq 2ND_{kl}$. When $\beta > 1$ and $N \gg 1$, the conditional PDF of $L_1$ given estimation values $\hat{\sigma}^2$ and $\hat{\beta}$ can be approximated by $f_{L_1|\hat{\sigma}^2,\hat{\beta}}(l|\hat{\sigma}^2,\hat{\beta})$ defined as follows:

$$f_{L_1|\hat{\sigma}^2,\hat{\beta}}(l|\hat{\sigma}^2,\hat{\beta}) \triangleq \frac{(l-2NC_1)^{-\frac{1}{2}}}{\sqrt{2\pi}c_5^{1/2}}\left(\frac{\hat{\sigma}}{\sigma}\right)^{1-\frac{\beta}{2}} e^{-\frac{l-2NC_1}{2c_5}\left(\frac{\hat{\sigma}}{\sigma}\right)^{2-\beta}} u(l-2NC_1), \tag{3.52}$$

where $u(\cdot)$ is the unit step function, and

$$C_1 \triangleq c_3 + \ln\frac{\sigma}{\hat{\sigma}} + c_4\left(\frac{\hat{\sigma}}{\sigma}\right)^{\beta} - \frac{1}{\hat{\beta}}.$$

*Proof.* With the help of (3.48) and (3.49), $L_1$ can be rewritten as

$$L_1 = 2NC_1 + c_5\left(\frac{\hat{\sigma}}{\sigma}\right)^{\beta-2} M^2. \tag{3.53}$$

Since $c_5 > 0$ for $\beta > 1$, the probability of $L_1 < l$ can be calculated as

$$\begin{aligned}\mathbb{P}(L_1 < l) &= \mathbb{P}(2NC_1 + c_5\left(\frac{\hat{\sigma}}{\sigma}\right)^{\beta-2} M^2 < l) \\ &= \mathbb{P}\left(M^2 < \frac{l-2NC_1}{c_5}\left(\frac{\hat{\sigma}}{\sigma}\right)^{2-\beta}\right).\end{aligned} \tag{3.54}$$

According to the distribution of $M$ in Lemma 3.2.3, it is easy to approximate the distribution of $M^2$ as follows

$$f_{M^2}(m) \approx \frac{1}{\sqrt{2\pi}} m^{-\frac{1}{2}} e^{-\frac{m}{2}}. \tag{3.55}$$

Thus, Eq. (3.54) can be further calculated as

$$\mathbb{P}(L_1 < l) \approx \int_{-\infty}^{\frac{l-2NC_1}{c_5}\left(\frac{\hat{\sigma}}{\sigma}\right)^{2-\beta}} \frac{1}{\sqrt{2\pi}} m^{-\frac{1}{2}} e^{-\frac{m}{2}} u(l - 2NC_1)\, \mathrm{d}m. \tag{3.56}$$

By taking the first order derivation over $l$ in Eq. (3.56), the approximated distribution of $L_1$ shown in Eq. (3.52) can be derived. □

Similarly, when the values of $\hat{m}, \hat{\beta}$ are seen as conditions, we use $f_{\hat{S}|\hat{m},\hat{\beta}}(s) \approx f_{\hat{S}}(s)$. With the help of variable $S$ defined in (3.23), the KLD in (3.51) can be further expressed as

$$D_{kl} = c_3 - \ln(S^{1/2}) + c_4 S^{\beta/2} + c_5 S^{\beta/2-1} \frac{(\hat{m} - m)^2}{2\sigma^2} - \frac{1}{\hat{\beta}}.$$

By invoking the Taylor expansion around the point $S = 1$, the terms with order higher than 2, namely, $o[(S-1)^2]$, diminish to 0 for $N \to \infty$ surely. So we have

$$D_{kl} \approx D_{kl}|_{\hat{m},\hat{\beta}}^{asy} \triangleq C_2 + C_3 \frac{(S-1)}{2\sqrt{Nc_2}} + C_4 \frac{(S-1)^2}{2c_2}, \tag{3.57}$$

where

$$C_2 \triangleq c_3 + c_4 + c_5 \frac{(\hat{m} - m)^2}{2\sigma^2} - \frac{1}{\hat{\beta}}; \tag{3.58}$$

$$C_3 \triangleq 2\sqrt{Nc_2} \left[ -\frac{1}{2} + \frac{\beta}{2} c_4 + \left( \frac{\beta}{2} - 1 \right) c_5 \frac{(\hat{m} - m)^2}{2\sigma^2} \right]; \tag{3.59}$$

$$C_4 \triangleq c_2 \left[ \frac{1}{2} + c_4 \frac{\beta}{2} \left( \frac{\beta}{2} - 1 \right) + c_5 \frac{(\hat{m} - m)^2}{2\sigma^2} \left( \frac{\beta}{2} - 1 \right) \left( \frac{\beta}{2} - 2 \right) \right]. \tag{3.60}$$

The following theorem is derived for the conditional distribution of KLD given $\hat{m}, \hat{\beta}$.

**Theorem 3.2.6.** When $\beta > 1$ and $N \gg 1$, the conditional PDF of $L_1$ given estimation values $\hat{m}$ and $\hat{\beta}$ can be approximated by $f_{L_1|\hat{m},\hat{\beta}}(l|\hat{m},\hat{\beta})$ defined as follows:

$$f_{L_1|\hat{m},\hat{\beta}}(l|\hat{m},\hat{\beta}) \triangleq \begin{cases} \cosh\left(\sqrt{\left|\frac{C_3}{2C_4}\right|\frac{l-2NC_2+\frac{C_3^2}{4C_4}}{C_4}}\right) u\left(\frac{l-2NC_2+\frac{C_3^2}{4C_4}}{C_4}\right) \\ \times \frac{\left|l-2NC_2+\frac{C_3^2}{4C_4}\right|^{-1/2}}{\sqrt{2\pi}|C_4|^{1/2}} e^{-\frac{1}{2}\left(\frac{l-2NC_2+\frac{C_3^2}{2C_4}}{C_4}\right)}, & \text{if } C_4 \neq 0; \\ \frac{1}{2\pi|C_3|} e^{-\frac{(l-2NC_2)^2}{2C_3^2}}, & \text{if } C_4 = 0, \end{cases} \tag{3.61}$$

where $\cosh(x) = e^x/2 + e^{-x}/2$ is the hyperbolic cosine function.

*Proof.* With the help of (3.57), $L_1$ can be rewritten as

$$L_1 = 2NC_2 + C_3\frac{\sqrt{N}(S-1)}{\sqrt{c_2}} + C_4\frac{N(S-1)^2}{c_2}.$$

In order to get the distribution of $L_1$, there are four cases to be considered.

When $C_4 \neq 0$, $L_1$ can be further rewritten as

$$L_1 = 2NC_2 - \frac{C_3^2}{4C_4} + C_4\left(\frac{\sqrt{N}(S-1)}{c_2} + \frac{C_3}{2C_4}\right)^2. \tag{3.62}$$

According to (4.10), it is obvious that $\sqrt{N}(S-1)/c_2 \xrightarrow[N\to\infty]{d} \mathcal{N}(0,1)$. Thus, the distribution of $L_1$ can be approximated as a linear transformation of a non-central chi-square distribution of one degree of freedom.

By defining $D(S) \triangleq \left(\frac{\sqrt{N}(S-1)}{c_2} + \frac{C_3}{2C_4}\right)^2$; the PDF of $D$, denoted as $f_D(s)$, can be obtained as

$$f_D(s) = \frac{1}{\sqrt{2\pi s}} e^{-\frac{1}{2}\left(s+\frac{C_3^2}{4C_4^2}\right)} \cosh\left(\sqrt{\left|\frac{C_3}{2C_4}\right| s}\right) u(s).$$

Thus, we have

$$\mathbb{P}(L_1 < l) = \mathbb{P}\left(2NC_2 - \frac{C_3^2}{4C_4} + C_4 D(S) < l\right).$$

If $C_4 > 0$, we have

$$\mathbb{P}(L_1 < l) = \mathbb{P}\left(D(S) < \frac{l - 2NC_2 + \frac{C_3^2}{4C_4}}{C_4}\right)$$
$$= \int_0^{\frac{l-2NC_2+\frac{C_3^2}{4C_4}}{C_4}} f_D(s) u\left(\frac{l - 2NC_2 + \frac{C_3^2}{4C_4}}{C_4}\right) \mathrm{d}s. \tag{3.63}$$

Similarly, if $C_4 < 0$, it can be obtained that

$$\mathbb{P}(L_1 < l) = \mathbb{P}\left(D(S) > \frac{l - 2NC_3 + \frac{C_4^2}{4C_4}}{C_4}\right)$$
$$= \int_{\frac{l-2NC_2+\frac{C_3^2}{4C_4}}{C_4}}^{+\infty} f_D(s) u\left(\frac{l - 2NC_2 + \frac{C_3^2}{4C_4}}{C_4}\right) \mathrm{d}s, \tag{3.64}$$

By the first order derivative over $l$ in (3.63) and (3.64), the distribution of $L_1$ in (3.61) for the case of $C_4 \neq 0$ can be derived by combining (3.63) and (3.64).

When $C_4 = 0$, since $\sqrt{N}(S-1)$ is Gaussian distributed, according to Lemma 3.2.3, it is easy to obtain the PDF of $L_1$ shown in (3.61). □

Particularly, if $C_3 = 0$ and $C_4 = 0$, we have $L_1 = 2NC_2$ as a constant, thus $f_{L_1|\hat{m},\hat{\beta}}(l|\hat{m},\hat{\beta}) = \delta(l - 2NC_2)$, or equivalently,

$$\mathbb{P}(l = 2NC_2|\hat{m},\hat{\beta}) = 1.$$

This is a special case and can be derived from the second case in (3.61).

Similar to the previous case, the threshold is chosen based on the level of false alarm probability $\gamma$. The threshold values $D_{th,m}$ and $D_{th,s}$ satisfy the following equations:

$$\mathbb{P}_{L_1|\hat{\sigma^2},\hat{\beta}}(D_{kl} \geq D_{th,m}|\hat{\sigma^2},\hat{\beta}) = \int_{2ND_{th,m}}^{\infty} f_{L_1|\hat{\sigma^2},\hat{\beta}}(l|\hat{\sigma^2},\hat{\beta})\,\mathrm{d}l = \gamma, \tag{3.65}$$

$$\mathbb{P}_{L_1|\hat{m},\hat{\beta}}(D_{kl} \geq D_{th,s}|\hat{m},\hat{\beta}) = \int_{2ND_{th,s}}^{\infty} f_{L_1|\hat{m},\hat{\beta}}(l|\hat{m},\hat{\beta})\,\mathrm{d}l = \gamma. \tag{3.66}$$

Via the results in Theorems 3.2.5 and 3.2.6, the values of $D_{th,m}$ and $D_{th,s}$ can be obtained by solving (3.65) and (3.66) via one dimensional line search methods. Notice that $D_{th,m}$ depends on the samples of the data set $\{x_i\}_{n=1}^N$ via dependence on $\hat{\sigma}^2, \hat{\beta}$, and similarly $D_{th,s}$ depends on $\hat{m}, \hat{\beta}$. Thus they vary with the data set.

## 3.3 Summary of the Detection Algorithms

In this section, the KLD based abnormality detection algorithms are summarized for constant threshold and adaptive threshold methods. Then discussions are given for several practical issues.

### 3.3.1 Detection Algorithm with a Constant Threshold

The detection algorithm with a constant threshold is summarized as Algorithm 2. There are three stages, namely, initialization, off-line training, and online detection.

Specifically, in the initialization stage, the data size $N$ and the false alarm rate level $\gamma$ are assigned. In the off-line training stage, the parameters $m, \sigma^2, \beta$ for the normal state are obtained from historical data via (3.3), (3.4) and (3.6). Further, with the help of Theorem 3.2.4, the threshold $D_{th}$ is calculated. For the online detection stage, a data set of size $N$ is collected. The mean and the variance for this set are estimated, i.e., $\hat{m}$ and $\hat{\sigma}^2$ are obtained by using (3.3) and (3.4). Then the KLD $D_{kl,\beta}^{asy}$ is calculated using (3.24) and compared with $D_{th}$. If the condition $D_{kl,\beta}^{asy} > D_{th}$ holds, the algorithm returns the alarm state 1 and saves the sample data set and the KLD value $D_{kl,\beta}^{asy}$ for further analysis; otherwise, the algorithm returns the alarm state 0. The algorithm can be conducted online continuously for every $N$ data samples.

### 3.3.2 Detection Algorithm with an Adaptive Threshold

The detection algorithm with an adaptive threshold is summarized in Algorithm 3, which also have three stages. The initialization stage is exactly the same as the constant threshold case, while the other two stages are different. In the offline training stage, the mean, the variance and the shape parameter are estimated based on historical data under the normal state. But the threshold calculation is conducted in the online detection stage. For the online detection stage, a data set of size $N$ is collected. The mean, variance, and shape parameter for this data set are estimated, i.e., $\hat{m}$, $\hat{\sigma}^2$, $\hat{\beta}$ are obtained by

**Algorithm 2** Detection algorithm with a constant threshold.

**Input**: Data collected under the normal condition and data inside a moving window $\boldsymbol{x} = \{x_i\}_{i=1}^{N}$ for detection.

**Output**: Hypothesis test results for (3.8), i.e., alarm states.

**Initialization**: Assign values to $N$ and $\gamma$.

**Off-line training**: Based on historical data for normal state, estimate $m$, $\sigma^2$ and $\beta$ by (3.3), (3.4) and (3.6); Find the threshold $D_{th}$ by (3.19).

**On-line detection**:

```
while N data samples are collected do
    Estimate m̂ and σ̂² by (3.3) and (3.4);
    Calculate D^{asy}_{kl,β} by (3.24);
    if D^{asy}_{kl,β} ≥ D_th then
        Return Alarm state as 1;
        Save x and D^{asy}_{kl,β} for further analysis;
    else
        Return Alarm state as 0;
    end
end
```

using (3.3), (3.4), and (3.6). Then the adaptive thresholds $D_{th,m}$ and $D_{th,s}$ are calculated from (3.65) and (3.66), respectively. Next, the KLD $D_{kl}$ in (3.9) are calculated and compared with $D_{th,m}$ and $D_{th,s}$. If the condition $D_{kl} > D_{th,m}$ or $D_{kl} > D_{th,s}$ holds, the algorithm returns the alarm state 1 and save the sample set and $D_{kl}$ for further analysis; otherwise, the algorithm returns the alarm state 0. The algorithm can be conducted online continuously for every $N$ data samples.

### 3.3.3 Discussions

There are two different aspects between the constant and adaptive threshold methods. Specifically, the main difference exists in when and how the thresholds are determined. In the constant threshold case, the threshold can be calculated off-line with the historical data under the normal condition, while in the adaptive threshold case, the threshold is derived on-line according to both historical data and current data inside a window. Moreover, in the constant threshold based method, the shape parameter of the data inside a

**Algorithm 3** Detection algorithm with an adaptive threshold.

**Input**: Data collected under the normal condition and data inside a moving window $\boldsymbol{x} = \{x_i\}_{i=1}^{N}$ for detection.

**Output**: Hypothesis test results for (3.8), i.e., alarm states.

**Initialization**: Assign values to $N$ and $\gamma$.

**Off-line training**: Based on historical data for the normal state, estimate $m$, $\sigma^2$ and $\beta$ by (3.3), (3.4) and (3.6).

**On-line detection**:

```
1 while N data samples are updated do
2     Estimate m̂, σ̂² and β̂ by (3.3), (3.4) and (3.6);
      Calculate D_{th,m} from (3.65) and D_{th,s} from (3.66);
      Calculate D_{kl} by (3.9);
      if D_{kl} ≥ D_{th,m}, OR D_{kl} ≥ D_{th,s}, then
3         Return Alarm state as 1;
          Save x and D_{kl} for further analysis;
4     else
5         Return Alarm state as 0;
6     end
7 end
```

window is not used in the detection due to the simplification of using $\beta$ in the data fitting instead of $\hat{\beta}$, which may affect its performance. For the adaptive threshold based method, the shape parameter is estimated for each data set and its difference to $\beta$ obtained from historical normal data is considered in the detection process. However, the on-line estimation of the shape parameter incurs higher computational load and requires a larger window size with longer detection delay.

In the application of the constant and adaptive threshold schemes, one practical issue is the selection of the window size $N$ in the initiation stage. On one hand, a large $N$ is needed for the precision of the parameter estimation and the validity of asymptotic analysis on the PDF s and the corresponding threshold selection; on the other hand, it also increases the computational load of the detection algorithm s.

Another practical problem is related to the offset between the derived PDFs for $L$, $L_1$ and those for data under the normal condition. Due to sample outliers or even errors in the distribution model, the derived PDFs of $L$ and $L_1$

may fail to describe the corresponding true distributions. In this case, the proposed methods can still be used for detection but the threshold selection needs to be adapted. A practical and general method is to use the kernel density estimation to find the PDF of the test statistic. For the constant threshold case, the KLD can be calculated numerically from (3.7), by either directly obtaining the two PDFs via the kernel density estimation or estimating the probability density ratio. More details can be found in [144] and [59]. Then the distribution of $L$ under the normal condition can be estimated and used for the constant threshold calculation in (3.19). As for the adaptive threshold case, the distributions of $M$ and $S$ can be derived numerically, so that the adaptive thresholds can be set.

Another assumption made in this chapter is that $x_i$'s are independent samples. For systems with correlated data, the distributions of $M$ and $S$ do not follow the results derived in Lemma 3.2.3. To tackle this problem, one option is to apply de-correlation methods, such as auto-regressive model and whitening transformation [108]. Alternatively, the kernel density estimation can be used to obtain the distribution of $L$ in the constant threshold case, and the distributions of $M$ and $S$ for the conditional threshold cases. Different from the kernel density estimation for data set with i.i.d. samples, the kernel bandwidth for correlated data needs to be carefully selected to achieve good trade-offs between the estimation bias and variance [27]. Also, considering that the values of $L$ and $S$ are nonnegative, the Gamma kernel [30] can be used instead of the Gaussian kernel, since the Gaussian kernel will cause estimation bias for samples near 0.

### 3.3.4 Extension to Multivariate Case

There are three possible ways to extend this work to multivariate cases. One straightforward way is to apply the KLD calculation and its distribution results directly for each variable and perform combining schemes (e.g. majority voting) to determine whether an alarm should be generated. The second

way is based on the multivariate GGD proposed in [53]. According to the definition, the multivariate generalized Gaussian variable consists of several i.i.d. univariate generalized Gaussian variables with the same shape parameter, and it can be modeled as a linear transformation from a standardized multivariate generalized Gaussian random variable by the singular value decomposition. Thus, the KLD for each univariate generalized Gaussian random variable can be calculated following similar steps in Appendix A. Then, by using the addition property of the KLD, the total KLD can be derived. Also, the results in Lemma 3.2.3 still hold for each generalized Gaussian random variable. By following similar steps in Appendixes C, D and E, the joint and conditional distributions of the KLD can be derived according to [110]. More generally, as the third way, by assuming that variables are mutually independent, the shape parameter for each variable can be different. Similarly to the second case, with the help of existing results in Lemma 3.2.3 and following similar steps in Appendixes A, C, D and E, the distributions of KLD can be derived according to [19] for the constant and adaptive threshold algorithms.

For all three ways, the results in the univariate case are fundamental for the multivariate extension. Considering the focus of this work, we only provide the above brief discussion on extensions to multivariate cases in this chapter and leave further analysis and case studies for the future.

## 3.4 Case Studies

In this section, simulation results are shown to validate the analytical results and to compare with other methods with both simulated and industrial data sets. Three widely used linear and non-linear filter detectors, namely, the mean absolute deviation (MAD) filter, the moving average (MA) filter, the median filter, are considered and compared with the proposed KLD based methods. Given the data sequence $\boldsymbol{x} = \{x_i\}_{i=1}^{N}$, the MA filter calculates

$$y_{MA} = \hat{m} = \frac{1}{N}\sum_{i=1}^{N} x_i; \tag{3.67}$$

the MAD filter calculates

$$y_{MAD} = \frac{1}{N}\sum_{i=1}^{N}|x_i - \hat{m}|^2; \tag{3.68}$$

and the median filter gets

$$y_M = x_{(v+1)} = \text{median}_N(\boldsymbol{x}), \tag{3.69}$$

where $x_{(i)}$ is the $i$th-order statistic, and if $N$ is odd, $v = (N-1)/2$; if $N$ is even, $v = N/2$ or $v = N/2 - 1$. Then the filter output $y_{MAD}$ (or $y_{MA}$ , $y_M$) is compared with a threshold, and an alarm is raised if the filter output is larger than the threshold. The MAD filter is a non-linear filter that takes both mean and variance changes into consideration to detect abnormalities. The MA filter is effective in capturing the mean difference in data set by keeping the mean and reducing the variance. It has been proved that the MA filter is optimal among linear filters under log-concave and symmetric data distributions [34]. The median filter, a non-linear filter, is shown to be effective in removing nuisance alarms, especially the chartering ones [116].

### 3.4.1 Simulation Results on KLD

In this section, the behaviors of the KLD are studied. The simulation results are compared with the derived analytical results in Theorem 3.2.1, Theorem 3.2.4, and Theorem 3.2.5.

#### KLD Calculation

The first to show is the KLD value under normal condition. There are two pairs of KLDs to be compared. The first pair is for the constant threshold case. One is $D_{kl,\beta}$ in (3.20) and the other is $D^{asy}_{kl,\beta}$ given in (3.24). And the second pair is under the adaptive threshold case, namely, $D_{kl}$ in (3.7) and $D_{kl}|^{asy}_{\hat{m},\hat{\beta}}$ shown in (3.57).

