5-2010

An Effective Approach to Nonparametric Quickest Detection and Its Decentralized Realization

Dayu Yang
University of Tennessee - Knoxville, dyang@utk.edu

To the Graduate Council:

I am submitting herewith a dissertation written by Dayu Yang entitled "An Effective Approach to Nonparametric Quickest Detection and Its Decentralized Realization." I have examined the final electronic copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Electrical Engineering.

Hairong Qi, Major Professor

We have read this dissertation and recommend its acceptance:

Husheng Li, Seddik M. Djouadi, Xiaobing Feng

Accepted for the Council:

Carolyn R. Hodges

Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)

To the Graduate Council:
I am submitting herewith a dissertation written by Dayu Yang entitled "An Effective Approach to Nonparametric Quickest Detection and Its Decentralized Realization". I have examined the final paper copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Electrical Engineering.

Hairong Qi, Major Professor

We have read this dissertation
and recommend its acceptance:

Husheng Li

Seddik M. Djouadi

Xiaobing Feng

Accepted for the Council:

Carolyn R. Hodges

Vice Provost and Dean of the
Graduate School

To the Graduate Council:
I am submitting herewith a dissertation written by Dayu Yang entitled "An Effective Approach to Nonparametric Quickest Detection and Its Decentralized Realization". I have examined the final paper copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Electrical Engineering.

Hairong Qi

Major Professor

We have read this dissertation
and recommend its acceptance:

Husheng Li

Seddik M. Djouadi

Xiaobing Feng

Accepted for the Council:

Carolyn R. Hodges

Vice Provost and Dean of the
Graduate School

(Original signatures are on file with official student records.)

An Effective Approach to Nonparametric Quickest Detection and Its Decentralized Realization

A Dissertation
Presented for the
Doctor of Philosophy Degree
The University of Tennessee, Knoxville

Dayu Yang
May 2010

Dedication

I dedicate my dissertation work to my loving wife, Sijie Yu, who has been proud and supportive of my work and who has shared the many uncertainties, challenges and sacrifices for completing this dissertation; to my parents, Yunmin Yang and Shunying Yu, and my big brother, Yongwei Yang, whose endless love have encouraged me during my graduate studies, during my whole life.

Acknowledgments

I would like to thank my advisor, Dr. Hairong Qi, who has made me who I am and who I will be in my professional career. Her willingness to support my work and her guidance throughout my studies has allowed me to develop my abilities to think through the problems and skills to find the answers. I thank her for that opportunity. Also, I would like to thank Dr. Husheng Li, Dr. Seddik M. Djouadi and Dr. Xiaobing Feng. Their advice and counsel have been of equal importance. I greatly appreciate their time and input to this dissertation.

Within the AICIP group, I owe many thanks to my fellow graduate labmates, Yang Bai, Sangwoo Moon, Mahmut Karakaya, Harika Tandra, Zhiliang Tu, Thomas P. Karnowski, Austin Albright, Paul Donnelly. I enjoyed the conversations and discussions regarding my research and all the other topics.

Finally, I must express my appreciation to the many people outside of my studies who have helped to relieve the sometimes stressful solitude of graduate school. Last but not least, I would like to thank all the faculties and students within the department whose encouragement and friendship have truly inspired me during my graduate study.

Abstract

This dissertation focuses on the study of nonparametric quickest detection and its decentralized implementation in a distributed environment. Quickest detection schemes are geared toward detecting a change in the state of a data stream or a real-time process. Classical quickest detection schemes invariably assume knowledge of the pre-change and post-change distributions that may not be available in many applications.

A distribution free nonparametric quickest detection procedure is presented based on a novel distance measure, referred to as the Q-Q distance calculated from the Quantile-Quantile plot. Theoretical analysis of the distance measure and detection procedure is presented to justify the proposed algorithm and provide performance guarantees. The Q-Q distance based detection procedure presents comparable performance compared to classical parametric detection procedure and better performance than other nonparametric procedures. The proposed procedure is most effective when detecting *small* changes.