The relative difference in percentage $\epsilon$ is used to quantify the offset of $D_{kl}$ by using its asymptotic approximation. For these two cases, the relative

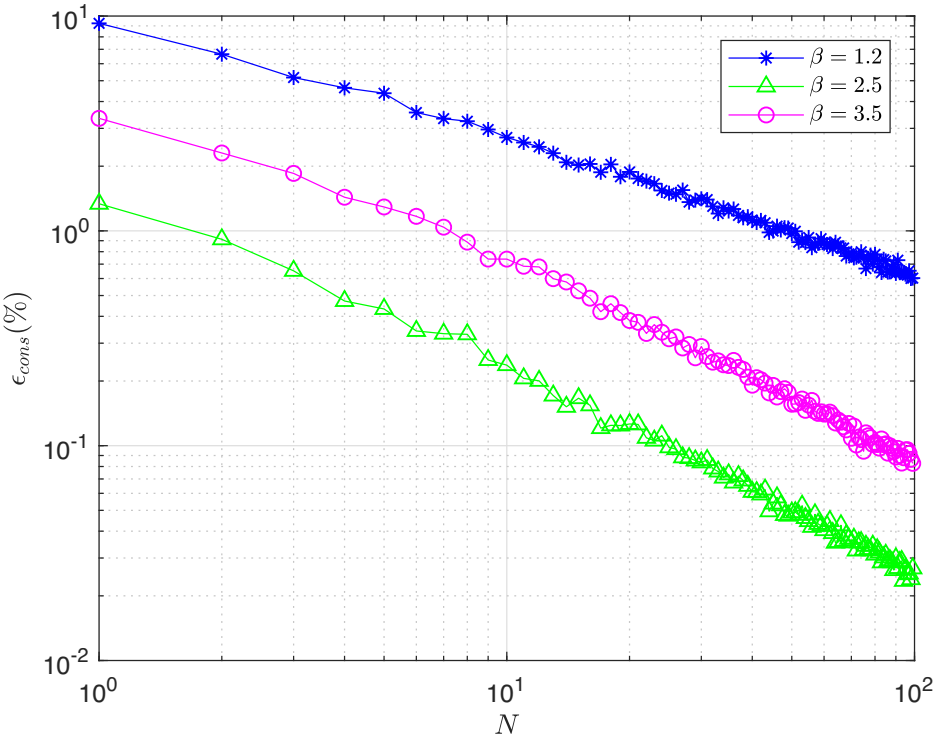



**Figure 3.2:** Relative difference of KLD value between $D_{kl,\beta}$ in (3.20) and $D_{kl,\beta}^{asy}$ in (3.24) for the constant threshold case.

differences are defined respectively as:

$$\epsilon_{cons} \triangleq \frac{\left|D_{kl,\beta} - D_{kl,\beta}^{asy}\right|}{D_{kl,\beta}} \times 100\%, \tag{3.70}$$

and

$$\epsilon_{adap} = \frac{\left|D_{kl} - D_{kl}|_{\hat{m},\hat{\beta}}^{asy}\right|}{D_{kl}} \times 100\%. \tag{3.71}$$

In the simulation, three distributions GGD$(0, 1, 1.2)$, GGD$(0, 1, 2.5)$ and GGD$(0, 1, 3.5)$ are considered. The relative difference in (3.70) for the constant threshold case is shown in Figure 3.2 for different window size $N$. For each $N$ value, 10000 randomly generated data sets are used to calculate the average difference. It can be seen that the difference is small even for small window size, e.g., less than 3% for $N \geq 10$. And the difference diminishes as $N$ increases. This validates the derived asymptotic results on KLD.

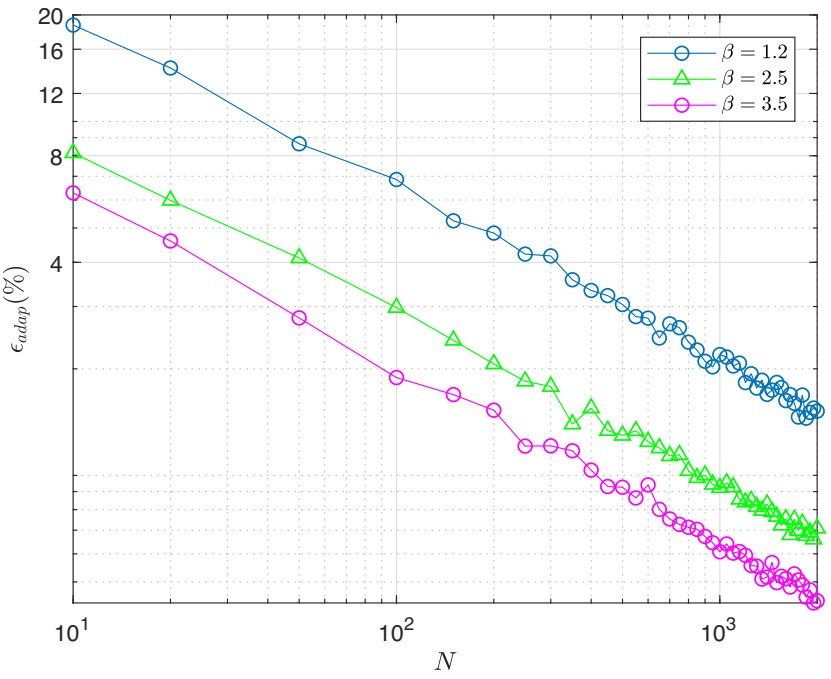



**Figure 3.3:** Relative difference of KLD value between $D_{kl}$ in (3.7) and $D_{kl}|^{asy}_{\hat{m},\hat{\beta}}$ in (3.57) for the conditional threshold case.

With the same parameter settings, the relative difference of the adaptive threshold case in (3.71) is shown in Figure 3.3. It can be observed that the relative difference decreases with larger window size. Also, the relative difference is less than 5% for window sizes greater than 200. Compared with the constant threshold case, the relative difference is higher with the same window size.

### KLD Distribution

The next to study is the PDF of the KLD. Figure 3.4 shows the relative $p$-percentile difference $\epsilon_p$ between the derived approximate PDF in Theorem 3.2.4 and the PDF obtained from the Monte Carlo simulations for data following GGD$(0, 1, 1.2)$, GGD$(0, 1, 2.5)$ and GGD$(0, 1, 3.5)$. The relative $p$th-

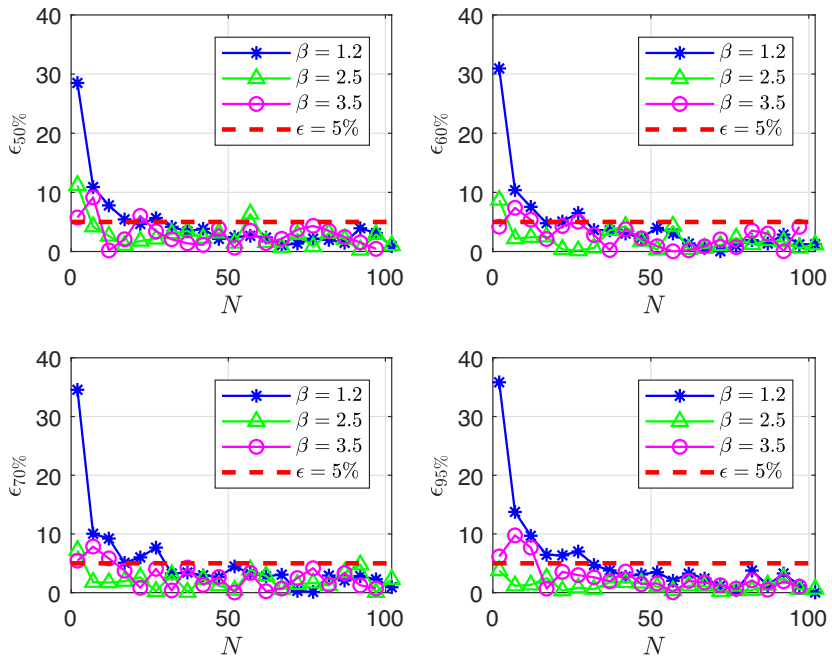



**Figure 3.4:** Relative PDF difference between asymptotic closed-form result in (3.34) and the value obtained from the Monte Carlo simulations.

percentile difference is defined as

$$\epsilon_p = \frac{\left|\hat{\psi}_p - \psi_p\right|}{\hat{\psi}_p} \times 100\%,$$

where $\psi_p$ is the $p$th-percentiles of the KLD distribution calculated from Theorem 3.2.4 (i.e.,$f_L(l)$), and $\hat{\psi}_p$ is that of the KLD distribution obtained from the Monte Carlo simulations. The results where $p$ is $50, 60, 70$ and $95$ are shown in the four sub-figures in Figure 3.4 respectively. For each $p$ value, the window size $N$ ranges from 1 to 100. And for each window size, the average of 1000 data sets is conducted.

Form Figure 3.4, it can be observed that the relative differences for the four $p$ values decrease as the window size increases up to 60. It is natural since the larger the window size, the more accurate the Taylor expansion in (3.9) and the derived PDF in (3.34). Also, it can be seen that the relative differences remain steady around 5% when the window size is larger than 60.

The analysis is accurate for large enough window size.

### 3.4.2 Abnormality Detection with Simulated Data

In this subsection, the proposed detection methods based on the KLD are compared with the MAD, MA and median filters under simulated data. Further, we compare with the case when the Gaussian distribution is used under the proposed constant and adaptive threshold schemes. In order to show the performance with different $\beta$, two cases with the same mean and variance but different shape parameters are considered under the normal condition, namely GGD$(1, 1, 1.2)$ and GGD$(1, 1, 5)$. As for the abnormal condition, seven types of faults are considered, including both symmetric and non-symmetric distributions. Specifically, for the Type-1 fault, the mean increases by 0.4, i.e., the distributions become GGD$(1.4, 1, 1.2)$ and GGD$(1.4, 1, 5)$. In the Type-2 fault, 1.25 and 0.69 increase in the variance is considered in the faulty distributions, namely, GGD$(1, 2.25, 1.2)$ and GGD$(1, 1.69, 5)$. As for the Type-3 fault, samples are generated from the distribution GGD$(1.4, 2.25, 1.2)$ and GGD$(1.4, 1.69, 5)$, respectively, where both the mean and the variance change. Also, the Gamma distribution Gamma$(1, 1.6)$ is considered as a unsymmetrical distribution in Type-4 fault. Additionally, three more faults are considered in order to show the performance in detecting abnormality with the shape parameter change combined with mean and/or variance changes, namely Type-5, Type-6 and Type-7 faults. Specifically, when $\beta = 1.2$ under normal condition, Type-5, Type-6 and Type-7 faults are GGD(1.4,1,5), GGD(1,2.25,5), and GGD(1.4,2.25,5); while when $\beta = 5$, GGD(1.4.1,8), GGD(1,1.69,8) and GGD(1.4,1.69,8) are used. The PDFs under the normal condition and the seven faults are depicted in Figures 3.5 and 3.6 when $\beta = 1.2$ and $\beta = 5$ under the normal condition, respectively.

For all involved detection methods, the window size or filter length is set as 60, i.e., $N = 60$. There are 10000 data sets generated for the normal condition and for each abnormal case. In order to compare the detection performance

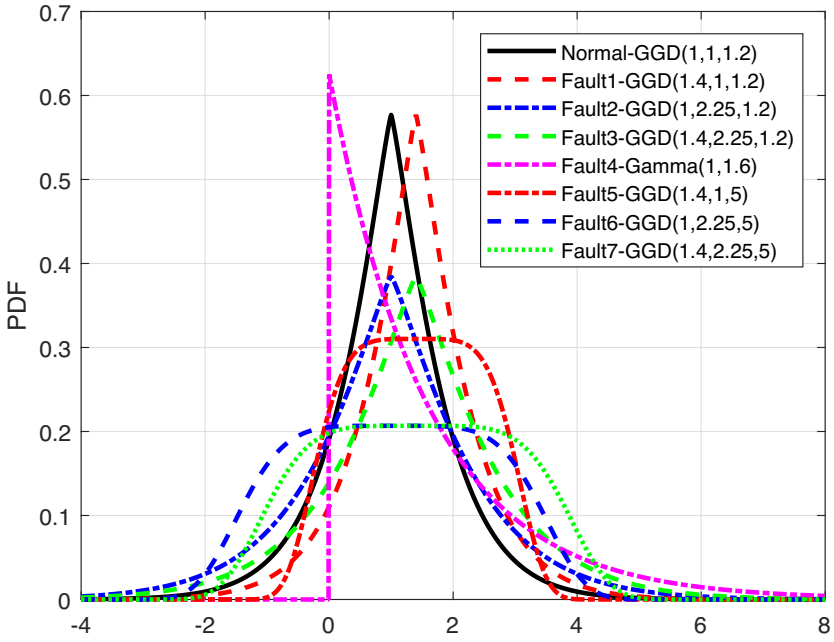



**Figure 3.5:** PDFs under the normal condition when $\beta = 1.2$ and seven types of faults.

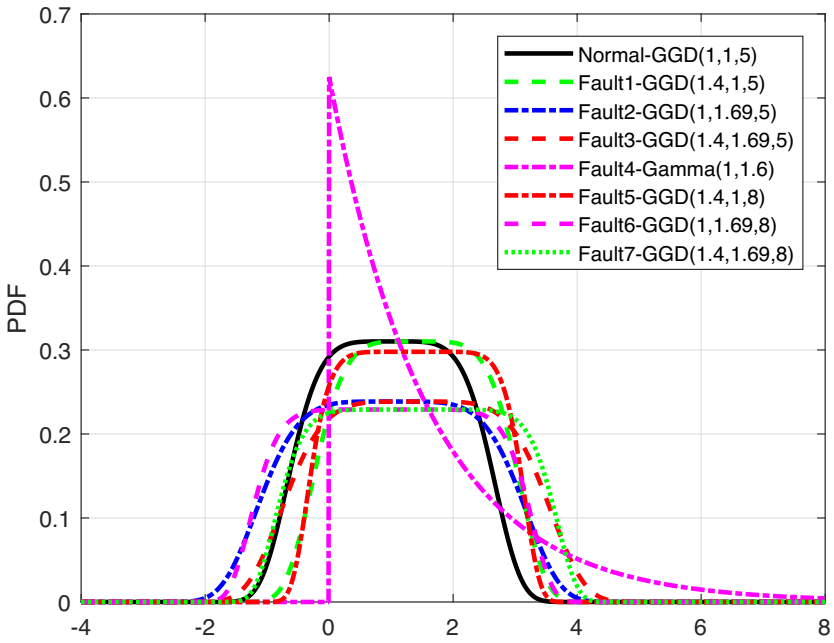



**Figure 3.6:** PDFs under the normal condition when $\beta = 5$ and seven types of faults.

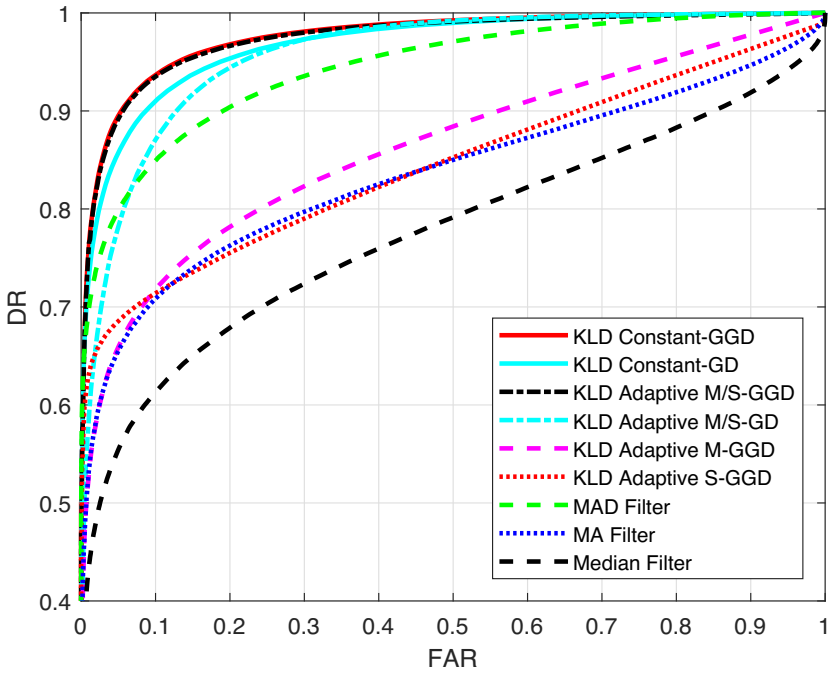



**Figure 3.7:** FAR v.s. DR with proposed KLD based methods and the MAD, MA filter when $\beta = 1.2$.

among different methods, the FAR versus DR curve is used. In the drawing, 1000 threshold values are used for each method, which generate 1000 pairs of FARs and DRs. The results for the proposed constant threshold scheme are obtained with the proposed Algorithm 1; while Algorithm 2 is used for the proposed adaptive threshold method considering $D_{th,m}$ and $D_{th,s}$ together through the 'OR' operation; thus it is denoted with 'KLD Adaptive M/S'. One can also apply the adaptive threshold method by using $D_{th,m}$ and $D_{th,s}$ separately. It follows the same steps in Algorithm 2, except it only considers either $D_{th,m}$ or $D_{th,s}$ in Step 5.

The FAR versus DR curves that comprehensively consider the all seven types of faults are shown in Figures 3.7 and 3.8. It can be observed that under the two GGDs, the two KLD based detection methods are apparently better than the other three filters. For example, when $\beta = 1.2$, the DRs at FAR=0.05 of the MAD, MA and median filters are 0.808, 0.668 and 0.598, while the proposed methods boost up the detection rates to 0.902 and 0.895,

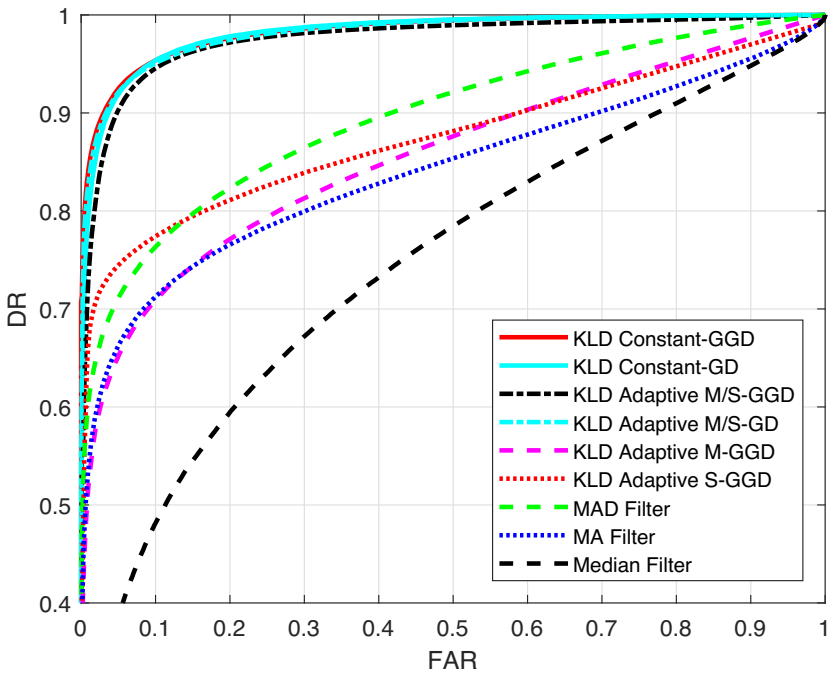



**Figure 3.8:** FAR v.s. DR with proposed KLD based methods and the MAD, MA filter when $\beta = 5$.

respectively. The figures further show that the use of the GGD model can achieve better performance than the Gaussian one when $\beta = 1.2$; while they have similar performance when $\beta = 5$. Also, compared with using $D_{th,m}$ and $D_{th,s}$ separately, the proposed adaptive threshold scheme that considers $D_{th,m}$ and $D_{th,s}$ together with the 'OR' operation shows better performance.

In order to show the performance under each fault, the DRs at the FAR of 0.05 are summarized in Table 3.1. The method with the best detection rate is marked as red under each fault. First, when $\beta = 1.2$ under the normal condition, it shows that the adaptive and constant threshold methods with the GGD model outperform the three filters and those with Gaussian model on average; when $\beta = 5$, the two proposed methods with the GGD model and Gaussian model have similar performance, which is still better than that of the three filters on average. Second, it can be seen that the MA filter outperforms other methods in detecting faults of types 1, 4 and 5, where the mean change dominant the fault feature; while the MAD filter performs generally better in

the faults of types 2, 3, 6 and 7, where the variance change dominants. Third, in a few cases the Gaussian model performs better than the GGD model, but in terms of the average performance, the GGD model is still better.

Additionally, the sensitivity of the proposed KLD based methods are tested under different noise levels. The signal-to-noise ratio (SNR) is used to quantify the noise level as SNR $= \log_{10}(\sigma_s^2/\sigma_e^2)$ dB, where $\sigma_s^2$ is the variance of the signal and $\sigma_e^2$ denotes the noise variance. Four SNR values are considered in this test, namely, 10, 20, 40 and $\infty$ dB. When SNR $= \infty$ dB, it is considered as the ideal noise-free case; when SNR = 40 dB, it is regarded as a normal noise environment, while SNR =10 dB is considered as a noisy circumstance. Under each SNR, the normal data is generated from the distributions GGD$(1, 1, 1.2)$ and GGD$(1, 1, 5)$. The faulty data consists of the same seven types of faults shown in Figures 3.5 and 3.6, respectively. The corresponding FAR versus DR curves are drawn in Figures 3.9 and 3.10. From the curve, it can be observed that the proposed methods can still achieve stable performance, even under noisy environment when SNR =10 dB.