As the technology advances, distributed sensing and detection become feasible. Existing decentralized detection approaches are largely parametric. The *decentralized* realization of Q-Q distance based nonparametric quickest detection scheme is further studied, where data streams are simultaneously collected from multiple channels located distributively to jointly reach a detection decision. Two implementation schemes, binary quickest detection and local decision fusion, are described. Experimental results show that the proposed method has a comparable performance to the benchmark parametric cumulative sum (CUSUM) test in binary detection. Finally the dissertation concludes with a summary of the contributions to the state of the art.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Quickest Detection

Quickest detection refers to real-time version of abrupt change detection. Figure 1.1 and 1.2 show two examples of the "change" we are discussing. In Figure 1.1 the change occurs on the mean value but the variance is unchanged. In Figure 1.2 the variance varies while the mean keeps the same. The goal of quickest detection is to detect the changes as soon as possible after they occur while keeping the false alarm rate below a given level. The research of abrupt change detection dates back to 1930s when Shewhart [Shewhart, 1931] first introduced this idea in manufacturing and business process quality control. It has attracted attention in a wide variety of fields recently including applications in network security, recognition-oriented signal processing, econometrics, environment modeling, finance, image analysis, medical diagnosis, fraud detection, counter-terrorism, and so on. This research assumes that the properties or parameters describing the data are normally constant or slowly time-varying in responding parametric models but these properties or parameters are subject to abrupt changes at some unknown time instants. Since most of the adaptive estimation algorithms are unable to catch the fast changes [Basseville and Nikiforov, 1993], investigations of finding effective detect procedures to follow the abrupt changes gets increasing attention.

The original formulation of the change detection problem is a single channel formulation, in which there is a sequence of independent and identical distributed (i.i.d) observations.