Furthermore, simulations are conducted in order to better compare the proposed two KLD based methods and explore how the computational load changes with the window size $N$ and the shape parameter $\beta$. In this part, the normal data is generated from GGD$(1, 1, 1.2)$ and GGD$(1, 1, 5)$, while three corresponding faults of types 5 to 7 are considered. The window sizes $N = 10, 30, 60, 100$ and 200 are used. The computational load is quantified by elapsed time using the 'tic toc' function in MATLAB software on a PC with an Intel core i7-4790 CPU 3.6 GHz. For the constant threshold algorithm, the computational time is counted during Step 2 and Step 3; while for the adaptive threshold case, there are three steps counted, i.e., Step 2 to Step 4. The computational load and DRs when the FAR is 0.05 are summarized in Table 3.2. It can be seen that compared with the constant threshold scheme, the one with the adaptive threshold performs generally better especially when $\beta = 1.2$, with the cost of a higher computational load. As shown by the two

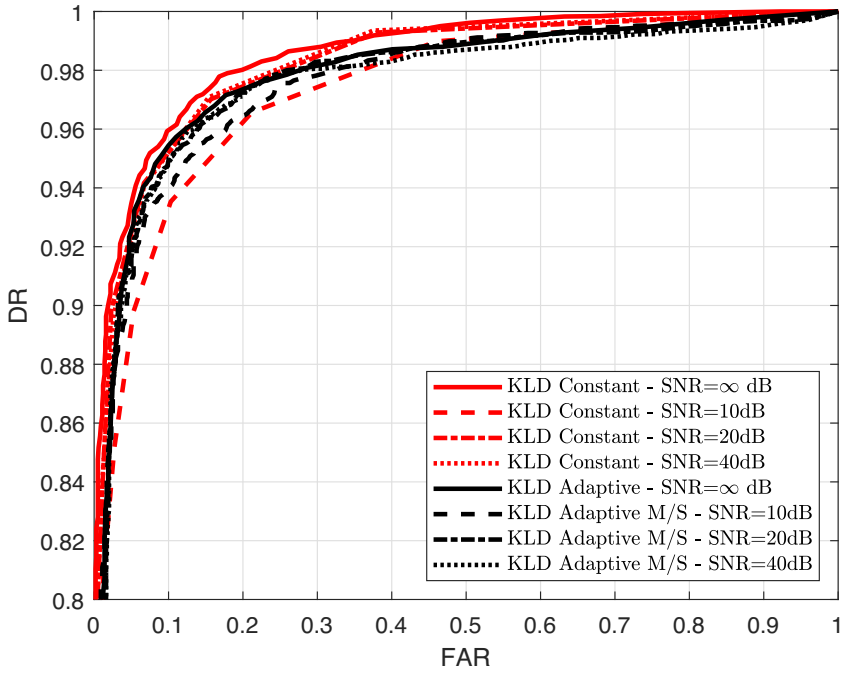



**Figure 3.9:** Sensitivity test under different SNRs in terms of FAR v.s. DR with proposed KLD based methods when $\beta = 1.2$.

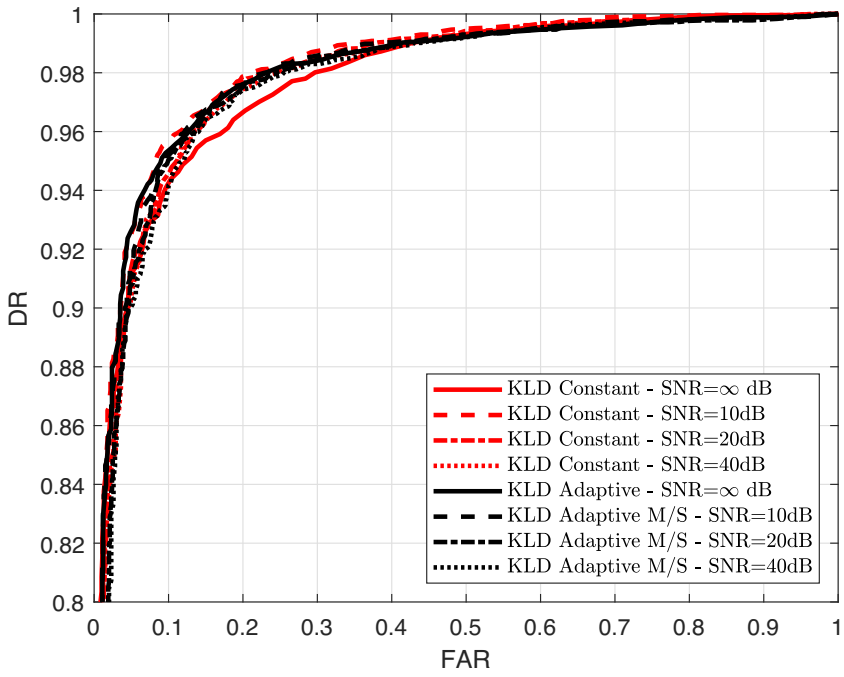



**Figure 3.10:** Sensitivity test under different SNRs in terms of FAR v.s. DR with proposed KLD based methods when $\beta = 5$.

**Table 3.1:** Detection performances for seven different faults when $\beta = 1.2$ and $\beta = 5$ under normal condition.

| | | KLD-GGD | | | | |
|---|---|---|---|---|---|---|
| $\beta$ | Faults | Constant | Adaptive M/S | Adaptive M | Adaptive S | |
| 1.2 | Type-1 | 0.820 | 0.713 | 0.872 | 0.075 | |
| | Type-2 | 0.830 | 0.944 | 0.190 | 0.947 | |
| | Type-3 | 0.965 | 0.977 | 0.775 | 0.907 | |
| | Type-4 | 0.963 | 0.954 | 0.962 | 0.909 | |
| | Type-5 | 0.788 | 0.705 | 0.870 | 0.020 | |
| | Type-6 | 0.958 | 0.978 | 0.189 | 0.977 | |
| | Type-7 | 0.989 | 0.991 | 0.775 | 0.973 | |
| | Average | 0.902 | 0.895 | 0.662 | 0.687 | |
| 5 | Type-1 | 0.809 | 0.855 | 0.872 | 0.238 | |
| | Type-2 | 0.964 | 0.877 | 0.131 | 0.965 | |
| | Type-3 | 0.993 | 0.973 | 0.808 | 0.973 | |
| | Type-4 | 0.984 | 0.990 | 0.962 | 0.995 | |
| | Type-5 | 0.804 | 0.853 | 0.869 | 0.175 | |
| | Type-6 | 0.973 | 0.824 | 0.132 | 0.928 | |
| | Type-7 | 0.994 | 0.964 | 0.807 | 0.936 | |
| | Average | 0.922 | 0.902 | 0.649 | 0.740 | |
| | | KLD-GD | | | | |
| $\beta$ | Faults | Constant | Adaptive M/S | MAD | MA | Median |
| 1.2 | Type-1 | 0.623 | 0.461 | 0.228 | 0.929 | 0.961 |
| | Type-2 | 0.899 | 0.935 | 0.961 | 0.138 | 0.135 |
| | Type-3 | 0.973 | 0.964 | 0.987 | 0.834 | 0.884 |
| | Type-4 | 0.955 | 0.935 | 0.924 | 0.978 | 0.346 |
| | Type-5 | 0.578 | 0.435 | 0.560 | 0.927 | 0.844 |
| | Type-6 | 0.990 | 0.997 | 0.998 | 0.137 | 0.264 |
| | Type-7 | 0.997 | 0.998 | 0.999 | 0.835 | 0.751 |
| | Average | 0.864 | 0.818 | 0.808 | 0.668 | 0.598 |
| 5 | Type-1 | 0.840 | 0.778 | 0.203 | 0.927 | 0.624 |
| | Type-2 | 0.905 | 0.967 | 0.945 | 0.104 | 0.103 |
| | Type-3 | 0.983 | 0.988 | 0.972 | 0.868 | 0.597 |
| | Type-4 | 0.980 | 0.977 | 0.816 | 0.978 | 0.152 |
| | Type-5 | 0.833 | 0.774 | 0.228 | 0.926 | 0.152 |
| | Type-6 | 0.922 | 0.976 | 0.964 | 0.103 | 0.113 |
| | Type-7 | 0.984 | 0.990 | 0.980 | 0.867 | 0.592 |
| | Average | 0.915 | 0.920 | 0.712 | 0.662 | 0.383 |

algorithms in Section 3.3, the extra computational time mainly comes from the shape parameter estimation process. Also, with larger window size, the

computational load gets heavier, but the detection rate is higher. However, there is no clear pattern on the relation between $\beta$ and the computational load.

**Table 3.2:** Computational load and DR with the proposed KLD based methods using different window size, when $\beta = 1.2$ and $\beta = 5$ under normal condition.

| | $N$ | 10 | 30 | 60 | 100 | 200 |
|---|---|---|---|---|---|---|
| Constant | DR (Normal $\beta = 1.2$) | 0.203 | 0.613 | 0.912 | 0.986 | 1.000 |
| | DR (Normal $\beta = 5$) | 0.450 | 0.752 | 0.923 | 0.985 | 1.000 |
| | Time (ms) $\beta = 1.2$ | 0.049 | 0.027 | 0.033 | 0.024 | 0.027 |
| | Time (ms) $\beta = 5$ | 0.028 | 0.024 | 0.024 | 0.026 | 0.029 |
| | Time (ms) $\beta = 8$ | 0.027 | 0.024 | 0.023 | 0.025 | 0.030 |
| Adaptive | DR (Normal $\beta = 1.2$) | 0.315 | 0.774 | 0.911 | 0.992 | 1.000 |
| | DR (Normal $\beta = 5$) | 0.427 | 0.754 | 0.908 | 0.989 | 1.000 |
| | Time (ms) $\beta = 1.2$ | 0.168 | 0.189 | 0.268 | 0.385 | 0.674 |
| | Time (ms) $\beta = 5$ | 0.139 | 0.201 | 0.271 | 0.402 | 0.688 |
| | Time (ms) $\beta = 8$ | 0.145 | 0.186 | 0.270 | 0.387 | 0.691 |

### 3.4.3 Abnormality Detection with Industrial Data

In this subsection, performance results under industrial data are shown for the two proposed KLD based methods as well as the MA, MAD and median filters for comparison. Data of a power plant drum level is used where the data points were sampled every 0.1 seconds. The time series are shown in Figure 3.11, where an abnormal water level fluctuation occurred after the sample instance shown with the red dotted line. By using the estimations in (3.3), (3.4), and (3.6) with 5000 data samples under the normal condition, the drum level data under normal condition is modeled by GGD$(49.80, 88.18, 2.37)^{\dagger}$. The estimated GGD is compared with the histogram of the data shown in Figure 3.12. It can be observed that the distribution of the drum level data under the normal condition reasonably matches the GGD.

†In the K-S test, the test statistic value is 0.0259 for samples to fit the estimated GGD, while it is 0.0295 with Gaussian distribution model. The threshold of the K-S test is selected as 0.0276 with significance level of 0.001 [89]. If the significant level is set as 0.01, neither of the two models pass the K-S test. But data under the normal condition fits GGD better than GD, and the proposed scheme under GGD model has better performance.

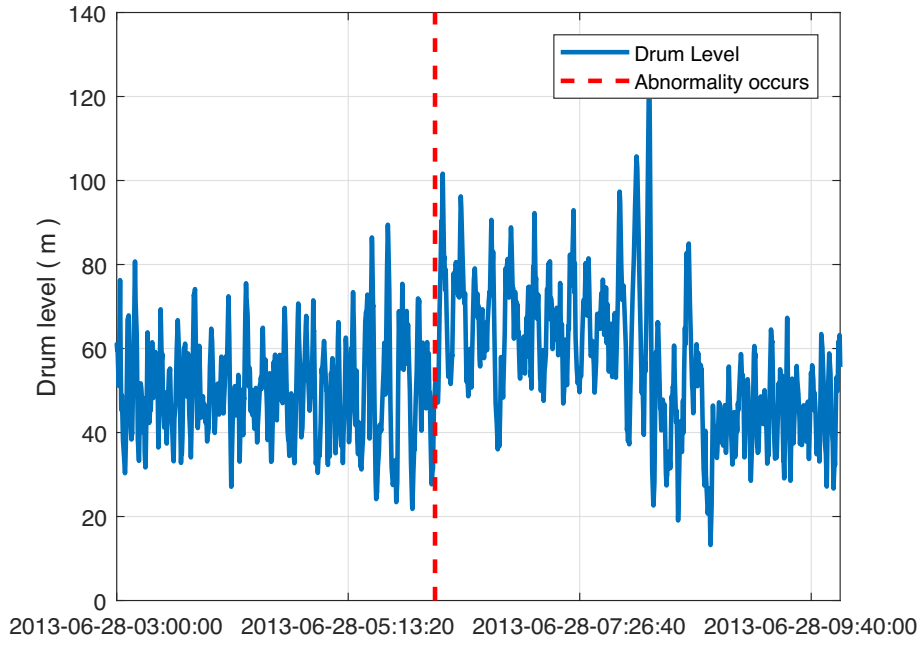



**Figure 3.11:** Industrial drum level data samples plot with sampling interval 0.1 second.

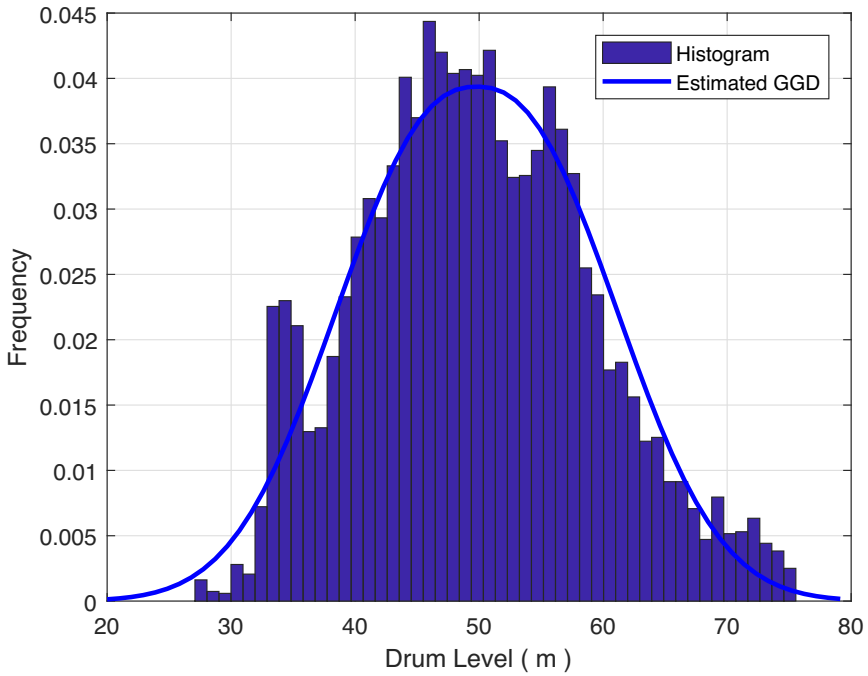



**Figure 3.12:** Estimated GGD for the drum level data compared with histogram.

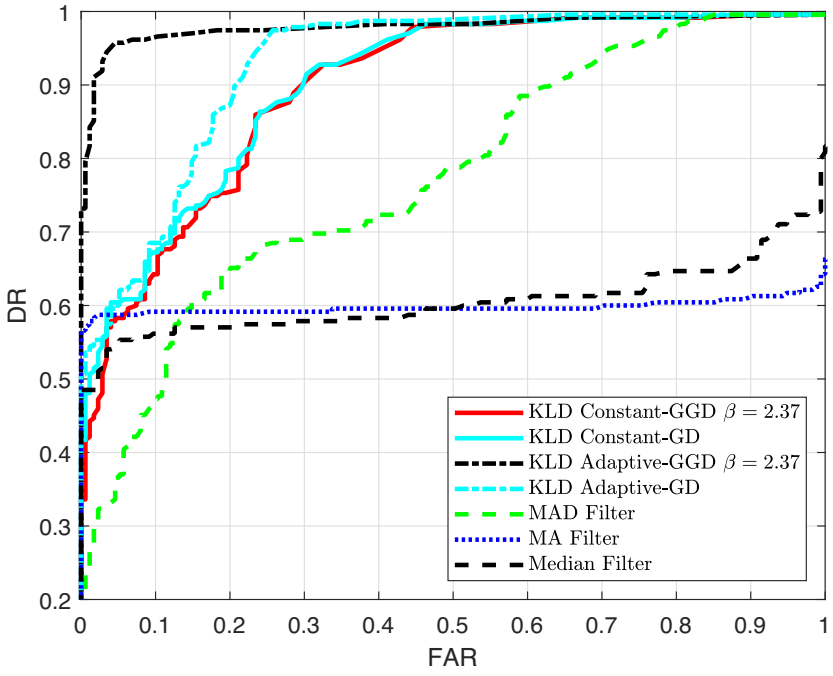



**Figure 3.13:** FAR v.s. DR with the proposed KLD based methods, the MAD, MA and median filters with industrial data.

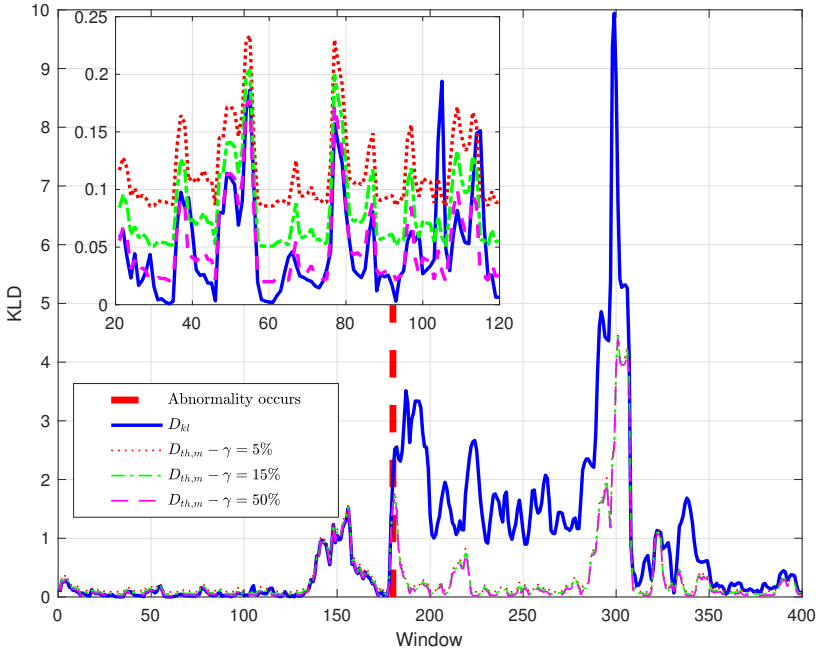



**Figure 3.14:** Monitoring statistics using the KLD based method with adaptive threshold.

By considering the slow change of the water level signal, in the simulation of the abnormality detection, the window size and the filter length is set as 600, i.e., $N = 600$, and a sliding window mechanism with the step size 60 is adopted to avoid long detection delay. Under this setting, the monitoring statistics of the KLD, the MA filter , the MAD filter, and the median filter can be calculated from (3.9), (3.67) , (3.68) and (3.69), respectively. Generated from 1000 pairs of FARs and DRs by moving the threshold of each monitoring statistics in full ranges, the FAR v.s. DR curves of different methods with both the GGD and Gaussian models are depicted in Figure 3.13. It can be first observed that the proposed KLD based methods outperform the three filters. Second, the adaptive threshold method with the GGD has better performance than other methods including that with the Gaussian model. As for the constant threshold scheme, even though the performance with the Gaussian model is slightly better than the GGD model, neither of the two can achieve good detection performance with the FAR of 0.05.

To better understand how the adaptive threshold method works, the statistic $D_{kl}$ in (3.9), and the adaptive thresholds $D_{th,m}$ with different $\gamma$ values are shown in Figure 3.14. To obtain the threshold, firstly, the distribution of $M^2$ is estimated with the kernel density estimation by using Gamma kernels. The smooth parameter is chosen as 0.02. Based on the estimated distribution of $M^2$, the distribution of $L_1$ is obtained via linear transformation. Thus, the adaptive thresholds can be calculated to meet the false alarm rate requirement. It can be observed that given a $\gamma$, the threshold values vary with respect to the data set and thus are adaptive to the current status of the system.

## 3.5 Summary

In this chapter, abnormality detection based on the KLD is studied for data with the GGD whose shape parameter is larger than 1. By adding the shape parameter to Gaussian distributions, the GGD covers more symmetric distributions such as Gaussian and uniform distributions. We first derived an

analytical formula of the KLD from the GGD estimated for current samples to that for the normal condition. Then, by considering the randomness in the sample mean and sample variance jointly and assuming there is no change in the shape parameter, a constant threshold scheme based on the KLD are derived with the help of the characteristic function and the central limit theorem. Further, an adaptive threshold method is proposed where the conditional KLD distributions are studied by separately considering the randomness in the sample mean and the sample variance while using the remaining parameter values as conditions.

In the simulation, firstly, the analytical expressions of the KLD are verified with the Monte Carlo simulations. It was shown that the relative difference between the KLD calculated by definition and the proposed formula is smaller than 5% for a large window size. So is the relative difference between the derived PDFs of KLD and those obtained from the histogram. Secondly, the proposed KLD based detection schemes are tested and compared with the MAD, MA and median filters with both simulated and industrial data. From the DR versus FAR curves, the KLD based detection scheme was shown to outperform the MAD, MA and median filters for both simulated and industrial cases.

# Chapter 4

# Abnormality Detection with the Rényi Divergence for Multivariate Gaussian Data*

By adding one more parameter, namely the divergence order, to the KLD, the Rényi divergence covers or relates to a family of divergences. It offers an additional degree of freedom for performance improvement. In this chapter, the distributions of the Rényi divergence under the normal condition and under off-set and scaling faults are derived analytically. Accordingly, the Rényi divergence based abnormality detection scheme with an adaptive divergence order is proposed for multivariate Gaussian distributed data.

## 4.1 Overview

In this chapter, we investigate the Rényi divergence based abnormality detection for the multivariate Gaussian distributed data. First, the singular value decomposition (SVD) is utilized to obtain uncorrelated processed data,

*A version of this chapter has been submitted as: Ying Xiong, Yindi Jing, Tongwen Chen, Abnormality detection with the Rényi divergence for multivariate Gaussian data, *IEEE Transactions on Signal Processing*, 2019. Part of this chapter has been accepted as: Ying Xiong, Yindi Jing, Tongwen Chen, Abnormality detection with Rényi divergence for univariate Gaussian data, *2019 IEEE Pacific Rim Conference on Communications, Computers and Signal Processing*, Victoria, Canada, 2019.

and the processed data samples are assumed to be i.i.d.. The Rényi divergence with the divergence order between 0 and 1 is adopted as detecting statistics. The reasons are twofold. First, this divergence order range still covers a family of divergences including the recommended divergence order 0.5 as per [22] and [111]. Second, the analytical results on the FAR and MAR with the divergence order greater than one has totally different formula comparing with those for the range between 0 and 1, causing even more complexity in the FAR and MAR calculations. Specifically, it leads to singularity in the divergence value, i.e., the Rényi divergence can approach to infinity when the divergence order is greater than 1. This occurs under both the normal and abnormal conditions, thus the calculation of the FAR and MAR become mathematically troublesome and non-trivial.

The remainder of this chapter is organized as follows. Section 4.2 illustrates the abnormality detection problem. The FAR and MAR are derived analytically in Section 4.3. The detection algorithm and discussions are explained in Section 4.4. Section 4.5 shows the case study results and Section 4.6 concludes this chapter.

## 4.2 Abnormality Detection Problem

Under the normal condition, without loss of generality as per [147], the $d$-dimensional measurement for each time instance is assumed to follow the multivariate Gaussian distribution with zero mean and covariance matrix $\mathbf{\Pi}$ denoted as $\mathcal{N}(\mathbf{0}, \mathbf{\Pi})$, and the measurements are time independent. We assume that $\mathbf{\Pi}$ is known, since it can be obtained via historical data under the normal condition. Let $\mathbf{\Pi} = \mathbf{P}\mathbf{\Lambda}\mathbf{P}^T$ be the SVD of $\mathbf{\Pi}$, where $\mathbf{P}$ is an orthogonal matrix and $\mathbf{\Lambda} = \text{diag}\{\sigma_1^2 \ \cdots \ \sigma_d^2\}$ is a diagonal matrix with non-negative entries. Let $\boldsymbol{x}_i = [x_{i1} \ \cdots \ x_{id}] \in \mathbb{R}^{1\times d}$ be the sample vector at time $i$ and $\mathbf{X} = \left[\boldsymbol{x}_1^T \ \cdots \ \boldsymbol{x}_N^T\right]^T \in \mathbb{R}^{N\times d}$ be the measured sample matrix containing $N$ sample vectors. By defining the processed signal matrix as $\mathbf{T} = \mathbf{X}\mathbf{P}$, we can conclude that the rows of $\mathbf{T}$, denoted as $\mathbf{t}_1, \cdots, \mathbf{t}_N$, are independent samples of

the multivariate Gaussian distribution with zero-mean and covariance matrix $\mathbf{\Lambda}$, i.e., $\mathcal{N}(\mathbf{0}, \mathbf{\Lambda})$.

### 4.2.1 Detection Problem Formulation

The sample mean denoted as $\hat{\boldsymbol{\mu}}$ and sample covariance denoted as $\hat{\mathbf{\Sigma}}$ with respect to the processed data can be calculated as follows:

$$\hat{\boldsymbol{\mu}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{t}_i; \tag{4.1}$$

$$\hat{\mathbf{\Sigma}} = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{t}_i - \hat{\boldsymbol{\mu}})(\mathbf{t}_i - \hat{\boldsymbol{\mu}})^T. \tag{4.2}$$

The abnormality detection problem can be expressed as follows

$$\begin{cases} H_0: & \boldsymbol{t}_i\text{'s follow the distribution } \mathcal{N}(\mathbf{0}, \mathbf{\Lambda}), \\ H_1: & \boldsymbol{t}_i\text{'s do not follow the distribution } \mathcal{N}(\mathbf{0}, \mathbf{\Lambda}). \end{cases} \tag{4.3}$$

The hypothesis test problem in (4.3) is to decide on whether the processed data set follows $\mathcal{N}(\mathbf{0}, \mathbf{\Lambda})$. A natural way is to fit the data with a multivariate Gaussian distribution with the sample mean and variance, i.e., $\mathcal{N}(\hat{\boldsymbol{\mu}}, \hat{\mathbf{\Sigma}})$, and check whether $\mathcal{N}(\hat{\boldsymbol{\mu}}, \hat{\mathbf{\Sigma}})$ is close enough to the known the distribution under the normal condition $(\mathbf{0}, \mathbf{\Lambda})$.

### 4.2.2 Rényi Divergence Based Detection Rule

In order to quantify the distance from one PDF to another, the Rényi divergence is adopted. According to [51], the Rényi divergence from $\mathcal{N}(\hat{\boldsymbol{\mu}}, \hat{\Sigma})$ to $\mathcal{N}(\mathbf{0}, \mathbf{\Lambda})$ of order $\alpha$ is defined as

$$D_\alpha\left[\mathcal{N}(\mathbf{0}, \mathbf{\Lambda})||\mathcal{N}(\hat{\boldsymbol{\mu}}, \hat{\mathbf{\Sigma}})\right] \triangleq \frac{1}{\alpha - 1}\ln\left[\int_{\mathbb{R}^d}\mathcal{N}(\mathbf{0}, \mathbf{\Lambda})^\alpha \mathcal{N}(\hat{\boldsymbol{\mu}}, \hat{\mathbf{\Sigma}})^{1-\alpha}\mathrm{d}x\right], \tag{4.4}$$

for $\alpha > 0$. The Rényi divergence covers a family of divergences with different $\alpha$ values. Specifically, when $\alpha$ approaches to 1, by taking the limit with L'Hopital rule, the Rényi divergence converges to the KLD. For the infinity order case, i.e., $\alpha \to \infty$, the Rényi divergence relates to the separation distance [9], which is used to bound the rate of convergence to the stationary distribution for

certain Markov chains. Additionally, the Rényi divergence becomes a function of the squared Hellinger distance [122] when $\alpha = 1/2$, and a function of the $\chi^2$-divergence for $\alpha = 2$. Moreover, the Rényi divergence is shown to relate with the total variation distance with the help of Pinsker's inequality [122] for $\alpha \in (0, 1)$.

When there is no confusion, we use the short notation $D_\alpha$ in the following content as the simplification for $D_\alpha \left[\mathcal{N}(\mathbf{0}, \mathbf{\Lambda})||\mathcal{N}(\hat{\boldsymbol{\mu}}, \hat{\mathbf{\Sigma}})\right]$. For multi-variant Gaussian PDFs, $D_\alpha$ can be calculated as [51]

$$D_\alpha = \begin{cases} \frac{\alpha}{2}\hat{\boldsymbol{\mu}}^T \mathbf{\Sigma}_\alpha^{-1} \hat{\boldsymbol{\mu}} - \frac{1}{2(\alpha-1)} \ln \frac{|\mathbf{\Sigma}_\alpha|}{|\hat{\mathbf{\Sigma}}|^\alpha |\mathbf{\Lambda}|^{1-\alpha}}, & \text{if } \mathbf{\Sigma}_\alpha \text{ is positive definite;} \\ +\infty, & \text{otherwise.} \end{cases} \tag{4.5}$$

where $\mathbf{\Sigma}_\alpha = \alpha\hat{\mathbf{\Sigma}} + (1-\alpha)\mathbf{\Lambda}$; $|\cdot|$ means the determinant of a matrix. Thus, the detection rule based on the Rényi divergence of divergence order $\alpha$ for the hypothesis test in (4.3) is expressed as follows:

$$\begin{cases} \text{Decide on } H_0 \text{ when } D_\alpha \leq D_{\alpha,th}; \\ \text{Decide on } H_1 \text{ when } D_\alpha > D_{\alpha,th}, \end{cases} \tag{4.6}$$

where $D_{\alpha,th}$ is the threshold for the detection.

In this chapter, we consider that $\alpha \in (0, 1)$ to avoid the singular case where the divergence value becomes infinity, i.e., the second case in (4.5). To include the $\alpha > 1$ range, the singular case needs to be considered separately causing more complexity in the performance analysis. Intuitively, when $D_\alpha = \infty$, the decision will be $H_1$ always, thus the false alarm probability and the probability of detection both will increase by the probability when $\Sigma_\alpha$ is not positive definite.

## 4.3 Performance Analysis

In this section, the FAR and MAR are derived analytically in the first and second subsections, respectively.

### 4.3.1 FAR Analysis

Given the processed data matrix $\mathbf{T} = \{\boldsymbol{t}_i\}_{i=1}^N$ under the normal condition, when $N$ is large, the sample covariance matrix is highly diagonal dominant since entries of $\boldsymbol{t}_i$'s are independent. It is reasonable to approximate the sample covariance matrix in (4.33) as a diagonal matrix, i.e., $\hat{\boldsymbol{\Sigma}} \approx \hat{\boldsymbol{\Sigma}}_{diag} = \text{diag}\{\hat{\sigma}_1^2 \ \cdots \ \hat{\sigma}_d^2\}$, where $\hat{\sigma}_j^2$ is the $j$th diagonal entry of $\hat{\boldsymbol{\Sigma}}$. It can be seen that the condition $\min_j\{\alpha\hat{\sigma_j}^2 + (1-\alpha)\sigma_j^2\} > 0$ holds true for $\alpha \in (0,1)$. Thus, from (4.5), the Rényi divergence can be calculated by

$$D_\alpha \approx \sum_{j=1}^{d} \left[ \frac{\alpha}{2} \frac{\hat{\mu}_j^2}{\alpha\hat{\sigma}_j^2 + (1-\alpha)\sigma_j^2} - \frac{1}{2(\alpha-1)} \ln \frac{\alpha\hat{\sigma_j}^2 + (1-\alpha)\sigma_j^2}{\sigma_j^{2-2\alpha}\hat{\sigma}_j^{2\alpha}} \right]. \tag{4.7}$$

By defining

$$S_j \triangleq \hat{\sigma}_j^2/\sigma_j^2 \quad \text{and} \quad M_j \triangleq \sqrt{N}\hat{\mu}_j/\sigma_j,$$

which are the normalized sample variance and the scaled mean, for $\alpha \in (0,1)$, $D_\alpha$ calculated in (4.7) can be further expressed as

$$D_\alpha \approx \frac{1}{2}\sum_{j=1}^{d} \left( \ln S_j + \frac{1}{\alpha-1} \ln \frac{S_j}{\alpha S_j + 1 - \alpha} \right) + \frac{1}{2}\sum_{j=1}^{d} \frac{\alpha M_j^2/N}{\alpha S_j + (1-\alpha)}. \tag{4.8}$$

To calculate the FAR, the distributions $S_j$ and $M_j$ for large $N$ under the normal condition are derived in Lemma 4.3.1 as per [133], followed by the FAR result in Theorem 4.3.2 for large $N$.

**Lemma 4.3.1.** Under the normal condition where the processed vector $\mathbf{t}_i$ follows $\mathcal{N}(\mathbf{0}, \boldsymbol{\Lambda})$, we have

$$M_j \xrightarrow[N\to\infty]{d} \mathcal{N}(0,1); \tag{4.9}$$

$$\sqrt{N}(S_j - 1) \xrightarrow[N\to\infty]{d} \mathcal{N}(0,2); \tag{4.10}$$

$$\text{Cov}(M_j, \sqrt{N}S_j) = 0, \tag{4.11}$$

where the notation $\xrightarrow[N\to\infty]{d}$ represents convergence in distribution when $N \to \infty$ and $\text{Cov(A, B)}$ is the covariance between variables $A$ and $B$.

**Theorem 4.3.2.** For $\alpha \in (0,1)$, the FAR denoted by $\mathrm{P}_F$ for the detection rule in (4.6) can be approximated as follows for large $N$:

$$\mathrm{P}_F = \frac{1}{\Gamma(d)}\Gamma\left(d, \frac{ND_{th}}{\alpha}\right), \tag{4.12}$$

where $\Gamma(s,x)$ is the upper incomplete gamma function.

*Proof.* According to (4.8), by defining $R = 2ND_\alpha/\alpha$,

$$A(S_j) = \frac{\ln S_j}{2} + \frac{1}{2(\alpha-1)}\ln\frac{S_j}{\alpha S_j + (1-\alpha)}, \tag{4.13}$$

$$B(S_j) = \frac{\alpha}{2\alpha S_j + 2(1-\alpha)}, \tag{4.14}$$

we have

$$\begin{aligned} R &= \frac{1}{\alpha}\sum_{j=1}^{d}\left(N\ln S_j + \frac{N}{\alpha-1}\ln\frac{S_j}{\alpha S_j + 1 - \alpha}\right) + \sum_{j=1}^{d}\frac{M_j^2}{\alpha S_j + (1-\alpha)} \\ &= \sum_{j=1}^{d}\left[2NA(S_j)/\alpha + 2M_j^2 B(S_j)/\alpha\right]. \end{aligned} \tag{4.15}$$

In order to get the PDF of $R$, firstly, we study the distribution of $R_j = 2NA(S_j)/\alpha + 2M_j^2 B(S_j)/\alpha$. The key is to get its characteristic function $\psi_{R_j}(t)$ as

$$\begin{aligned} \psi_{R_j}(t) &= \mathbb{E}(e^{it[2NA(S_j)/\alpha + 2M_j^2 B(S_j)/\alpha]}) \\ &= \mathbb{E}_{S_j}\left[\mathbb{E}_{M_j}\left(e^{it\left[2NA(S_j)/\alpha + 2M_j^2 B(S_j)/\alpha\right]} \mid S_i\right)\right]. \end{aligned} \tag{4.16}$$

As is shown in Lemma 4.3.1, the correlation between $M$ and $S$ is zero. So (4.16) can be further calculated as

$$\psi_{R_j}(t) = \mathbb{E}_{S_j}\left[e^{2itNA(S_j)/\alpha}\,\mathbb{E}_{M_j}\left(e^{itM_j^2 B(S_j)/\alpha} \mid S_j\right)\right]. \tag{4.17}$$

Since $M_j$ converges in distribution to $\mathcal{N}(0,1)$ when $N$ is large, it is easy to conclude that $M_j^2$ follows chi-square distribution of one degree of freedom for $N \to \infty$. Thus, the characteristic function of $M_j^2$ can be calculated as

$$\psi_{M_j^2}(t) \approx (1-2it)^{-\frac{1}{2}}. \tag{4.18}$$

By using the linear operation rule of the characteristic function, the characteristic function of $M_j^2 B(S_j)/\alpha$ can be obtained as

$$\psi_{M_j^2 B(S_j)}(t) = \mathbb{E}_{M_j}\left(e^{itM_j^2 B(S_j)/\alpha} \mid S_j\right) \approx \left[1 - 2itB(S_j)/\alpha\right]^{-\frac{1}{2}}. \tag{4.19}$$

Thus, by substituting (4.17) into (4.19), we have

$$\psi_{L_i}(t) \approx \int_{-\infty}^{+\infty} e^{2itNA(S_i)\alpha} \left[1 - 2itB(S_i)/\alpha\right]^{-\frac{1}{2}} f_{S_i}(S_i)\mathrm{d}S_i, \tag{4.20}$$

where $f_S(S)$ is the PDF of the variable $S$. In order to solve the integral in (4.20), the Taylor expansion is performed around $S_j = 1$ for $2NA(S_j)/\alpha$ and $\left[1 - 2itB(S_j)/\alpha\right]^{-\frac{1}{2}}$ respectively as follows:

$$2NA(S_j)/\alpha = \frac{N}{2}(S_j - 1)^2 + o\left[(S_j - 1)^2\right]; \tag{4.21}$$

$$\left[1 - 2itB(S_j)/\alpha\right]^{-\frac{1}{2}} = (1 - 2it)^{-\frac{1}{2}} - \frac{1}{2}it\alpha(1 - 2it)^{-\frac{3}{2}}(S_j - 1) + o(S_j - 1). \tag{4.22}$$

Since the higher order term approaches to zero when $N$ is large, by plugging (4.22) into (4.20) and denoting the characteristic function of $2NA(S_j)/\alpha$ as $\psi_{A(S_j)}(t)$, we have

$$\psi_{R_j}(t) \approx (1 - 2it)^{-\frac{1}{2}} \int_{-\infty}^{+\infty} e^{2itNA(S_j)/\alpha} f_{S_j}(S_j)\mathrm{d}S_j \tag{4.23}$$

$$= (1 - 2it)^{-\frac{1}{2}} \psi_{A(S_j)}(t). \tag{4.24}$$

Recall Lemma 4.3.1, since $\sqrt{N}(S_j - 1)/\sqrt{2} \xrightarrow[N\to\infty]{d} \mathcal{N}(0, 1)$, $A(S_j)$ expressed in (4.21) approximately follows the chi-square distribution of one degree of freedom when $N$ is large. Thus, we have $\psi_{A(S_j)}(t) \approx (1 - 2it)^{-\frac{1}{2}}$ and $\psi_{R_j}(t) \approx (1 - 2it)^{-1}$, which is the characteristic function of chi-square distribution of two degrees of freedom. So the results in Theorem 4.3.2 is derived. □

### 4.3.2 MAR Analysis

We consider two types of frequently discussed faults, namely, the offset faults causing mean changes only and the scaling faults causing variance changes only.

**Under Off-set Faults**

The off-set faults include the constant bias faults that bring up time-invariant mean change, and the drift faults which cause time-varying mean shifts. Once the off-set fault occurs, the distribution of the sample vectors becomes $\mathcal{N}(\mathbf{\Delta}, \mathbf{\Pi})$ and the distribution of the processed sample vectors becomes $\mathcal{N}(\mathbf{\Delta P}, \mathbf{\Lambda})$.

As the variance under the abnormal condition still have the same diagonal structure, the Rényi divergence calculation in (4.7) still holds under the constant bias fault. We have the following theorem for the distribution of the Rényi divergence of order $\alpha \in (0, 1)$ under off-set faults.

**Theorem 4.3.3.** For $\alpha \in (0, 1)$, the MAR denoted by $\mathrm{P}_{M1}$ under off-set faults for the detection rule in (4.6) with threshold $D_{\alpha,th}$ can be approximated as follows for large $N$:

$$\mathrm{P}_{M1} = 1 - Q_d\left(\sqrt{N\|\mathbf{\Delta\Lambda}^{-1}\mathbf{P}\|_2^2}, \sqrt{ND_{\alpha,th}/\alpha}\right), \tag{4.25}$$

where $Q_M(a, b)$ is the Marcum Q-function.

*Proof.* In this case, the $i$th processed row vector is denoted as $\boldsymbol{t}_{f,i}$, which can be calculated as

$$\boldsymbol{t}_{f,i} = \boldsymbol{t}_i + \mathbf{\Delta P}. \tag{4.26}$$

Thus, the distribution of the processed sample vectors becomes $\mathcal{N}(\mathbf{\Delta P}, \mathbf{\Lambda})$. Recall the definition of $M_j$ under the normal condition, i.e., $M_j \triangleq \sqrt{N}\hat{\mu}_j/\sigma_j$, by denoting the $j$th column of $\mathbf{P}$ as $\boldsymbol{p}_j$, under the abnormal condition, $M_j$ can be calculated as follows,

$$M_j = \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\frac{t_{ij}}{\sigma_j} + \frac{\sqrt{N}\mathbf{\Delta}\boldsymbol{p}_j}{\sigma_j}. \tag{4.27}$$

With the help of the central limit theorem, $M_j$ converges in distribution to $\mathcal{N}(\sqrt{N}\mathbf{\Delta}\boldsymbol{p}_j/\sigma_j, 1)$ for large $N$, i.e., $M_{f,j} \xrightarrow[N\to\infty]{d} \mathcal{N}(\sqrt{N}\mathbf{\Delta}\boldsymbol{p}_j/\sigma_j, 1)$. Thus, $M_j^2$

follows a non-centralized chi-square distribution with the non-central parameter $N(\boldsymbol{\Delta p}_j)^2/\sigma_j^2$, and the characteristic function of $M_j^2$ can be calculated as

$$\psi_{M_j}(t) \approx \exp\left[\frac{itN\left(\boldsymbol{\Delta p}_j\right)^2}{\sigma_j^2(1-2it)}\right](1-2it)^{-\frac{1}{2}}. \tag{4.28}$$

By following the same steps in the proof of Theorem 4.3.2, the characteristic function of $R_j$ can be derived as

$$\psi_{R_j}(t) \approx \exp\left[\frac{itN\left(\boldsymbol{\Delta p}_j\right)^2}{\sigma_j^2(1-2it)}\right](1-2it)^{-1}. \tag{4.29}$$

Since the variables in $\boldsymbol{t}_i$'s are uncorrelated and follow Gaussian distribution, we can see that the $d$ variables are mutually independent. Thus, the characteristic function of $R$ can be obtained by

$$\begin{aligned}\psi_R(t) &\approx \exp\left[\frac{itN\sum_{j=1}^{d}\left(\boldsymbol{\Delta p}_j\right)^2}{\sigma_j^2(1-2it)}\right](1-2it)^{-d} \\ &= \exp\left[\frac{itN\|\boldsymbol{\Delta\Lambda}^{-1}\mathbf{P}\|_2}{1-2it}\right](1-2it)^{-d}.\end{aligned} \tag{4.30}$$

So the results in Theorem 4.3.3 can be derived according to the property of the non-centralized chi-square distribution. □

**Under Scaling Faults**

Different from the off-set faults, the scaling faults change the variance of the measurement but keeps the mean, such as the well-know precision degradation and multiplicative faults in sensors. When this type of faults occurs, the distribution of the sample vectors becomes $\mathcal{N}(\mathbf{0}, \mathbf{K\Pi K})$, where $\mathbf{K} = \mathrm{diag}(\kappa_1, \cdots, \kappa_j, \cdots, \kappa_d)$. The distribution of the processed sample vector is thus $\mathcal{N}(\mathbf{0}, \mathbf{P}^T\mathbf{K\Pi K P})$, which is the same as $\mathcal{N}(\mathbf{0}, (\mathbf{P}^T\mathbf{K P})\boldsymbol{\Lambda}(\mathbf{P}^T\mathbf{K P}))$.