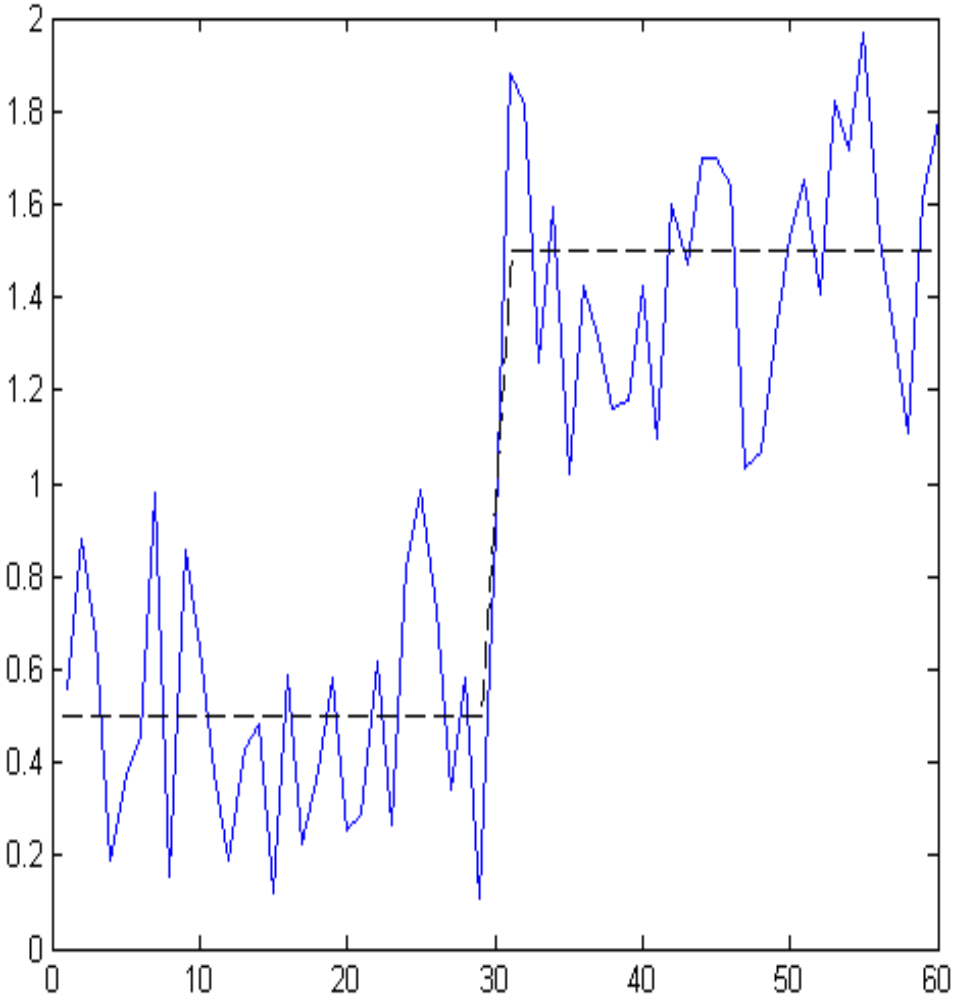

Figure 1.1: Change in mean value

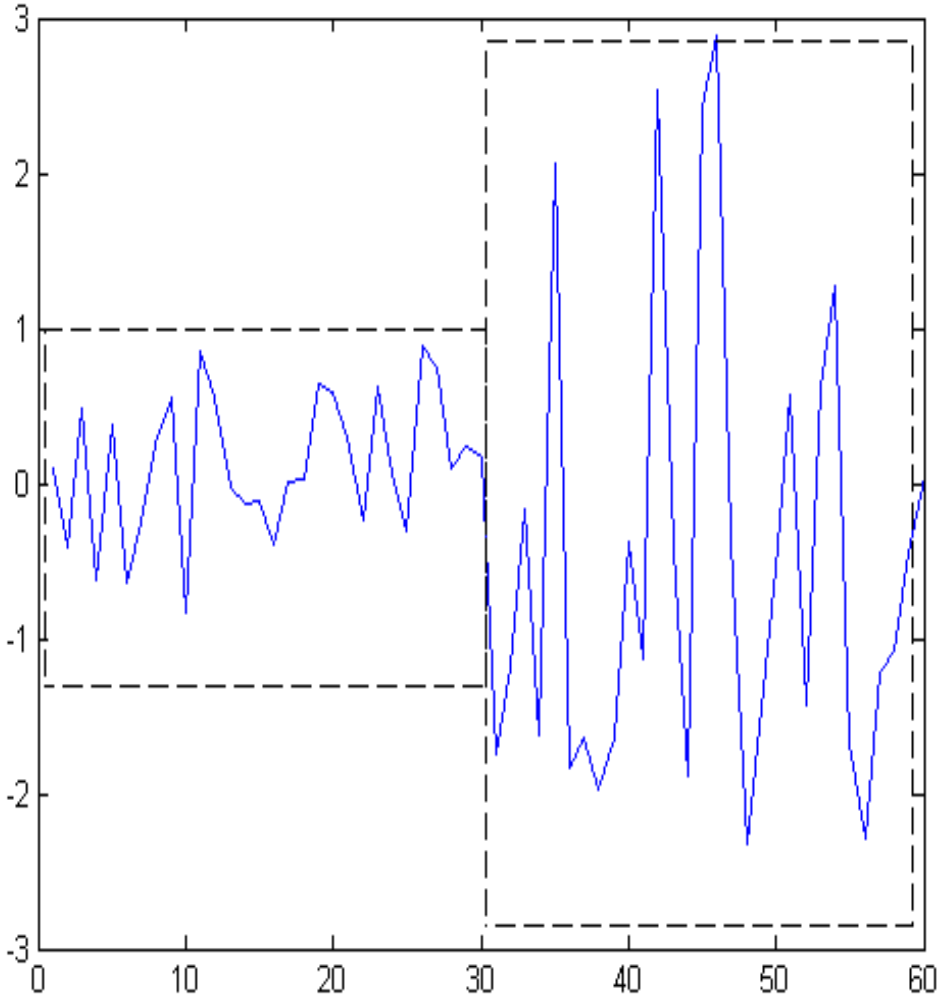

Figure 1.2: Change in variance

The distribution of the sequence changes at some unknown point in time and the pre-change and post-change distributions are assumed known.

On one hand, among those applications some only deal with off-line analysis to detect changes in statistical behavior during a predefined frame of time or space; on the other, many applications require on-line (i.e. real time) detection of such changes in a way that minimizes the delay between the time change occurs and the time it is detected. This on-line detection problem is known as the *Quickest Detection* problem. In other words, quickest detection is to detect changes happening at unknown points in time, as rapidly as possible, while maintaining the false alarm rate (FAR) at a given level.

The formulation of the quickest detection problem is based on the assumption that the pre-change and post-change probability distributions of the data are known and the classical detection schemes invariably assume knowledge of the distributions. We call those detection procedures *parametric quickest detection*. However, in a variety of applications such as surveillance or intrusion detection systems, the assumption is hardly true. In such situations, an alternative to the existing schemes must be found. As the opposite of parametric detection, nonparametric quickest detection does not depend on the data fitting any parametrized distributions.

In this dissertation, the observations dealt with by quickest detection are referred as data stream instead of data. The reasons are: 1. the observation set is not a static dataset and the underlying process (distribution) that generates the data stream is changing over time; 2. the observations are not to be examined all at the same time but one by one by the detection scheme, observations generated after current detecting time will have no impact onto the detecting result.

1.2 Applications

In this section, we describe some intensively investigated applications where quickest detection problems occur.

1.2.1 Intrusion Detection

The idea of intrusion detection appeared in 1980 [Anderson, 1980] and an early abstract intrusion detection model was proposed in 1987 by Denning [Denning, 1987]. As a countermeasure of the vast number of security incidents that occur on a network, network intrusion detection systems (NIDSs), have become an important component in the security infrastructure. NIDSs detect suspicious activities that may compromise networks security and alert the networks administrator to respond to the threat. Based on the techniques used, NIDSs can be classified as either signature detection systems or anomaly detection systems. Signature detection recognizes an intrusion based on known intrusions or attack characteristics or signatures. It identifies intruders who are trying to break in with some known techniques. The detection decision is made based on the knowledge of the model intrusive processes and what traces the detector should find in the observed system. On the other hand, anomaly detection, which is based on the assumption that something abnormal is most likely to be an intrusion, identifies an intrusion by calculating a deviation from normal system behavior.

The main references for intrusion detection applications include [Cardenas et al., 2004], [Chang, 2002], [Kim et al., 2004], [Tartakovsky et al., 2006b] and [Tartakovsky et al., 2006a].

1.2.2 Econometrics

To monitor the current and predict the future states of economy or business processes, some leading economic indicators are used as stochastic and time-varying structural parameters to build interdependent econometric models. By detecting the turning point time of these parameters, one can better understand the future behavior of the economy.

The main references for econometrics applications include [Andersson et al., 2004], [Andersson et al., 2006], [Andreou and Ghyels, 2004], [Andrews et al., 1996], [Berkes et al., 2004], and [Brännäs and Westlund, 1980].

1.2.3 Quality Control

The first application change detection dealt with was quality control and it is still an active field of quickest detection research. Quality control is crucial in preventing product failures or catastrophic events. During the detection procedure, real-time decisions which tell the current state (*in control* or *out of control*) of a process are made sequentially in order to guarantee the products or services meet industrial or customer requirements.

The main references for quality control applications include [Girshick and Rubin, 1952], [Lai, 1995], [Basseville and Nikiforov, 1993], and [Yashchin, 1997].