Notice that in this case, $\mathbf{P}^T\mathbf{K P}$ is not a diagonal matrix in general, thus $\hat{\boldsymbol{\Sigma}}$ is not diagonal dominant even for large $N$ and the Rényi divergence calculation in (4.7) no longer holds. The first term in (4.5) is related to the mean change,

thus it approaches to zero when $N$ is large. With this, the Rényi divergence under scaling faults can be approximated as

$$D_\alpha \approx -\frac{1}{2(\alpha-1)} \ln \frac{|\alpha\hat{\mathbf{\Sigma}} + (1-\alpha)\mathbf{\Lambda}|}{|\hat{\mathbf{\Sigma}}|^\alpha |\mathbf{\Lambda}|^{1-\alpha}}. \tag{4.31}$$

To help the calculations and analysis on (4.31), we further define $\tilde{\mathbf{t}}_i = \mathbf{t}_i(\mathbf{P}^T\mathbf{K}\mathbf{P})^{-1} = \mathbf{t}_i(\mathbf{P}^T\mathbf{K}^{-1}\mathbf{P})$, and $\tilde{\mathbf{t}}_i$'s are samples of $\mathcal{N}(\mathbf{0}, \mathbf{\Lambda})$. Define

$$\hat{\tilde{\boldsymbol{\mu}}} = \frac{1}{N}\sum_{i=1}^{N} \tilde{\mathbf{t}}_i; \tag{4.32}$$

$$\hat{\tilde{\mathbf{\Sigma}}} = \frac{1}{N}\sum_{i=1}^{N} (\tilde{\mathbf{t}}_i - \hat{\tilde{\boldsymbol{\mu}}})(\tilde{\mathbf{t}}_i - \hat{\tilde{\boldsymbol{\mu}}})^T. \tag{4.33}$$

We have that $\hat{\boldsymbol{\mu}} = \hat{\tilde{\boldsymbol{\mu}}}(\mathbf{P}^T\mathbf{K}\mathbf{P})$ and $\hat{\mathbf{\Sigma}} = (\mathbf{P}^T\mathbf{K}\mathbf{P})\hat{\tilde{\mathbf{\Sigma}}}(\mathbf{P}^T\mathbf{K}\mathbf{P})$. For large $N$, $\hat{\tilde{\mathbf{\Sigma}}}$ is diagonal dominant, from which we have $\hat{\tilde{\mathbf{\Sigma}}} \approx \text{diag}\{\hat{\tilde{\mathbf{\Sigma}}}\} = \text{diag}\{\hat{\tilde{\sigma}}_1^2 \ \cdots \ \hat{\tilde{\sigma}}_d^2\}$ and $\hat{\mathbf{\Sigma}} \approx (\mathbf{P}^T\mathbf{K}\mathbf{P})\text{diag}\{\hat{\tilde{\mathbf{\Sigma}}}\}(\mathbf{P}^T\mathbf{K}\mathbf{P})$. By using this approximation in (4.31), we have that

$$\begin{aligned}
D_\alpha &\approx -\frac{1}{2(\alpha-1)} \ln \frac{|\alpha(\mathbf{P}^T\mathbf{K}\mathbf{P})\text{diag}\{\hat{\tilde{\mathbf{\Sigma}}}\}(\mathbf{P}^T\mathbf{K}\mathbf{P}) + (1-\alpha)\mathbf{\Lambda}|}{|\mathbf{K}|^{2\alpha}|\text{diag}\{\hat{\tilde{\mathbf{\Sigma}}}\}|^\alpha|\mathbf{\Lambda}|^{1-\alpha}} \\
&= -\frac{1}{2(\alpha-1)} \ln \frac{|\mathbf{K}|^2 \cdot |\alpha\text{diag}\{\hat{\tilde{\mathbf{\Sigma}}}\} + (1-\alpha)(\mathbf{P}^T\mathbf{K}^{-1}\mathbf{P})\mathbf{\Lambda}(\mathbf{P}^T\mathbf{K}^{-1}\mathbf{P})|}{|\mathbf{K}|^{2\alpha}|\text{diag}\{\hat{\tilde{\mathbf{\Sigma}}}\}|^\alpha|\mathbf{\Lambda}|^{1-\alpha}} \\
&= -\frac{1}{2(\alpha-1)} \ln \frac{|\mathbf{K}|^{2(1-\alpha)} \cdot |\alpha\text{diag}\{\hat{\tilde{\mathbf{\Sigma}}}\} + (1-\alpha)\tilde{\mathbf{\Pi}}|}{|\text{diag}\{\hat{\tilde{\mathbf{\Sigma}}}\}|^\alpha|\mathbf{\Lambda}|^{1-\alpha}},
\end{aligned}$$

where $\tilde{\mathbf{\Pi}} \triangleq (\mathbf{P}^T\mathbf{K}^{-1}\mathbf{P})\mathbf{\Lambda}(\mathbf{P}^T\mathbf{K}^{-1}\mathbf{P}) = \mathbf{P}^T\mathbf{K}^{-1}\mathbf{\Pi}\ \mathbf{K}^{-1}\mathbf{P}$.

When $\alpha \in (0,1)$, we have $\alpha\text{diag}\{\hat{\tilde{\mathbf{\Sigma}}}\} + (1-\alpha)\tilde{\mathbf{\Pi}} > 0$, the Hadamard's inequality can be applied in the above expression to obtain

$$\begin{aligned}
D_\alpha \le D_\alpha^{up} &\triangleq -\frac{1}{2(\alpha-1)} \ln \frac{\prod_{j=1}^d \kappa_j^{2(1-\alpha)} \prod_{j=1}^d \left[\alpha\hat{\tilde{\sigma}}_j^2 + (1-\alpha)\tilde{\pi}_{jj}\right]}{\prod_{j=1}^d \hat{\tilde{\sigma}}_j^{2\alpha}\sigma_j^{2(1-\alpha)}} \\
&= \frac{1}{2(\alpha-1)} \sum_{i=1}^{d} \left[\ln \kappa_j^{2(\alpha-1)} + \alpha \ln \frac{\hat{\tilde{\sigma}}_j^2}{\sigma_j^2} - \ln \frac{\alpha\hat{\tilde{\sigma}}_j^2 + (1-\alpha)\tilde{\pi}_{jj}}{\sigma_j^2}\right],
\end{aligned}$$

where $\tilde{\pi}_{jj}$ is the $(j,j)$th entry of $\tilde{\mathbf{\Pi}}$.

With slight abuse of notation, re-define $S_j \triangleq \hat{\tilde{\sigma}}_j^2/\sigma_j^2$. As per Lemma 4.3.1, we have that $\sqrt{N}(S_j-1)$ approximately follows Gaussian distribution $\mathcal{N}(0,2)$ when $N$ is large, i.e., for large $N$, $\sqrt{N}(S_j-1)/\sqrt{2} \sim \mathcal{N}(0,1)$. The upper bound of the Rényi divergence can be calculated as

$$D_\alpha^{up} = \frac{1}{2(\alpha-1)} \sum_{j=1}^{d} \left\{ \ln \kappa_j^{2\alpha-2} + \alpha \ln S_j - \ln \left[ \alpha S_j + (1-\alpha)\tilde{\pi}_{jj}/\sigma_j^2 \right] \right\}.$$

With the help of Taylor series around $S_j = 1$, we have

$$D_\alpha^{up} = \sum_{j=1}^{d} \left[ a_j \left[ \frac{\sqrt{N}(S_j-1)}{\sqrt{2}} \right]^2 + b_j \frac{\sqrt{N}(S_j-1)}{\sqrt{2}} + c_j \right] + o\left[ (S_j-1)^2 \right], \tag{4.34}$$

where we have made the following definitions:

$$a_j \triangleq \frac{\alpha}{2N(1-\alpha)} \left\{ 1 - \frac{\alpha}{\left[ \frac{\tilde{\pi}_{jj}}{\sigma_j^2} + \alpha \left( 1 - \frac{\tilde{\pi}_{jj}}{\sigma_j^2} \right) \right]^2} \right\}, \tag{4.35}$$

$$b_j \triangleq \frac{\alpha}{\sqrt{2N}} \frac{1 - \frac{\tilde{\pi}_{jj}}{\sigma_j^2}}{\frac{\tilde{\pi}_{jj}}{\sigma_j^2} + \alpha \left( 1 - \frac{\tilde{\pi}_{jj}}{\sigma_j^2} \right)}, \tag{4.36}$$

$$c_j \triangleq \frac{1}{2(1-\alpha)} \ln \kappa_j^{2(1-\alpha)} + \frac{1}{2(1-\alpha)} \ln \left[ \frac{\tilde{\pi}_{jj}}{\sigma_j^2} + \alpha \left( 1 - \frac{\tilde{\pi}_{jj}}{\sigma_j^2} \right) \right]. \tag{4.37}$$

Let $\mathcal{S}_0 = \{j | a_j = 0\}$, $\mathcal{S}_+ = \{j | a_j > 0\}$ and $\mathcal{S}_- = \{j | a_j < 0\}$. $D_\alpha^{up}$ can be rewritten as

$$D_\alpha^{up} \approx \sum_{j \in \mathcal{S}_0} \left[ b_j \frac{\sqrt{N}(S_j-1)}{\sqrt{2}} \right] + \underbrace{\sum_{j \in \mathcal{S}_+} \left\{ a_j \left[ \frac{\sqrt{N}(S_j-1)}{\sqrt{2}} + \frac{b_j}{2a_j} \right]^2 \right\}}_{D_1}$$
$$\underbrace{- \sum_{j \in \mathcal{S}_-} \left\{ -a_j \left[ \frac{\sqrt{N}(S_j-1)}{\sqrt{2}} + \frac{b_j}{2a_j} \right]^2 \right\}}_{D_2} + C, \tag{4.38}$$

where

$$C \triangleq \sum_{j=1}^{d} c_j - \frac{b_j^2}{4a_j}. \tag{4.39}$$

Since when $N$ is large, $\sqrt{\frac{N}{2}}(S_j - 1) \sim \mathcal{N}(0,1)$ and $\sqrt{\frac{N}{2}}(S_j - 1) + \frac{b_j}{2a_j} \sim \mathcal{N}\left(\frac{b_j}{2a_j}, 1\right)$, we have that $\left[\frac{\sqrt{N}(S_j-1)}{\sqrt{2}} + \frac{b_j}{2a_j}\right]^2$ follows non-central Chi-square distribution with one degree of freedom and non-centrality parameter $\frac{b_j^2}{4a_j^2}$, i.e., $\chi^2\left(1, \frac{b_j^2}{4a_j^2}\right)$. Since the terms in $D_1$ and $D_2$ are mutually independent for different $j$, we have that $D_1$ and $D_2$ are linear combinations of independent non-central Chi-square random variables. Further, $D_\alpha^{up}$ can be regarded as a linear combination of Gaussian, positive and negative weighted non-central Chi-square variables, but the distribution of $D_\alpha^{up}$ is intractable and remains an open problem. To avoid the cases when Gaussian variables are included, i.e., $\mathcal{S}_0 \neq \varnothing$, a subset of $(0,1)$ is considered as the new range for $\alpha$. Specifically, it should be noted that when Gaussian variables are provoked, i.e., at least one $\alpha$ satisfies $a_j(\alpha) = 0$, those $\alpha$ values can be obtained as a function of $\tilde{\pi}_{jj}$ for given $N$ and $\sigma_j$ according to (4.35). By denoting those $\alpha$ values as $\mathcal{A}_{=0} = \{\alpha | a_j(\alpha) = 0\}$, the new set of $\alpha$ can be expressed as $\mathcal{A}_{\neq 0} = \{\alpha | \alpha \in (0,1), \alpha \notin \mathcal{A}_{=0}\}$, which is considered as the new $\alpha$ range. For $\alpha \in \mathcal{A}_{\neq 0}$, $D_\alpha^{up}$ can be further expressed as $D_\alpha^{up} = D_1 - D_2 + C$. Via Lemma 2 in [78], by setting

$$\lambda_{*1} = \frac{\sum_{j\in\mathcal{S}_+}(2a_j^2 + b_j^2)}{\sum_{j\in\mathcal{S}_+}\left(2a_j + \frac{b_j^2}{a_j}\right)}, \quad \lambda_{*2} = -\frac{\sum_{j\in\mathcal{S}_-}(2a_j^2 + b_j^2)}{\sum_{j\in\mathcal{S}_-}\left(2a_j + \frac{b_j^2}{a_j}\right)}, \tag{4.40}$$

we have $Y_1 \triangleq D_1/\lambda_{*1} \sim \chi^2(\nu_{*1}, w_{*1})$ and $Y_2 \triangleq D_2/\lambda_{*2} \sim \chi^2(\nu_{*2}, w_{*2})$ approximately, where

$$\nu_{*1} = \frac{(\sum_{j\in\mathcal{S}_+} a_j)\left[\sum_{j\in\mathcal{S}_+}\left(2a_j + \frac{b_j^2}{a_j}\right)\right]}{\sum_{j\in\mathcal{S}_+}(2a_j^2 + b_j^2)}, \tag{4.41}$$

$$w_{*1} = \frac{1}{4}\frac{\left(\sum_{j\in\mathcal{S}_+}\frac{b_j^2}{a_j}\right)\left[\sum_{j\in\mathcal{S}_+}\left(2a_j + \frac{b_j^2}{a_j}\right)\right]}{\sum_{j\in\mathcal{S}_+}(2a_j^2 + b_j^2)}, \tag{4.42}$$

$$\nu_{*2} = \frac{(\sum_{j\in\mathcal{S}_-} a_j)\left[\sum_{j\in\mathcal{S}_-}\left(2a_j + \frac{b_j^2}{a_j}\right)\right]}{\sum_{j\in\mathcal{S}_-}(2a_j^2 + b_j^2)}, \tag{4.43}$$

$$w_{*2} = \frac{1}{4}\frac{\left(\sum_{j\in\mathcal{S}_-}\frac{b_j^2}{a_j}\right)\left[\sum_{j\in\mathcal{S}_-}\left(2a_j + \frac{b_j^2}{a_j}\right)\right]}{\sum_{j\in\mathcal{S}_-}(2a_j^2 + b_j^2)}. \tag{4.44}$$

Thus, $D_\alpha^{up}$ can be written as $D_\alpha^{up} = \lambda_{*1} Y_1 - \lambda_{*2} Y_2 + C$. By denoting $P_{M2}$ as the MAR under scaling faults, $P_{M2}$ can be calculated as

$$\begin{aligned} P_{M2} =& \mathbb{P}\left(D_\alpha^{up} > D_{\alpha,th}\right) \\ =& \mathbb{P}\left(Y_1 > \frac{\lambda_{*2}}{\lambda_{*1}} Y_2 + \frac{\alpha D_{\alpha,th} - C}{\lambda_{*1}}\right) \\ =& \int_{-\infty}^{+\infty} \int_{\frac{\lambda_{*2}}{\lambda_{*1}} y_2 + \frac{\alpha D_{\alpha,th} - C}{\lambda_{*1}}}^{+\infty} f_{Y_1}(y_1) f_{Y_2}(y_2)\, \mathrm{d}y_1\, \mathrm{d}y_2. \end{aligned}$$

Depending on whether $D_{\alpha,th} - C$ is positive, we have that

$$P_{M2} = \begin{cases} \displaystyle\int_0^{+\infty} \left[1 - Q_{\frac{\nu_{*1}}{2}}\left(\sqrt{w_{*1}}, \sqrt{\frac{\lambda_{*2}}{\lambda_{*1}} y_2 + \frac{\alpha D_{\alpha,th} - C}{\lambda_{*1}}}\right)\right] f_{Y_2}(y_2)\, \mathrm{d}y_2, \\ \qquad\qquad \alpha D_{\alpha,th} - C > 0; \\ \displaystyle\int_{\frac{\alpha D_{\alpha,th} - C}{\lambda_{*1}}}^{+\infty} \left[1 - Q_{\frac{\nu_{*1}}{2}}\left(\sqrt{w_{*1}}, \sqrt{\frac{\lambda_{*2}}{\lambda_{*1}} y_2 + \frac{\alpha D_{\alpha,th} - C}{\lambda_{*1}}}\right)\right] f_{Y_2}(y_2)\, \mathrm{d}y_2, \\ \qquad\qquad \alpha D_{\alpha,th} - C \leq 0. \end{cases} \tag{4.45}$$

By plugging the distribution of $Y_2$, $P_{M2}$ can be further calculated as

$$\begin{aligned} P_{M2} = & \int_{\frac{(\alpha D_{\alpha,th} - C)^+}{\lambda_{*2}}}^{+\infty} \left[1 - \frac{1}{2} Q_{\frac{\nu_{*1}}{2}}\left(\sqrt{w_{*1}}, \sqrt{\frac{\lambda_{*2}}{\lambda_{*1}} y_2 + \frac{\alpha D_{\alpha,th} - C}{\lambda_{*1}}}\right)\right] \\ & \times e^{-\frac{y_2 + \nu_{*2}}{2}} \left(\frac{y_2}{\nu_{*2}}\right)^{w_{*2}/4 - 1/2} I_{w_{*2}/2 - 1}\left(\sqrt{\nu_{*2} y_2}\right) \mathrm{d}y_2, \end{aligned} \tag{4.46}$$

where $I_v(y)$ is a modified Bessel function of the first kind, and $(\alpha D_{\alpha,th} - C)^+ = \max\left(0, \alpha D_{\alpha,th} - C\right)$.

## 4.4 Proposed Detection Algorithm with Adaptive Divergence Order

In this section, we propose a detection algorithm considering the optimization of the divergence order $\alpha$ in terms of the FAR and MAR for multivariate Gaussian data.

Under off-set faults, with respect to the NP-testing problem, $\alpha$ value does not change the FAR and MAR. Specifically, it can be observed that $D_{\alpha,th}$ and $\alpha$ appear as a ratio in both the FAR expression in (4.12) and the MAR

expression in (4.25). When $\alpha$ changes, by adjusting $D_{\alpha,th}$, the same FAR and MAR can be obtained.

As for the scaling faults, the value of $\alpha$ has decoupled effects on the FAR and MAR, thus its optimization is useful. We can follow the NP-testing problem formulation to consider the joint optimization of $\alpha$ and $D_{\alpha,th}$. With the help of the derived analytical formulas on the FAR and the MAR in (4.12) and (4.46), the joint optimization problem can be formulated to optimize the MAR under scaling faults subject to the given FAR requirement $\gamma$, i.e.,

$$\begin{aligned} \min_{D_{\alpha,th},\alpha} \quad & \mathrm{P}_{M2} \\ \text{s.t.} \quad & \mathrm{P}_F \leq \gamma. \end{aligned} \tag{4.47}$$

The Lagrange method can be directly applied to solve this joint optimization problem with respect to $\alpha$ and $D_{\alpha,th}$. Alternatively, it can be transfered as a univariate unconstrained problem regarding to the divergence order $\alpha$ to relieve computational burdens. Specifically, for the case when the constraint is tight, i.e., $\mathrm{P}_F = \gamma$, from which $D_{\alpha,th}$ can be expressed as a function of $\alpha$. By replacing $D_{\alpha,th}$ with the function of $\alpha$, the optimization problem in (4.47) becomes univariate and unconstrained, which can be solved by exhaustive search. When the FAR constraint is not tight for given $\alpha$, the threshold can be shifted to tight the constraint and reduce the MAR.

In practice, it needs to be judged whether to trigger the optimization problem in (4.47), since it is set up under scaling faults where the variance change dominates the abnormal features. In order to obtain the preliminary knowledge about the occurrence of variance changes, the $k$-sigma rule is adopted and applied on $S_j = \hat{\sigma}_j^2/\sigma_j^2$, since $S_j$ conveys the variance change information in the observed data and follows Gaussian distributions under the normal condition as per Lemma 4.3.1. Specifically, if any $S_j$'s in $n$ consecutive windows satisfies that $\sqrt{N}|S_j - 1|/\sqrt{2} > k$, it is considered as the occurrence of variance change faults and the optimization problem in (4.47) is triggered. For other cases, $\alpha$ is selected as 0.001 considering that excessive false alarms are major problems in many practical alarming systems. Since the Rényi divergence is

a monotonic increase function with respect to $\alpha$, a smaller divergence order under the normal condition induces a lower FAR.