1.3 Motivation

Quickest detection has received increasing attention in recent years. While the parametric approaches are quite successful in solving problems with known pre-change and post-change distributions, from a practical point of view, it is most interesting if both the pre-change and post-change distributions are unknown but it is also most challenging area in quickest detection research. Existing nonparametric methods such as rank-based approach [Gordon and Pollak, 1994] and score function based approach [Tartakovsky et al., 2006b] only work well in some predefined environment. Furthermore, in a distributed detection environment, there is virtually no working nonparametric detection procedures and it is unclear how this problem can be addressed. We need to find an effective scheme as well as its decentralized implementation which makes no assumption about the distributions and is capable of detecting a change only based on the data it examines.

1.4 State of the Art

Generally speaking, the research on quickest detection can be divided into four categories, i.e., Bayesian framework, non-Bayesian framework, nonparametric detection, and decentralized detection. In this section, we identify the current research on this topic and highlight issues exist.

1.4.1 Bayesian Formulation

Although the change detection problem was first introduced in the 1930s and early research by Page [Page, 1954,Page, 1955], and Girshick and Rubin [Girshick and Rubin, 1952] was published in the 1950s, the first precise mathematical formulation on quickest detection, i.e., the Bayesian formulation, in which the unknown change point is assumed to have a geometric prior distribution, was proposed by Kolmogorov and Shiryaev [Shiryaev, 1963] in the early 1960s. By optimizing the two performance indices, the mean detection delay and the probability of false alarm, Bayesian procedure decides the change point at the first upcrossing of a properly chosen threshold by the posterior probability of the change based on the past and current observations. Under the Bayesian framework, different cost functions for the optimization problem were also proposed such as by Pelkowitz [Pelkowitz, 1987] and Poor [Poor, 1998] and they are essentially equivalent to the Shiryaev problem [Karatzas, 2003]. The solution to Bayesian formulation can be obtained through the Shiryaev test [Shiryaev, 1963,Shiryaev, 1978] and it is proven to be optimal. However, the drawbacks of the Bayesian formulation are that its assumption of a prior on the change point is sometimes not true. For example, in intrusion detection systems, there is no statistical model for the starting time of attacks.

1.4.2 Non-Bayesian Formulation

Lorden [Lorden, 1971] proposed the first non-Bayesian (i.e., Minimax) formulation of quickest detection, in which the change point is assumed to be a non-random quantity but no prior knowledge of the change point is known. Lorden's formulation optimizes the worst case delay conditioned on previous observations while keeping the constraint of a given average mean time between false alarms. Lorden, Moustakides [Moustakides, 1986], and Ritov [Ritov, 1990] showed that Page's CUSUM algorithm [Page, 1954] provides the optimal stopping time to the Lorden problem. Pollak [Pollak, 1985] proposed another widely used formulation in which the worst case delay is conditioned on the stopping time happens after the change time. It shares the same false alarm constraint as the Lorden formulation. The Shiryaev-Roberts (SR) test [Roberts, 1966] is considered the optimal solution for the Pollak problem. Although both the Bayesian and non-Bayesian

formulations offer excellent ways to handle the quickest detection problem, they fail to deal with the situations where the distributions of pre-change and post-change are unknown or implicit.

1.4.3 Nonparametric Detection

Nonparametric detection remains one of the most challenging topics in quickest detection. The majority of published studies try to construct a sequence of statistics to replace the statistics of the CUSUM or SR procedures. We have identified three papers that represent the state of the art.

Gordon and Pollak [Gordon and Pollak, 1994] proposed a rank and sign based sequential likelihood ratio approach. They use a finite sequence of the likelihood ratio of signs and ranks of absolute values, which is called nonparametric Shiryaev-Roberts statistics (NPSR), to replace the likelihood ratios in a parametric SR procedure. Though its detection statistic has an explicit form so the computation of such statistic is feasible, the rank based method is not suitable for real-time applications because it has to estimate the induced distributions of the pre-change and post-change of the signs and ranks of the observations at each time moment, which makes the distributions variant all the time. Also, it requires symmetrical pre-change distribution which makes itself unavailable to many applications.