According to the discussions above, our proposed detection algorithm is summarized in Algorithm 4. There are two stages involved, namely, the off-line training stage where historical normal data is processed and the on-line detection stage in which observed data is used for abnormality detection in a moving window. In the off-line training stage, the mean and covariance matrices are firstly calculated from historical data by (4.32) and (4.33). The matrix $\mathbf{P}$ is derived by the SVD, and the processed signal matrix can be calculated as $\mathbf{T} = \mathbf{XP}$, whose dimension can be reduced by the given threshold of the percentage variance explained in $\mathbf{T}$. Specifically, by sorting the elements in $\mathbf{\Lambda} = \text{diag}\{\sigma_1^2 \ \cdots \ \sigma_d^2\}$ in a descending order, the explained percentage variance by using the first $d_m$ elements can be calculated as $\sum_{i=1}^{d_m} \sigma_i^2 / \sum_{i=1}^{d} \sigma_i^2$. For a given percentage variance threshold $V_{th}$, the reduced dimension $d_m$ can be determined as the minimum number of elements that satisfies $\sum_{i=1}^{d_m} \sigma_i^2 / \sum_{i=1}^{d} \sigma_i^2 \geq V_{th}$, i.e., $d_m = \min_{d_m} \left\{ d_m | \sum_{i=1}^{d_m} \sigma_i^2 / \sum_{i=1}^{d} \sigma_i^2 \geq V_{th} \right\}$. More details can be referred to [43]. In the on-line detection stage, for each window of observed data, the data matrix is first projected by the matrix $\mathbf{P}$, then the mean vector and covariance matrix of the estimated Gaussian distribution is obtained by (4.32) and (4.33). If the variance change is detected with the help of $k$-sigma rule, the fault coefficients $\kappa_j$'s are approximately estimated by matching the diagonal elements of $(\mathbf{P}^T\mathbf{KP})\mathbf{\Lambda}(\mathbf{P}^T\mathbf{KP})$ with the covariance matrix of the observed data $\hat{\mathbf{\Sigma}}$, which is equivalent to assuming the abnormal distribution has the same diagonal elements as the covariance matrix of observed data. Thus, we have that $\kappa_j = S_j(\mathbf{p}_j^T\mathbf{p}_j)^{-1}$, and the optimization problem in (4.47) is triggered to obtain $\alpha$ that minimizes the MAR. Otherwise, the constant 0.001 is assigned to $\alpha$. Based on the $\alpha$ value, the threshold $D_{\alpha,th}$ is calculated by setting $\text{P}_F = \gamma$, and then used to determine whether an alarm should be raised.

**Algorithm 4** Detection algorithm with adaptive divergence order and threshold

**Input**: Training data under normal condition and observed data for detection.
**Output**: Hypothesis test results for (4.3), i.e., alarm states.
**Initialization**: Assign values to $N$ and $\gamma$.
**Off-line training**: Estimate $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$ by (4.32) and (4.33);
Find the matrix $\mathbf{P}$ and $d_m$. **Close**
**On-line detection**:

**while** *the window is filled by N new samples* **do**
  Calculate the processed observed data matrix $\mathbf{T}$ with $\mathbf{T} = \mathbf{XP}$;
  Estimate $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$ of $\mathbf{T}$ by Eqs. (4.32) and (4.33);
  **if** *Variance changes* **then**
    $\alpha$ is obtained by solving (4.47);
  **else**
    $\alpha = 0.001$;
  **end**
  Calculate $D_\alpha$ by (4.5);
  Obtain $D_{\alpha,th}$ by solving $\mathrm{P}_F = \gamma$ with (4.12);
  **if** $D_\alpha > D_{\alpha,th}$ **then**
    **Return** Alarm state as 1; Save $\mathbf{X}$ and $D_\alpha$ for further analysis;
  **else**
    **Return** Alarm state as 0;
  **end**
**end**
Back to Step 1;

## 4.5 Case Studies

In this section, the asymptotic behaviors of the Rényi divergence are studied in the first subsection, while detection results are illustrated with one-dimensional and multi-dimensional data in the next two subsections, respectively.

### 4.5.1 Asymptotic Behavior Analysis

In this subsection, the asymptotic behavior the Rényi divergence under the normal condition and under the off-set and scaling faults is studied, respectively. In the following simulation, four cases with different dimensions

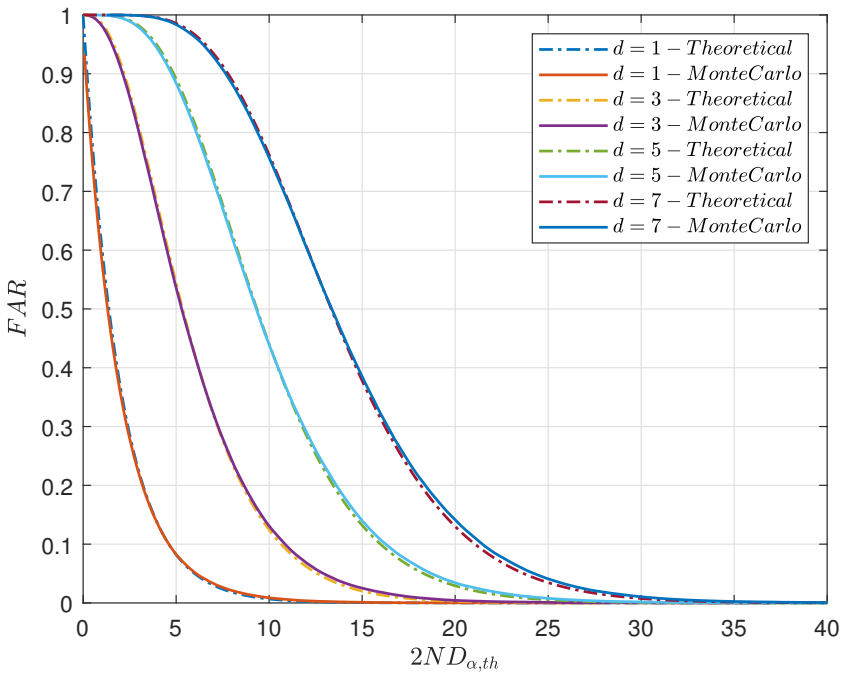



**Figure 4.1:** Comparison between the theoretical FAR result in (4.12) and the FAR obtained from the Monte Carlo simulations.

are considered, namely $d = 1, 3, 5, 7$. The divergence order $\alpha$ is fixed as 0.6 and the window size is set as 100.

First, we study the FAR calculation. The data matrix $\mathbf{X}$ is generated from the multivariate Gaussian distribution with zero mean and identical variance matrix, i.e., $\mathbf{P} = \boldsymbol{I}$, which is equivalent to assuming that the entries of $\boldsymbol{x}_i$'s are i.i.d. and follow $\mathcal{N}(0, 1)$. In the Monte Carlo simulations, there are 10000 Rényi divergence values calculated to obtain the FAR curve. The comparison between the theoretical FAR result calculated in (4.12) and that obtained from the Monte Carlo simulations is shown in Figure 4.1. From Figure 4.1, we can see that the theoretical FAR derived in (4.12) matches well with that obtained from the Monte Carlo simulations.

Second, the off-set fault is studied. In the simulation, the normal condition is set as the same as that under the FAR comparison, while the bias parameter vector is set as $\boldsymbol{\Delta}(j) = 0.1j,\ j = 1 \cdots d$, which means that the $j$th entry of $\boldsymbol{x}_i$'s are i.i.d. and follow $\mathcal{N}(0.1j, 1)$. In the Monte Carlo simulation, there are

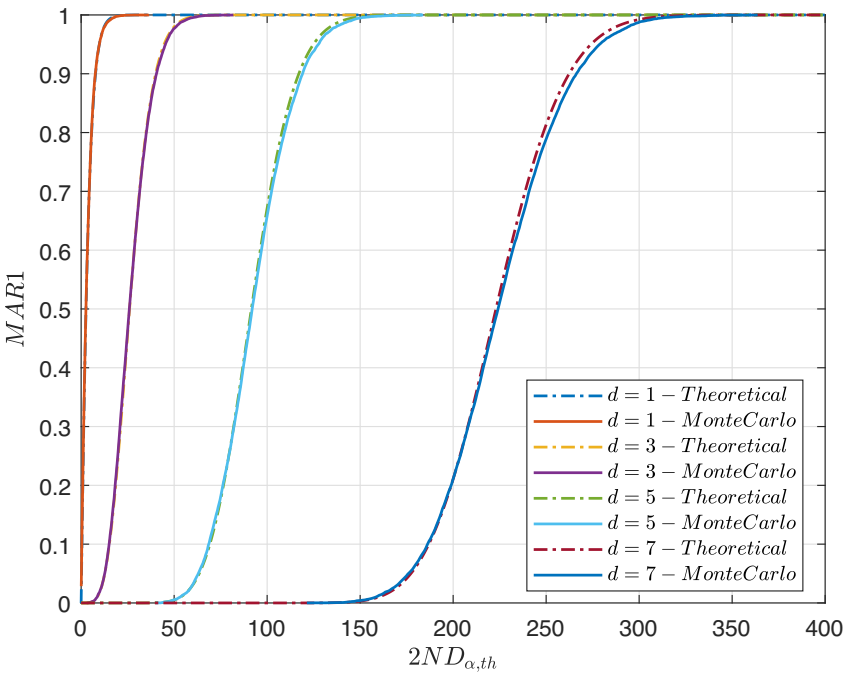



**Figure 4.2:** Comparison between the theoretical MAR result in (4.25) and the MAR obtained from the Monte Carlo simulations.

10000 windows of data generated to obtain the MAR curve. The comparison results between the theoretical MAR calculation in (4.25) and the Monte Carlo simulation results is depicted in Figure 4.2. It can be seen from Figure 4.2 that the theoretical MAR curve derived from (4.25) almost overlaps with that from the Monte Carlo simulation.

Third, the scaling fault is considered. The comparison between the theoretical MAR result in (4.46) and the MAR result obtained from the Monte Carlo simulations is shown in Figure 4.3. The fault parameter $\kappa_j$ in $\mathbf{K}$ is set as $\kappa_j = 0.6 + 0.15j,\ j = 1 \cdots d$. The multivariate Gaussian distribution $\mathcal{N}(\mathbf{0}, \mathbf{K})$ is used to generate 10000 windows of data to calculate the Rényi divergence. It can be seen that the theoretical calculation matches that from the Monte Carlo simulations but the error is greater than the two former cases. This phenomenon is possibly owing to the additional Taylor expansion applied for $D_\alpha^{up}$.

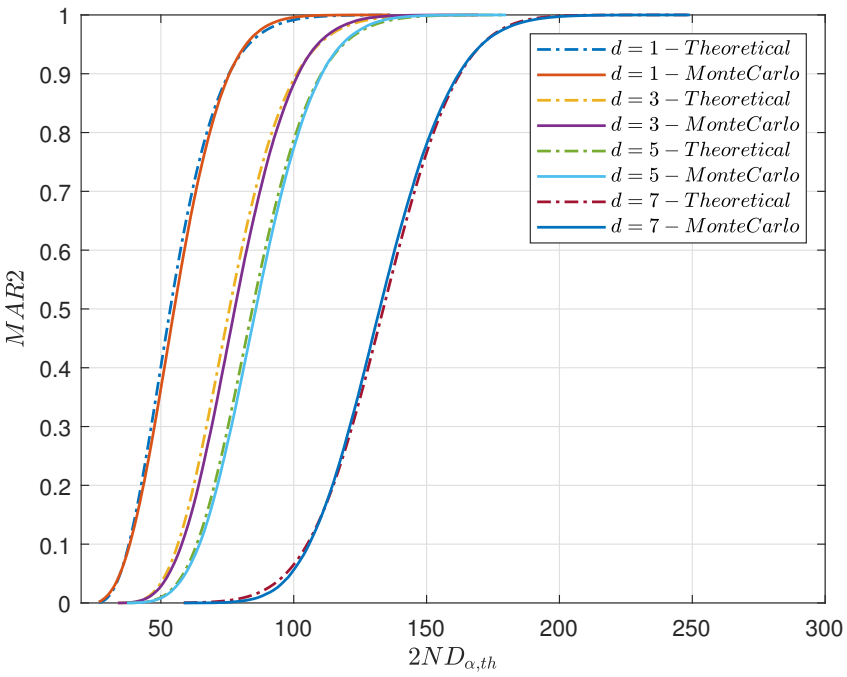



**Figure 4.3:** Comparison between the theoretical MAR result in (4.46) and the MAR obtained from the Monte Carlo simulations..

### 4.5.2 Detection with One-Dimensional Data

In this subsection, the proposed Algorithm 4 is applied to the one-dimensional data. In the simulation, data under the normal condition follows Gaussian distribution denoted as $\mathcal{N}(0, 1)$. The window size is chosen as 100, i.e., $N = 100$.

First, the optimization result of (4.47) under different $\kappa$ values is depicted with Figure 4.4, where the FAR is required to be less than 5%, i.e., $\gamma = 5\%$. It can be seen that when $\kappa \in (0.79, 1)$, the optimal $\alpha$ is an increasing function of $\kappa$ until $\kappa = 0.89$, then it keeps stable around 1. For $\kappa \in (1, 1.21)$, it approaches to 0, followed by an increasing trend to the value of 0.5 when $\kappa$ rises from 1.21 to around 1.63. When the variance change becomes significant, i.e., $\kappa \in (0, 0.79)$ or $\kappa \in (1.21, 2)$, the optimal $\alpha$ is 0.5.

Second, the detection performance of the proposed Rényi based method is compared with the KLD based method. In the one-dimensional case, the KLD between two Gaussian distributions $\mathcal{N}(\mu, \sigma^2)$ and $\mathcal{N}(\hat{\mu}, \hat{\sigma}^2)$ can be calculated

as per [144]:

$$D_{kl} = \ln \frac{\sigma}{\hat{\sigma}} + \frac{\hat{\sigma}^2 + (\mu - \hat{\mu})^2}{2\sigma^2} - \frac{1}{2}. \tag{4.48}$$

In this experiment, four abnormal cases where data follows the Gaussian distributions $\mathcal{N}(0.4, 1)$, $\mathcal{N}(0, 0.75^2)$, $\mathcal{N}(0, 1.25^2)$ and $\mathcal{N}(0.1, 0.8^2)$ are considered, which correspond to one constant bias fault, two multiplicative faults with $\kappa$ range of $(0, 0.79)$ and $(1.21, 2)$, and one combining the constant bias and multiplicative faults, respectively. The number of windows is 1000 under the normal condition and under each of the four abnormal conditions. In detecting the variance change, 2-sigma rule is applied in 2 consecutive windows, i.e., $k = n = 2$.

The FAR versus DR curves are shown in Figure 4.5. In Figure 4.5 (a), it shows the comparison under the off-set fault $\mathcal{N}(0.4, 1)$. We can see that the performance of the Rényi divergence and the KLD based method is exactly the same, which agrees with our analysis that different divergence orders do not change the FAR-MAR curve. The comprehensive performance under the two multiplicative faults, namely $\mathcal{N}(0, 0.75^2)$ and $\mathcal{N}(0, 1.25^2)$, are depicted in Figure 4.5 (b), while Figure 4.5 (c) shows the results under the combined fault $\mathcal{N}(0.1, 0.8^2)$. The overall performance under the above four faults are depicted in Figure 4.5 (d). We can see that under the latter three faults, the proposed algorithm achieves better DR than the KLD based method.

### 4.5.3 Detection with Multi-Dimensional Data

In this subsection, the detection performance of the proposed scheme in Algorithm 4 with multi-dimensional data is shown and compared with the KLD, the Hotelling's $T^2$ and the SPE statistics. The KLD for two multivariate Gaussian distributions denoted as $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\mathcal{N}(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$ can be expressed according to [144] as

$$D_{kl} = \frac{1}{2} \sum_{i=1}^{d} \left[ \ln \left( \frac{\sigma_i^2}{\hat{\sigma}_i^2} \right) + \frac{\hat{\sigma}_i^2}{\sigma_i^2} + \frac{(\mu_i - \hat{\mu}_i)^2}{\sigma_i^2} - 1 \right]. \tag{4.49}$$

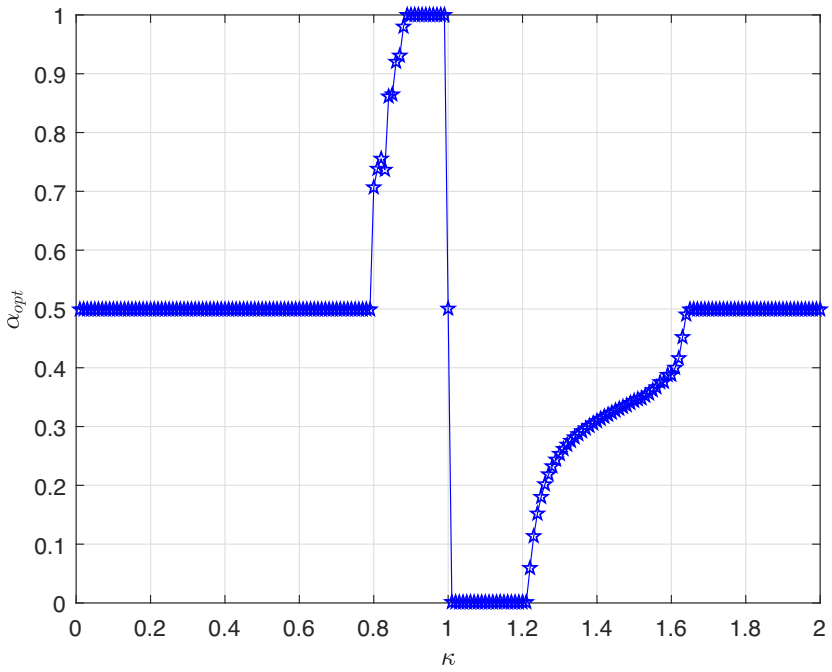



**Figure 4.4:** Optimal $\alpha$ with different $\kappa$.

The Hotelling's $T^2$ measures the magnitude of the variation remaining in each sample after projection through the PCA model. It can be calculated as the sum of the normalized squared scores as

$$T^2 = \boldsymbol{x}\mathbf{P}\boldsymbol{\Lambda}^{-1}\mathbf{P}^T\boldsymbol{x}^T. \tag{4.50}$$

The SPE quantifies the residuals of lack of fit with the PCA model, which can be derived as

$$SPE = \boldsymbol{x}(\boldsymbol{I} - \mathbf{P}\mathbf{P}^T)\boldsymbol{x}^T. \tag{4.51}$$

The multi-dimensional data model can be expressed as follows:

$$\boldsymbol{x} = \boldsymbol{A}\boldsymbol{s} + \boldsymbol{\epsilon}, \tag{4.52}$$

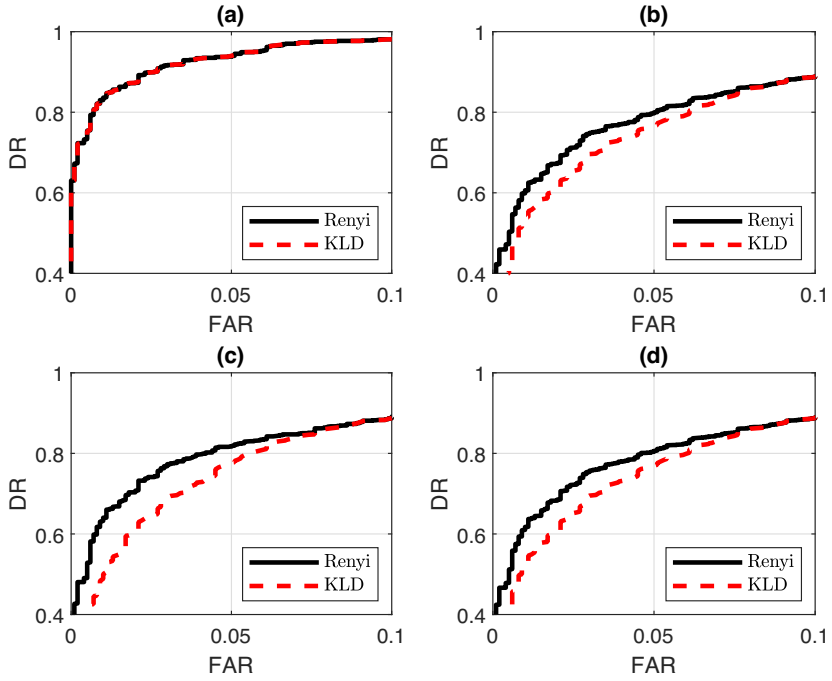



**Figure 4.5:** FAR v.s. DR comparison between the Renyi divergence and the KLD: (a) under the constant bias fault; (b) under the multiplicative fault; (c) under the combined fault; (d) overall performance

where

$$\boldsymbol{A} = \begin{bmatrix} -0.31 & -0.082 & -0.96 & 0.001 \\ -0.324 & 0.736 & -0.216 & 0.001 \\ -0.007 & -0.396 & -0.571 & 0.001 \\ -0.409 & -0.344 & -0.005 & 0.001 \\ -0.006 & 0.011 & 0.637 & 0.001 \\ -0.566 & -0.016 & 0.021 & 0.001 \\ 0.001 & 0.003 & 0.013 & -0.568 \\ -0.003 & 0.004 & 0.009 & 0.737 \end{bmatrix},$$

$\boldsymbol{x}$ is 8-dimensional measured data, $\boldsymbol{s}$ is a four-dimensional multivariate Gaussian distributed variable, $\boldsymbol{\epsilon}$ is added in the measured data to mimic the noise with zero-mean and signal-to-noise ratio (SNR) 10 dB. Under the normal condition, $\boldsymbol{s}$ follows the standard Gaussian distribution, i.e., $\boldsymbol{s} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$; while a multiplicative fault with $\mathbf{K} = \text{diag}(0.6, 0.7, 1.1, 0.8, 1.2, 1, 1, 1)$ is considered under the abnormal condition. The window size is chosen as 100, i.e., $N = 100$. There are 10000 windows of data generated under both the normal and abnormal conditions from (4.52).

In the training stage, the SVD is firstly performed on 30000 data points generated under the normal condition. From the SVD, we have that $\mathbf{\Lambda} = \text{diag}(2.71, 1.91, 1.65, 1.37, 0.093, 0.092, 0.091, 0.089)$. By setting the variance percentage threshold as 95%, i.e., $V_{th} = 95\%$, the dimension is reduced from 8 to 4, i.e., $d = 8$ and $d_m = 4$. Based on the loading matrix $\mathbf{P}$ obtained from the SVD, the SPE and the Hotelling's $T^2$ statistics are calculated with (4.50) and (4.51), respectively. The KLD is also applied on the projected data and calculated by (4.49). In the Step 5 of the proposed algorithm, two consecutive estimated variances ($n = 2$) are examined with the help of 2-sigma rule ($k = 2$) to trigger the optimization process, i.e., any $S_j$'s in two consecutive windows are out of the 2-sigma zone, the variance change is detected.

The FAR versus DR curves are shown in Figure 4.6. There are 1000 FAR-DR points obtained for each method. It can be seen that the proposed method based on the Rényi divergence with an adaptive order outperforms the rest three methods. Particularly, when the FAR is required to be less than 5%, i.e., $\gamma = 0.05$, the true FARs are 4.6%, 4.7%, 5.0% and 4.9% for the proposed method, the KLD based method, the SPE and the Hotelling's $T^2$, respectively; while the proposed method achieves the highest DR (76%), which is 8% higher than the KLD based method. Both the proposed and the KLD based methods are way better than the SPE and the Hotelling's $T^2$.

## 4.6 Summary

In this chapter, the Rényi divergence with the divergence order between 0 and 1 is proposed as the detection statistic with the multivariate Gaussian distributed data. By adding the divergence order to the KLD, the Rényi divergence covers a family of divergences and offers one more degree of freedom to be tuned for performance improvement. The FAR and MAR are considered as the detection performance measures. We first analytically derived the FAR formula based on the result that the scaled Rényi divergence asymptotically follows chi-square distribution when the number of samples is large. Under off-

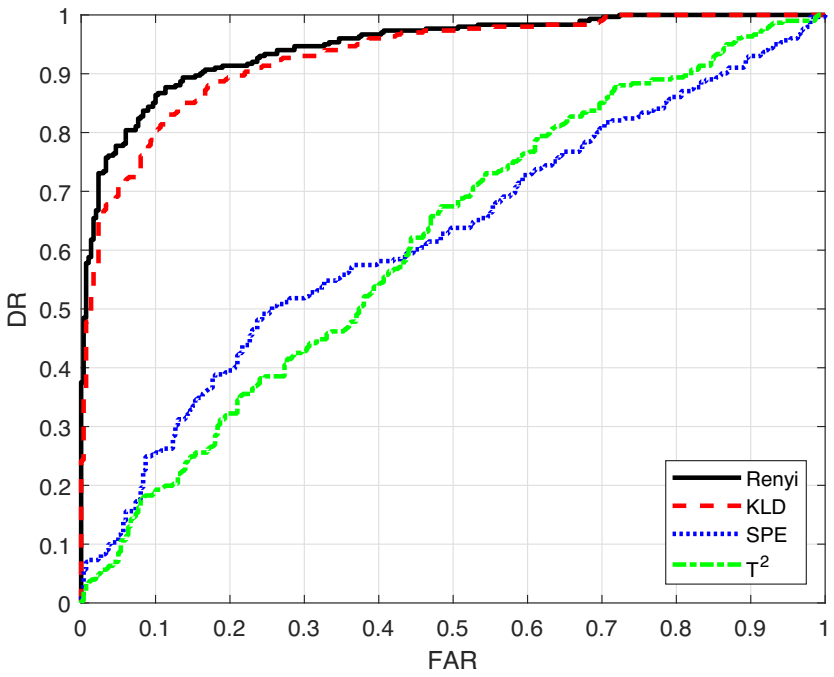



**Figure 4.6:** FAR v.s. DR comparison among the Renyi divergence, the KLD, the SPE and the Hotelling's $T^2$.

set faults, it has been proved that the Rényi divergence follows non-centralized chi-square distribution for large sample size, according to which the MAR is derived analytically. It is observed the divergence order and the threshold appear as a ratio in both the FAR and MAR under off-set faults. Otherwise stated, different divergence orders will not affect the FAR and MAR, since the threshold can be adjusted accordingly. As for the scaling faults, the MAR is formulated by regarding the Rényi divergence as a sum of positive and negative weighted non-centralized chi-squared random variables. Since the divergence order and the threshold have decouple effects on the FAR and MAR, the optimization problem with respect to the divergence order is modeled as to minimize the MAR under scaling faults subject to the given FAR requirement. In order to acquire preliminary knowledge on the occurrence of scaling faults, the $k$-sigma rule is utilized to detect the variance changes. Triggered by the $k$-sigma rule, the optimization process of the adaptive divergence order is incorporated in the proposed detection algorithm considering both the FAR

and MAR.

In the simulation, first, the theoretical results on the FAR and MAR under the normal condition and two types of faults are verified with Monte Carlo simulations for multivariate Gaussian data. It shows that the theoretical FAR and MAR results match well with those obtained from the Monte Carlo simulations. Second, the proposed algorithm is applied on one-dimensional Gaussian data and compared with the KLD based method. It shows that the Rényi divergence based method outperforms the KLD based one with higher DR under the same FAR requirement. Third, under the multi-dimensional data, the proposed method is compared with the KLD based method and two benchmark statistics, namely the Hotelling's $T^2$ and the SPE. The proposed Rényi divergence based method achieves better DR than the rest three methods with the same FAR requirement.

# Chapter 5

# Conclusions and Future Work

This chapter concludes the thesis. A summary of the main results is presented in Section 5.1, while Section 5.2 contains potential future directions.

## 5.1 Conclusions

Targeting at improving effectiveness and reliability of abnormality detection systems, this thesis conducts performance analysis and proposes new detection design schemes for linear alarm filters, the KLD based methods and the Rényi divergence based methods. The contents in this thesis can be summarized as follows:

1. The elasticity, a local sensitivity measure, is proposed to quantify the sensitivity of the detection performance over uncertainties in the trip point and over measurement noise. Analytical results on the sensitivity are derived using the Gaussian kernel based method, where there is no constraint on data distributions. Then, a linear filter design problem is formulated based on these analytical sensitivity measures in a constrained minimization structure. In the optimization problem, the analytical sensitivity results are considered as constraints, based on which the goal is to minimize the summation of the FAR and MAR. The grid search is used to find the optimal sensitivity-based linear filter.

2. The KLD based abnormality detection scheme is investigated for i.i.d.

data from GGDs with shape parameters larger than 1. First, an analytical formula of the KLD from the estimated GGD of current samples to that of historical data under the normal condition is derived analytically. Second, by considering the randomness in the sample mean and sample variance jointly and assuming there is no change in the shape parameter, a constant threshold scheme is derived with the help of the characteristic function and the central limit theorem. Further, an adaptive threshold method is proposed where the conditional KLD distributions are derived by considering the randomness in the sample mean and the sample variance separately.

3. A new detection scheme based on the Rényi divergence with the divergence order between 0 and 1 is proposed for i.i.d. multi-variate Gaussian data. First, the FAR formula is analytically derived based on the result that the scaled Rényi divergence asymptotically follows chi-square distribution when the number of samples is large. As for the MAR calculation, both off-set and scaling faults are considered. Specifically, under off-set faults, it has been analytically shown that the Rényi divergence asymptotically follows a non-centralized chi-square distribution, and different divergence orders do not affect the FAR and MAR. Under scaling faults, the MAR is formulated analytically as a function of the divergence order and threshold, which have decoupled effects on the FAR and MAR. Then, a detection scheme is proposed with an adaptive divergence order, which is calculated to minimize the MAR under the scaling faults subject to the given FAR requirement.

Intensive case studies have been conducted to verify the derived analytical performance results and show the effectiveness of the proposed abnormality detection schemes.

## 5.2 Future work

Several future directions are demonstrated in this section. A possible extension of the sensitivity analysis results is introduced in the first subsection, while the next two subsections illustrate potential new aspects and methods.

### 5.2.1 Robust Detector Design

The sensitivity measures used in this thesis are the local sensitivity measures that consider the derivative of the detection performance over measurement errors. However, it is also crucial to find a proper global sensitivity measure to quantify the overall robustness in an interval of inferences. Since the local sensitivity measure used in Chapter 2 can only quantify the response behavior against slight variable changes, global sensitivity measures called influence function and sensitivity curve [103] can be used in the filter sensitivity analysis to measure the system robustness over the entire parameter changing range. This includes two potential directions, i.e., to apply the global sensitivity measures into linear filters and to conduct sensitivity analysis for the divergence based methods.

The focus here will be the global sensitivity measure selection and the optimization problem formulation. Specifically, since both linear filters and the divergence based methods are essentially data-based statistical methods, it is possible to conduct sensitivity analysis with the help of robust theory in the robust statistical field. Actually, the influence function and the sensitivity curve are widely used in the robust statistic, which is defined to be statistics with good performance for data drawn from a wide range of probability distributions, especially for distributions that are not normal [65]. For example, the influence function in [107] describes the effect of an infinitesimal contamination at the collected data point, standardized by the mass of the contamination; and the sensitivity curve in [98] represents the effect of shifting an observation slightly to a neighbor point, i.e., add an observation at one point and remove one at another point. In the robust statistic theory,

both the influence function and the sensitivity curve should be bounded, thus a robust detector can be designed.

### 5.2.2 Performance Analysis and Design for Combined Data-Based Methods

The performance analysis and design schemes in this thesis are proposed based on simple abnormality detection schemes. It is also interesting to explore the performance of combined data-based methods.

In recent literature, the combination of two or more data based methods are gaining more attention. Precisely, in [11], a PCA based method combined with fuzzy logic filters was proposed to improve the FAR without sacrificing the detection delay. In [28], a probability-relevant PCA method was proposed and combined with the support vector machine to achieve good detection results with applications on high-speed trains. In [40], a combined method of hidden Markov and Bayesian network models was proposed and applied on the Tennessee Eastman process to show its effectiveness. However, the detection performance was evaluated from case studies instead of analytical analysis in [11, 28, 40].

According to the above articles, it is possible that a combined detection scheme with the joint design of parameters can outperform each of the individual detection scheme. In other words, it is meaningful to explore a properly combined method under the analytical performance analysis and design structure, which can be done with the help of statistical tools and existing results for each single detection scheme. The analysis can be conducted starting from basic but powerful detection schemes, for which the analytical results on the FAR and MAR are available in the literature. For example, as three basic method in alarm monitoring, filters, dead-bands and delay-timers are studied in [116], [5] and [3] respectively, where the FAR, MAR and expected detection delay were derived analytically. Besides, the analytical performance of the PCA based detection methods was studied in [147]. According to the signal flow of the combined methods and the performance analysis steps in the lit-

erature, performance measures of a combined method can be derived, based on which the corresponding design problem can be formulated.

### 5.2.3 Exploring Model-Data-Based Methods

Despite of the combination among data-based methods, it is also promising to combine the model-based and data-based methods, which are still at the preliminary stage. It has been shown in [76] that proper data-based methods, which process residuals in the model-based methods, can enhance the detection performance of a model-based detection method in terms of dealing with model uncertainties and measurement noise. In the literature, there are only a few articles in this area. Specifically, in [48], the model-based methods are combined with linear filters to achieve robust performance under exogenous disturbances for small unmanned aerial vehicles. In [76], a residual selection algorithm was developed that combined the model-based and data-based methods. The combination maximizes detection performance by finding proper residual generators.

Other than the combinations mentioned above, the particle filter is still among one of the most promising tools to combine the model-based and data-based methods, since it bridges system models and process data with the help of Bayesian frameworks. Essentially, the particle filter use sequential Monte Carlo methods to approximate the optimal filtering by representing the PDF with a swarm of particles; thus, it is able to handle any functional nonlinearity or measurement noise of any probability distributions. It has attracted much attention in the nonlinear non-Gaussian state estimation field. In the literature, an adaptive particle filter was firstly proposed in [85], where simulation results showed that the adaptive particle filter was superior than the extended Kalman filter. Later, a re-sampling method was applied into the particle filter to avoid the particle degeneracy, where after several iterations the whole probability mass was focused on a few particles in [121]. From the articles above, we can see that the particle filter contains model knowledge

via probability information. So it is possible to be related with other data-based methods such as divergence based methods and likelihood ratio based methods, where probability information is extracted as abnormal features.

# Bibliography

[1] Abdelaziz, T. H. (2015). Robust pole assignment using velocity–acceleration feedback for second-order dynamical systems with singular mass matrix. *ISA Transactions*, 57:71–84.

[2] Adnan, N. A., Izadi, I., and Chen, T. (2011a). Computing detection delays in industrial alarm systems. In *Proceedings of the 2011 American Control Conference*, pages 786–791. IEEE.

[3] Adnan, N. A., Izadi, I., and Chen, T. (2011b). On expected detection delays for alarm systems with deadbands and delay-timers. *Journal of Process Control*, 21(9):1318–1331.

[4] Afzal, M. S. (2017). *Data Driven Methods for Analysis and Design of Industrial Alarm Systems*. PhD thesis, University of Alberta.

[5] Afzal, M. S., Chen, T., Bandehkhoda, A., and Izadi, I. (2018). Analysis and design of time-deadbands for univariate alarm systems. *Control Engineering Practice*, 71:96–107.

[6] Agarwal, M., Biswas, S., and Nandi, S. (2019). Discrete event system framework for fault diagnosis with measurement inconsistency: case study of rogue DHCP attack. *IEEE/CAA Journal of Automatica Sinica*, 6(3):789–806.

[7] Aiazzi, B., Alparone, L., and Baronti, S. (1999). Estimation based on entropy matching for generalized Gaussian PDF modeling. *IEEE Signal Processing Letters*, 6(6):138–140.

[8] Alauddin, M., Khan, F., Imtiaz, S., and Ahmed, S. (2018). A bibliometric review and analysis of data-driven fault detection and diagnosis methods for process systems. *Industrial & Engineering Chemistry Research*, 57(32):10719–10735.

[9] Aldous, D. J. and Diaconis, P. (1987). Strong uniform times and finite random walks. *Advances in Applied Mathematics*, 8(1):69 – 97.

[10] Amin, M. T., Khan, F., and Imtiaz, S. (2019). Fault detection and pathway analysis using a dynamic Bayesian network. *Chemical Engineering Science*, 195:777–790.

[11] Ammiche, M., Kouadri, A., and Bakdi, A. (2018). A combined monitoring scheme with fuzzy logic filter for plant-wide Tennessee Eastman process fault detection. *Chemical Engineering Science*, 187:269–279.

[12] Arifin, B., Li, Z., and Shah, S. L. (2015). Pipeline leak detection using particle filters. *IFAC-PapersOnLine*, 48(8):76–81.

[13] Bang, M., Engelsgaard, S. S., Alexandersen, E. K., Skydt, M. R., Shaker, H. R., and Jradi, M. (2019). Novel real-time model-based fault detection method for automatic identification of abnormal energy performance in building ventilation units. *Energy and Buildings*, 183:238 – 251.

[14] Basseville, M. and Nikiforov, I. V. (1993). *Detection of Abrupt Changes: Theory and Application*. Englewood Cliffs, N.J.: Prentice Hall.

[15] Behal, S. and Kumar, K. (2017). Detection of DDoS attacks and flash events using novel information theory metrics. *Computer Networks*, 116:96–110.

[16] Behal, S., Kumar, K., and Sachdeva, M. (2018). A generalized detection system to detect distributed denial of service attacks and flash events for information theory metrics. *Turkish Journal of Electrical Engineering & Computer Sciences*, 26(4):1759–1770.

[17] Benkouider, A., Kessas, R., Yahiaoui, A., Buvat, J., and Guella, S. (2012). A hybrid approach to faults detection and diagnosis in batch and semi-batch reactors by using EKF and neural network classifier. *Journal of Loss Prevention in the Process Industries*, 25(4):694–702.

[18] Bhuyan, M. H., Bhattacharyya, D., and Kalita, J. K. (2015). An empirical evaluation of information metrics for low-rate and high-rate DDoS attack detection. *Pattern Recognition Letters*, 51:1–7.

[19] Bodenham, D. A. and Adams, N. M. (2016). A comparison of efficient approximations for a weighted sum of chi-squared random variables. *Statistics and Computing*, 26(4):917–928.

[20] Borgonovo, E. and Apostolakis, G. E. (2001). A new importance measure for risk-informed decision making. *Reliability Engineering & System Safety*, 72(2):193–212.

[21] Borgonovo, E. and Plischke, E. (2016). Sensitivity analysis: A review of recent advances. *European Journal of Operational Research*, 248(3):869–887.

[22] Boškoski, P., Gašperin, M., Petelin, D., and Juričić, Đ. (2015). Bearing fault prognostics using Rényi entropy based features and Gaussian process models. *Mechanical Systems and Signal Processing*, 52:327–337.

[23] Brandsæter, A., Vanem, E., and Glad, I. K. (2019). Efficient on-line anomaly detection for ship systems in operation. *Expert Systems with Applications*, 121:418–437.

[24] Ćesić, J., Marković, I., and Petrović, I. (2016). Moving objects tracking on the unit sphere using a multiple-camera system on a mobile robot. In *Intelligent Autonomous Systems 13*, pages 899–911. Springer.

[25] Chang, C. T. and Hwang, J. I. (1998). Simplification techniques for EKF

computations in fault diagnosis—suboptimal gains. *Chemical Engineering Science*, 53(22):3853–3862.

[26] Chang, H.-C., Jheng, Y.-M., Kuo, C.-C., and Huang, L.-B. (2016). On-line motor condition monitoring system for abnormality detection. *Computers & Electrical Engineering*, 51:255–269.

[27] Chen, E. J. and Kelton, W. D. (2014). Density estimation from correlated data. *Journal of Simulation*, 8(4):281–292.

[28] Chen, H., Jiang, B., Chen, W., and Yi, H. (2019). Data-driven detection and diagnosis of incipient faults in electrical drives of high-speed trains. *IEEE Transactions on Industrial Electronics*, 66(6):4716–4725.

[29] Chen, H., Jiang, B., and Lu, N. (2018a). An improved incipient fault detection method based on Kullback-Leibler divergence. *ISA Transactions*, 79:127–136.

[30] Chen, S. X. (2000). Probability density function estimation using Gamma kernels. *Annals of the Institute of Statistical Mathematics*, 52(3):471–480.

[31] Chen, Z., Han, F., Wu, L., Yu, J., Cheng, S., Lin, P., and Chen, H. (2018b). Random forest based intelligent fault diagnosis for PV arrays using array voltage and string currents. *Energy Conversion and Management*, 178:250–264.

[32] Cheng, Y. (2013). *Data-driven Techniques on Alarm System Analysis and Improvement*. PhD thesis, University of Alberta.

[33] Cheng, Y., Izadi, I., and Chen, T. (2011). On optimal alarm filter design. In *2011 International Symposium on Advanced Control of Industrial Processes (ADCONIP)*, pages 139–145. IEEE.

[34] Cheng, Y., Izadi, I., and Chen, T. (2013). Optimal alarm signal processing: Filter design and performance analysis. *IEEE Transactions on Automation Science and Engineering*, 10(2):446–451.

[35] Chetouani, Y. (2007). Use of cumulative sum (CUSUM) test for detecting abrupt changes in the process dynamics. *International Journal of Reliability, Quality and Safety Engineering*, 14(01):65–80.

[36] Cho, S. and Jiang, J. (2016). Detection of sensor abnormalities in a pressurizer by means of analytical redundancy. *IEEE Transactions on Nuclear Science*, 63(6):2925–2933.

[37] Clark, R. N. (1979). The dedicated observer approach to instrument failure detection. In *1979 18th IEEE Conference on Decision and Control including the Symposium on Adaptive Processes*, volume 2, pages 237–241. IEEE.

[38] Delpha, C., Diallo, D., and Youssef, A. (2017). Kullback-Leibler divergence for fault estimation and isolation: Application to Gamma distributed data. *Mechanical Systems and Signal Processing*, 93:118–135.

[39] Do, M. N. and Vetterli, M. (2002). Wavelet-based texture retrieval using generalized Gaussian density and Kullback-Leibler distance. *IEEE Transactions on Image Processing*, 11(2):146–158.

[40] Don, M. G. and Khan, F. (2019). Dynamic process fault detection and diagnosis based on a combined approach of hidden Markov and Bayesian network model. *Chemical Engineering Science*, 201:82–96.