Tartakovsky et al. [Tartakovsky et al., 2006b] offered an adaptive quickest detection. They construct specified score functions for the changes they want to detect to replace the log-likelihood ratios in CUSUM test. For example, if the mean value of a process is under monitoring, the score function would include an estimated pre-change mean value and an adaptively estimated future mean value. This method is efficient in some applications such as detecting denial-of-service (DoS) attack in computer networks but it does not work well for small changes. Furthermore, the performance largely depends on the estimated parameters whose accuracies are hardly guaranteed.

Kifer et al. [Kifer et al., 2004] adopted the Vapnik-Chervonenkis theory to define an $\mathcal{A}$-distance which is used to measure the distance between two probability distributions inferred from the pre-change and post-change observations. They statistically guarantee

the detection of real changes in high probabilities but they do not really apply this distance measure into a quickest detection framework. Also, the $\mathcal{A}$-distance has a loose bound compared to the proposed Q-Q distance.

1.4.4 Decentralized Detection

Decentralized quickest detection problem draws increasing interest recently due to the fast development of distributed systems. With the constraints on communication bandwidth usage and power consumption, centralized schemes are no longer suitable for the detection task in a distributed environment where a distributed L-sensor (monitoring channel) system observes an L-component stochastic process. There are two main detecting scenarios [Poor and Hadjiliadis, 2009], [Tartakovsky and Veeravalli, 2008] for decentralized quickest detection. In the first scenario, each sensor sends a sequence of compressed or quantized observations to a fusion center, where a detection procedure is carried out to determine the true hypothesis. In the second scenario, detection procedure is performed at each sensor and all local decisions are sent to the fusion center for integration.

Tenney and Sandell [Tenney and Sandell, 1981] are perhaps the first to introduce the extensions from classical centralized framework to decentralized framework. They propose a decentralized (binary) hypothesis test in the local channels and provide theoretic results in decision fusion using an example of two Gaussian observations.

Veeravalli and Tartakovsky, among others [Veeravalli, 1999, Veeravalli, 2001], [Tartakovsky and Veeravalli, 2003, Tartakovsky and Kim, 2006, Tartakovsky and Polunchenko, 2008, Tartakovsky and Veeravalli, 2008] developed corresponding decentralized binary versions of classical Bayesian and minimax procedures. They also derive the optimal properties of the quantizer and propose fusion rules. However, to date, no nonparametric binary detection procedure has been successfully implemented. The binary scheme that we present in the dissertation represents the first nonparametric binary quickest detection procedure.

There are also other formulations concerning the decentralized detection. For example, Teneketzis and Varaiya [Teneketzis and Varaiya, 1984] formulated the detection of the states of a Markov chain process and the decision fusion in a dynamic programming framework. Raghavan and Veeravalli [Raghavan and Veeravalli, 2008] used a similar formulation

but assumed that each sensor's observations may change at different points in time instead of all at the same time.

1.5 Contributions

The algorithms we have developed extend the above state of the art. We have developed an effective nonparametric quickest detection procedure that overcomes the drawbacks of some existing nonparametric procedures. Our research contributions are listed as follows.

Distance Measure: The most significant contribution is the development of a novel distance measure (Q-Q distance) for distributional change using the Quantile-Quantile plot technique which compares the distributions inferred directly from data sets generated from the distributions. To date, most nonparametric distance measures are derived from estimated distributions and thus have errors embedded. The distance measure we defined avoids the process of estimating distributions and comes from the data sets directly, which is more suitable for the real world applications.

Nonparametric Detection: The second major contribution is the development of a distribution free algorithm to perform quickest detection in data stream. With the new defined distance measure replacing the likelihood ratio in the benchmark CUSUM procedure, we present in this dissertation a novel nonparametric quickest detection procedure. Our algorithm outperforms other popular nonparametric quickest detection procedures and have comparable performance to the benchmark parametric CUSUM test.

Nonparametric Decentralized Detection: The third contribution is the development of a binary nonparametric quickest detection scheme which sends binary version of observations to the fusion center where the detection procedure is carried out. We also propose to perform quickest detection locally and send all local decisions to the fusion center to aggregate the decisions. To the best of our knowledge, there has been no decentralized nonparametric quickest detection procedures successfully implemented.

Performance Guarantee: The final distribution is the study of the optimality property of our detection procedure. We show that our Q-Q distance converges to zero *almost surely* if the two distributions in question are actually identical, and we determine a lower bound on the sample size using the Dvoretzky-Kiefer-Wolfowitz Inequality [Dvoret-

zky et al., 1956].

1.6 Dissertation Organization

The remainder of this dissertation documents the details of our algorithms and the above contributions. Chapter 2 reviews existing research work relevant to this dissertation, including classical formulations, nonparametric quickest detection, and distributed quickest detection. Then we describe our nonparametric detection algorithm for quickest detection in Chapter 3. Chapter 4 gives theoretical analysis and mathematical guarantees for our nonparametric detection algorithm. We present the decentralized quickest detection procedures in Chapter 5. Chapter 6 shows all experiment results demonstrating the capabilities of our algorithms. Finally, we conclude this dissertation with a summary of accomplished and future work in Chapter 7.

Chapter 2

Quickest Detection

2.1 Problem Description

Suppose there is a single channel data stream $\mathbf{X}(n)$, $n \geq 1$, that is chosen for monitoring. At an unknown point in time λ, ($\lambda \geq 1$), a change happens and the distribution of the channel is changed. All elements in the stream are assumed to be independent and identically distributed (i.i.d) before and after the change.

Here we introduce the hypotheses

$$\begin{cases} H_\infty : \lambda = \infty & \text{the change does not occur} \\ H_k : \lambda = k & \text{the change occurs in a single channel at time } \lambda = k \end{cases} \tag{2.1}$$

Let $\mathbf{P}_\infty$ denote the probability measure when the change never happens ($\lambda = \infty$), and let $\mathbf{P}_\lambda$ denote the probability measure when the change happens at time λ. Correspondingly $\mathbf{E}_\infty$ and $\mathbf{E}_\lambda$ denote the expectations. Suppose the change happens at time k such that $\lambda = k$ and the observations $\{\mathbf{X}(1), \mathbf{X}(2), \ldots, \mathbf{X}(k-1)\}$ follow a distribution $F^{(0)}$ with a probability density function $f^{(0)}$, while the observations $\{\mathbf{X}(k), \mathbf{X}(k+1), \ldots\}$ follow a distribution $F^{(1)}$ with a probability density function $f^{(1)}$. The task is to locate the change point as early as possible, while keeping the rate of false alarm under a given level. In the quickest detection framework, the solutions of the quickest detection problem can be seen as the results of optimizing the tradeoff between two performance criteria including the *detection delay*, the time between a change occurs and it is detected, which measures the

ability of the detection scheme to fire an alarm after a change actually happens; and the *probability of false alarm*, which is related to the detection accuracy.

There are two major mathematical models, Gaussian and Minimax, for the change time λ. Formulations based on these models will be introduced below.

2.2 Bayesian Formulation and Solution

In the Bayesian formulation, the change point λ is assumed to be a random variable with a known probability distribution. Here we follow [Poor and Hadjiliadis, 2009]. Define the sequence $\{\pi_k\}$ by

$$\pi_k = \mathbf{P}(\lambda \leq k | \mathcal{F}_k^X), \quad k = 0, 1, 2, \ldots \tag{2.2}$$

where $\mathcal{F}_k^X$ is the sigma-algebra generated by $\{\mathbf{X}(1), \mathbf{X}(2), \ldots, \mathbf{X}(k)\}$. Also, define $\mathbf{P}^\pi$ as the average probability measure

$$\mathbf{P}^\pi(\bullet) = \sum_{k=0}^{\infty} \mathbf{P}_k(\bullet)\pi_k \tag{2.3}$$

and $\mathbf{E}^\pi$ denotes the corresponding expectation.