[41] Dong, G., Chongguang, W., Zhang, B., and Xin, M. (2010). Signed directed graph and qualitative trend analysis based fault diagnosis in chemical industry. *Chinese Journal of Chemical Engineering*, 18(2):265–276.

[42] Dugan, J. B., Bavuso, S. J., and Boyd, M. A. (1992). Dynamic fault-tree models for fault-tolerant computer systems. *IEEE Transactions on Reliability*, 41(3):363–377.

[43] Farrell, M. D. and Mersereau, R. M. (2005). On the impact of KDE

dimension reduction for hyperspectral detection of difficult targets. *IEEE Geoscience and Remote Sensing Letters*, 2(2):192–195.

[44] Ferracuti, F., Giantomassi, A., Iarlori, S., Ippoliti, G., and Longhi, S. (2013). Induction motor fault detection and diagnosis using KDE and Kullback-Leibler divergence. In *2013 39th Annual Conference in Industrial Electronics Society of the IEEE (IECON)*, pages 2923–2928. IEEE.

[45] Fezai, R., Mansouri, M., Trabelsi, M., Hajji, M., Nounou, H., and Nounou, M. (2019). Online reduced kernel GLRT technique for improved fault detection in photovoltaic systems. *Energy*, 179:1133–1154.

[46] Frank, P. M. (1990). Fault diagnosis in dynamic systems using analytical and knowledge-based redundancy: A survey and some new results. *Automatica*, 26(3):459–474.

[47] Frank, P. M. (1996). Analytical and qualitative model-based fault diagnosis–a survey and some new results. *European Journal of Control*, 2(1):6–28.

[48] Freeman, P., Pandita, R., Srivastava, N., and Balas, G. J. (2013). Model-based and data-driven fault detection performance for a small UAV. *IEEE/ASME Transactions on Mechatronics*, 18(4):1300–1309.

[49] Gao, Y., Xiao, F., Liu, J., and Wang, R. (2019). Distributed soft fault detection for interval type-2 fuzzy-model-based stochastic systems with wireless sensor networks. *IEEE Transactions on Industrial Informatics*, 15(1):334–347.

[50] Gao, Z., Cecati, C., and Ding, S. X. (2015). A survey of fault diagnosis and fault-tolerant techniques—part I: Fault diagnosis with model-based and signal-based approaches. *IEEE Transactions on Industrial Electronics*, 62(6):3757–3767.

[51] Gil, M., Alajaji, F., and Linder, T. (2013). Rényi divergence measures for commonly used univariate continuous distributions. *Information Sciences*, 249:124–131.

[52] Gonzalez, R., Huang, B., and Lau, E. (2015). Process monitoring using kernel density estimation and Bayesian networking with an industrial case study. *ISA Transactions*, 58:330–347.

[53] Goodman, I. R. and Kotz, S. (1973). Multivariate $\theta$-generalized normal distributions. *Journal of Multivariate Analysis*, 3(2):204–219.

[54] Gosumbonggot, J. and Fujita, G. (2019). Global maximum power point tracking under shading condition and hotspot detection algorithms for photovoltaic systems. *Energies*, 12(5):882.

[55] Gradshteyn, I. S. and Ryzhik, I. M. (2014). *Table of Integrals, Series, and Products*. Academic Press.

[56] Gunawan, S. and Azarm, S. (2004). Non-gradient based parameter sensitivity estimation for single objective robust design optimization. *Journal of Mechanical Design*, 126(3):395–402.

[57] Guo, Y., Wang, J., Chen, H., Li, G., Huang, R., Yuan, Y., Ahmad, T., and Sun, S. (2019). An expert rule-based fault diagnosis strategy for variable refrigerant flow air conditioning systems. *Applied Thermal Engineering*, 149:1223–1235.

[58] Hall, P., Sheather, S. J., Jones, M., and Marron, J. S. (1991). On optimal data-based bandwidth selection in kernel density estimation. *Biometrika*, 78(2):263–269.

[59] Hamadouche, A., Kouadri, A., and Bakdi, A. (2017). A modified Kullback-Leibler divergence for direct fault detection in large scale systems. *Journal of Process Control*, 59:28–36.

[60] Hao, H., Wang, M., Tang, Y., and Li, Q. (2018). Research on data fusion of multi-sensors based on fuzzy preference relations. *Neural Computing and Applications*, pages 1–10.

[61] He, Y., Hamza, A. B., and Krim, H. (2003). A generalized divergence measure for robust image registration. *IEEE Transactions on Signal Processing*, 51(5):1211–1220.

[62] Hobza, T., Molina, I., and Morales, D. (2007). *Rényi Statistics for Testing Hypotheses with s Samples*. Research report I-2007-30 in the Operations Research Center of the Miguel Hernández University of Elche.

[63] Hossain, M. P., Sanusi, R. A., Omar, M. H., and Riaz, M. (2019). On designing maxwell cusum control chart: an efficient way to monitor failure rates in boring processes. *The International Journal of Advanced Manufacturing Technology*, 100(5-8):1923–1930.

[64] Huang, C.-Y. (2018). Applying multivariate analysis to analyze and improve component rejection by pick and place machines. *The International Journal of Advanced Manufacturing Technology*, 96(1-4):1265–1281.

[65] Huber, P. J. (2011). *Robust Statistics*. Springer.

[66] Ibrahim, R. K., Watson, S. J., Djurović, S., and Crabtree, C. J. (2018). An effective approach for rotor electrical asymmetry detection in wind turbine DFIGs. *IEEE Transactions on Industrial Electronics*, 65(11):8872–8881.

[67] Isermann, R. (1984). Process fault detection based on modeling and estimation methods–a survey. *Automatica*, 20(4):387–404.

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

[69] Isermann, R. (2006). *Fault-Diagnosis Systems: An Introduction from Fault Detection to Fault Tolerance*. Springer Science & Business Media.

[70] Izadi, I., Shah, S. L., Shook, D. S., Kondaveeti, S. R., and Chen, T. (2009). A framework for optimal design of alarm systems. In *Proceedings of 7th IFAC Symposium on Fault Detection, Supervision and Safety of Technical Processes*, pages 651–656.

[71] James, A. T., Gandhi, O., and Deshmukh, S. (2018). Fault diagnosis of automobile systems using fault tree based on digraph modeling. *International Journal of System Assurance Engineering and Management*, 9:1–15.

[72] Jiang, B., Luo, Y., and Lu, Q. (2019a). Maximized mutual information analysis based on stochastic representation for process monitoring. *IEEE Transactions on Industrial Informatics*, 15(3):1579–1587.

[73] Jiang, G., He, H., Yan, J., and Xie, P. (2019b). Multiscale convolutional neural networks for fault diagnosis of wind turbine gearbox. *IEEE Transactions on Industrial Electronics*, 66(4):3196–3207.

[74] Jiang, Q., Gao, F., Yan, X., and Yi, H. (2019). Multiobjective two-dimensional CCA-based monitoring for successive batch processes with industrial injection molding application. *IEEE Transactions on Industrial Electronics*, 66(5):3825–3834.

[75] Jiang, W., Shu, L., and Apley, D. W. (2008). Adaptive CUSUM procedures with EWMA-based shift estimators. *IIE Transactions*, 40(10):992–1003.

[76] Jung, D. and Sundström, C. (2017). A combined data-driven and model-based residual selection algorithm for fault detection and isolation. *IEEE Transactions on Control Systems Technology*, (99):1–15.

[77] Karnavas, W. J., Sanchez, P. J., and Bahill, A. T. (1993). Sensitivity analyses of continuous and discrete systems in the time and frequency domains. *IEEE Transactions on Systems, Man and Cybernetics*, 23(2):488–501.

[78] Kim, H. Y., Gribbin, M. J., Muller, K. E., and Taylor, D. J. (2006). Analytic, computational, and approximate forms for ratios of noncentral and central Gaussian quadratic forms. *Journal of Computational and Graphical Statistics*, 15(2):443–459.

[79] Kramer, M. A. and Palowitch, B. (1987). A rule-based approach to fault diagnosis using the signed directed graph. *AIChE Journal*, 33(7):1067–1078.

[80] Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, 22(1):79–86.

[81] Leite, G. d. N. P., Araújo, A. M., Rosas, P. A. C., Stosic, T., and Stosic, B. (2019). Entropy measures for early detection of bearing faults. *Physica A: Statistical Mechanics and its Applications*, 514:458–472.

[82] Li, B., Jing, Y., and Xu, W. (2017). A generic waveform abnormality detection method for utility equipment condition monitoring. *IEEE Transactions on Power Delivery*, 32(1):162–171.

[83] Li, L., Ding, S. X., Qiu, J., Yang, Y., and Zhang, Y. (2016). Weighted fuzzy observer-based fault detection approach for discrete-time nonlinear systems via piecewise-fuzzy Lyapunov functions. *IEEE Transactions on Fuzzy Systems*, 24(6):1320–1333.

[84] Lin, C. E., Ling, J.-M., and Huang, C.-L. (1993). An expert system for transformer fault diagnosis using dissolved gas analysis. *IEEE Transactions on Power Delivery*, 8(1):231–238.

[85] Liu, M., Zang, S., and Zhou, D. (2005). Fast leak detection and location of gas pipelines based on an adaptive particle filter. *International Journal of Applied Mathematics and Computer Science*, 15(4):541.

[86] Liu, W., Wang, Y.-L., Liu, J., Huang, L., and Hao, C. (2017). Performance analysis of adaptive detectors for point targets in subspace interfer-

ence and Gaussian noise. *IEEE Transactions on Aerospace and Electronic Systems*, 54(1):429–441.

[87] Ma, S., Chu, F., and Han, Q. (2019). Deep residual learning with demodulated time-frequency features for fault diagnosis of planetary gearbox under nonstationary running conditions. *Mechanical Systems and Signal Processing*, 127:190–201.

[88] Mansouri, M., Al-Khazraji, A., Hajji, M., Harkat, M. F., Nounou, H., and Nounou, M. (2018). Wavelet optimized EWMA for fault detection and application to photovoltaic systems. *Solar Energy*, 167:125–136.

[89] Marks, N. B. (2007). Kolmogorov–Smirnov test statistic and critical values for the Erlang-3 and Erlang-4 distributions. *Journal of Applied Statistics*, 34(8):899–906.

[90] Meskin, N., Khorasani, K., and Rabbath, C. A. (2010). A hybrid fault detection and isolation strategy for a network of unmanned vehicles in presence of large environmental disturbances. *IEEE Transactions on Control Systems Technology*, 18(6):1422–1429.

[91] Miao, Y., Zhao, M., Makis, V., and Lin, J. (2019). Optimal swarm decomposition with whale optimization algorithm for weak feature extraction from multicomponent modulation signal. *Mechanical Systems and Signal Processing*, 122:673–691.

[92] Muralidharan, V. and Sugumaran, V. (2012). A comparative study of naïve bayes classifier and bayes net classifier for fault diagnosis of monoblock centrifugal pump using wavelet analysis. *Applied Soft Computing*, 12(8):2023–2029.

[93] Murvay, P.-S. and Silea, I. (2012). A survey on gas leak detection and localization techniques. *Journal of Loss Prevention in the Process Industries*, 25(6):966–973.

[94] Newhart, K. B., Holloway, R. W., Hering, A. S., and Cath, T. Y. (2019). Data-driven performance analyses of wastewater treatment plants: A review. *Water research*.

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

[96] Parzen, E. (1962). On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33(3):1065–1076.

[97] Platon, R., Martel, J., Woodruff, N., and Chau, T. Y. (2015). Online fault detection in PV systems. *IEEE Transactions on Sustainable Energy*, 6(4):1200–1207.

[98] Quintern, L., Furusawa, Y., Fukutsu, K., and Holtschmidt, H. (1997). Characterization and application of UV detector spore films: the sensitivity curve of a new detector system provides good similarity to the action spectrum for UV-induced erythema in human skin. *Journal of Photochemistry and Photobiology B: Biology*, 37(1-2):158–166.

[99] Ren, H., Yin, A., Zhou, Q., Li, J., and Hu, Y. (2019). A wind turbine bearing performance evaluation method based on similarity analysis of fuzzy k-principal curves in manifold space. *IEEE Access*, 7:36154–36163.

[100] Rényi, A. (1961). On measures of entropy and information. Technical report, Hungarian Academy of Sciences Budapest Hungary.

[101] Roberts, S. (1959). Control chart tests based on geometric moving averages. *Technometrics*, 1(3):239–250.

[102] Robinson, P. and Ho, T. Y. (1978). Average run lengths of geometric moving average charts by numerical methods. *Technometrics*, 20(1):85–93.

[103] Rousseeuw, P. J. and Ronchetti, E. (1981). Influence curves of general statistics. *Journal of Computational and Applied Mathematics*, 7(3):161–166.

[104] Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M., and Tarantola, S. (2008). *Global Sensitivity Analysis: the Primer*. John Wiley & Sons.

[105] Sason, I. (2016). On the Rényi divergence, joint range of relative entropies, and a channel coding theorem. *IEEE Transactions on Information Theory*, 62(1):23–34.

[106] Schmid, W., Schöne, A., et al. (1997). Some properties of the EWMA control chart in the presence of autocorrelation. *The Annals of Statistics*, 25(3):1277–1283.

[107] Sheorey, P., Loui, J., Singh, K., and Singh, S. K. (2000). Ground subsidence observations and a modified influence function method for complete subsidence prediction. *International Journal of Rock Mechanics and Mining Sciences*, 37(5):801–818.

[108] Shon, S., Mun, S., and Ko, H. (2017). Recursive whitening transformation for speaker recognition on language mismatched condition. *arXiv preprint arXiv:1708.01232*.

[109] Shuqing, Z., Tianye, G., Xu, H., Jian, J., and Zhongdong, W. (2009). Research on pipeline leak detection based on Hilbert-Huang transform. In *2009 International Conference on Energy and Environment Technology*, volume 3, pages 500–503. IEEE.

[110] Simon, M. K. (2007). *Probability Distributions Involving Gaussian Random Variables: A Handbook for Engineers and Scientists*. Springer Science & Business Media.

[111] Singh, J., Darpe, A., and Singh, S. (2017). Bearing damage assessment using Jensen-Rényi divergence based on EEMD. *Mechanical Systems and Signal Processing*, 87:307–339.

[112] Smaili, R., El Harabi, R., and Abdelkrim, M. (2017). Design of fault monitoring framework for multi-energy systems using signed directed graph. *IFAC-PapersOnLine*, 50(1):15734–15739.

[113] Srivastava, M. and Wu, Y. (1997). Evaluation of optimum weights and average run lenghts in EWMA control schemes. *Communications in Statistics-Theory and Methods*, 26(5):1253–1267.

[114] St-Onge, X. F., Cameron, J., Saleh, S., and Scheme, E. J. (2019). A symmetrical component feature extraction method for fault detection in induction machines. *IEEE Transactions on Industrial Electronics*, 66(9):7281–7289.

[115] Sun, C., Wang, P., Yan, R., Gao, R. X., and Chen, X. (2019). Machine health monitoring based on locally linear embedding with kernel sparse representation for neighborhood optimization. *Mechanical Systems and Signal Processing*, 114:25–34.

[116] Sun, Y., Tan, W., and Chen, T. (2018). A method to remove chattering alarms using median filters. *ISA Transactions*, 73:201–207.

[117] Taghizadeh-Alisaraei, A. and Mahdavian, A. (2019). Fault detection of injectors in diesel engines using vibration time-frequency analysis. *Applied Acoustics*, 143:48–58.

[118] Tan, W., Sun, Y., Azad, I. I., and Chen, T. (2017). Design of univariate alarm systems via rank order filters. *Control Engineering Practice*, 59:55–63.

[119] Tong, Z. Y., Dong, Z. Y., and Li, M. (2016). A new entropy bi-cepstrum based-method for dc motor brush abnormality recognition. *IEEE Sensors Journal*, 17(3):745–754.

[120] Tsay, R. S. (2005). *Analysis of Financial Time Series*. John Wiley & Sons.

[121] Uilhoorn, F. E. (2015). State-space estimation with a Bayesian filter in a coupled PDE system for transient gas flows. *Applied Mathematical Modelling*, 39(2):682–692.

[122] Van Erven, T. and Harremos, P. (2014). Rényi divergence and Kullback-Leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820.

[123] Van Nguyen, H., Rezatofighi, S. H., Vo, B.-N., and Ranasinghe, D. C. (2018). Online UAV path planning for joint detection and tracking of multiple radio-tagged objects. *arXiv preprint arXiv:1808.04445*.

[124] Varanasi, M. K. and Aazhang, B. (1989). Parametric generalized Gaussian density estimation. *The Journal of the Acoustical Society of America*, 86(4):1404–1415.

[125] Venkatasubramanian, V., Rengaswamy, R., and Kavuri, S. N. (2003a). A review of process fault detection and diagnosis: Part II: Qualitative models and search strategies. *Computers & Chemical Engineering*, 27(3):313–326.

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

[127] Walck, C. (2007). *Handbook on statistical distributions for experimentalists*. University of Stockholm Internal Report SUF-PFY/96-01.

[128] Wang, J., Yang, F., Chen, T., and Shah, S. L. (2015). An overview of industrial alarm systems: Main causes for alarm overloading, research status, and open problems. *IEEE Transactions on Automation Science and Engineering*, 13(2):1045–1061.

[129] Wang, R., Edgar, T. F., Baldea, M., Nixon, M., Wojsznis, W., and Dunia, R. (2018). A geometric method for batch data visualization, process monitoring and fault detection. *Journal of Process Control*, 67:197–205.

[130] Wang, Z. and Shang, H. (2015). Kalman filter based fault detection for two-dimensional systems. *Journal of Process Control*, 28:83–94.

[131] Wu, M. S. Y. (1995). Evaluation of optimum weights and average run lengths in EWMA control schemes. Technical report, University of Toronto.

[132] Xiong, Y., Jing, Y., and Chen, T. (2017). Performance sensitivity analysis of linear alarm filters. In *Proceedings of the 2017 American Control Conference*, pages 4424–4429. IEEE.

[133] Xiong, Y., Jing, Y., and Chen, T. (2019). Abnormality detection based on the Kullback–Leibler divergence for generalized Gaussian data. *Control Engineering Practice*, 85:257–270.

[134] Yang, F., Shah, S. L., and Xiao, D. (2010). SDG (signed directed graph) based process description and fault propagation analysis for a tailings pumping process. *IFAC Proceedings Volumes*, 43(9):50–55.

[135] Yang, H. and Saif, M. (1995). Nonlinear adaptive observer design for fault detection. In *Proceedings of the 1995 American Control Conference*, volume 2, pages 1136–1139. IEEE.

[136] Yang, X., Chen, W., Li, A., Yang, C., Xie, Z., and Dong, H. (2019). BA-PNN-based methods for power transformer fault diagnosis. *Advanced Engineering Informatics*, 39:178–185.

[137] Yiakopoulos, C., Gryllias, K., Chioua, M., Hollender, M., and Antoniadis, I. (2016). An on-line SAX and HMM-based anomaly detection and visualization tool for early disturbance discovery in a dynamic industrial process. *Journal of Process Control*, 44:134–159.

[138] Yin, S., Ding, S. X., Haghani, A., Hao, H., and Zhang, P. (2012). A comparison study of basic data-driven fault diagnosis and process monitoring methods on the benchmark Tennessee Eastman process. *Journal of process control*, 22(9):1567–1581.

[139] Youssef, A., Delpha, C., and Diallo, D. (2016). An optimal fault detection threshold for early detection using Kullback–Leibler divergence for unknown distribution data. *Signal Processing*, 120:266–279.

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

[141] Yu, J. and Yan, X. (2019). Active features extracted by deep belief network for process monitoring. *ISA Transactions*, 84:247–261.

[142] Yu, Y., Zhu, D., Wang, J., and Zhao, Y. (2017). Abnormal data detection for multivariate alarm systems based on correlation directions. *Journal of Loss Prevention in the Process Industries*, 45:43–55.

[143] Zames, G. and Francis, B. (1983). Feedback, minimax sensitivity, and optimal robustness. *IEEE Transactions on Automatic Control*, 28(5):585–601.

[144] Zeng, J., Kruger, U., Geluk, J., Wang, X., and Xie, L. (2014). Detecting abnormal situations using the Kullback–Leibler divergence. *Automatica*, 50(11):2777–2786.

[145] Zeng, Z., Tan, W., and Zhou, R. (2017). Computation of performance indices for generalized delay-timers. In *Proceedings of the 2017 American Control Conference*, pages 4803–4808. IEEE.

[146] Zhang, K., Hao, H., Chen, Z., Ding, S. X., and Peng, K. (2015). A comparison and evaluation of key performance indicator-based multivariate statistics process monitoring approaches. *Journal of Process Control*, 33:112–126.

[147] Zhang, K., Shardt, Y. A., Chen, Z., and Peng, K. (2017). Using the expected detection delay to assess the performance of different multivariate

statistical process monitoring methods for multiplicative and drift faults. *ISA Transactions*, 67:56–66.

[148] Zhang, X., Wang, J., Liu, Z., and Wang, J. (2019). Weak feature enhancement in machinery fault diagnosis using empirical wavelet transform and an improved adaptive bistable stochastic resonance. *ISA Transactions*, 84:283–295.

[149] Zhang, Y., Chen, S., Li, J., and Jin, S. (2014). Leak detection monitoring system of long distance oil pipeline based on dynamic pressure transmitter. *Measurement*, 49:382–389.

[150] Zhao, Y., Liu, P., Wang, Z., Zhang, L., and Hong, J. (2017). Fault and defect diagnosis of battery for electric vehicles based on big data analysis methods. *Applied Energy*, 207:354–362.