It is a natural choice to adopt the expected delay, also called average detection delay (ADD), as the performance measure of detection delay, which is:

$$ADD^\pi(\tau) = \mathbf{E}^\pi(\tau - \lambda + 1)^+ \tag{2.4}$$

where τ is the stopping time and $x^+ = \max\{0, x\}$. Similarly, as a measure of false alarm rate, the probability of false alarm (PFA) is adopted as:

$$PFA^\pi(\tau) = \mathbf{P}^\pi(\tau < \lambda) = \sum_{k=0}^{\infty} \mathbf{P}_k(\tau < k)\pi_k \tag{2.5}$$

then the optimization problem can be formed as follow [Shiryaev, 1963]:

$$\inf_{\tau \in \mathcal{T}} \{\mathbf{P}^\pi(\tau < \lambda) + c\mathbf{E}^\pi(\tau - \lambda + 1)^+\} \tag{2.6}$$

where $\mathcal{T}$ denotes the set of all possible stopping times and c is a positive constant controlling

the weight of the two performance measures.

For every stopping time τ, the following equality holds:

$$\mathbf{P}^{\pi}(\tau < \lambda) = \mathbf{E}^{\pi}\{1_{(\tau<\lambda)}\} = \mathbf{E}^{\pi}(1 - \pi_{\tau}) \tag{2.7}$$

and by [Poor and Hadjiliadis, 2009],

$$\mathbf{E}^{\pi}(\tau - \lambda + 1)^{+} = \mathbf{E}^{\pi}\left(\sum_{m=0}^{\tau} \pi_m\right) \tag{2.8}$$

then Eq. 2.6 can be rewritten as:

$$\inf_{\tau \in \mathcal{T}} \mathbf{E}^{\pi}\left\{1 - \pi_{\tau} + c\sum_{m=0}^{\tau} \pi_m\right\} \tag{2.9}$$

In [Shiryaev, 1963, Shiryaev, 1978], Shiryaev proved that given a prior distribution on the change point λ geometric, that is

$$\mathbf{P}(\lambda = k) = \begin{cases} \pi & \text{if} \quad k = 0 \\ (1-\pi)\rho(1-\rho)^{k-1} & \text{if} \quad k = 1, 2, \ldots \end{cases} \tag{2.10}$$

and an appropriately chosen threshold $\pi^* \in [0, 1]$, the optimal solution to Eq. 2.9 is:

$$\tau_{opt} = \inf\{k \geq 0 \mid \pi_k \geq \pi^*\} \tag{2.11}$$

Based on the results above, we can see the optimal stopping time is the first time when the sequence $\{\pi_k\}$ reaches a suitable threshold. There comes the Shiryaev detection procedure. Let

$$p^n = \mathbf{P}(\lambda \leq n | X(n)), \quad n = 0, 1, 2, \ldots \tag{2.12}$$

be the posterior probability that in the i-th channel a change happens before time n. The optimal detection procedure is

$$\nu(A) = \min\{n \geq 1 : \ p^n \geq A\} \tag{2.13}$$

where the threshold A should be chosen in such a way that $PFA^{\pi}(\nu) \leq \alpha$, which is the

prescribed false alarm rate, and setting $A = 1 - \alpha$ guarantees the inequality. [1]

2.3 Minimax Formulations and Solutions

In Bayesian quickest detection, the unknown change point is assumed to be a random variable with a prior distribution. However, there are many applications in which the assumption of a prior on the change point is hardly true. For example, in network intrusion detection system, there is no pre-existing statistical model for the normal online behavior and intrusions. An alternative for Bayesian method, the Minimax formulation was first proposed by Lordon [Lorden, 1971]. In the Minimax formulation, the change point λ is assumed to be a fixed but unknown non-random quantity that can be any value in the positive integers.

Consider a measurable space $(\Omega, \mathcal{F})$ with sample space Ω and a σ-field $\mathcal{F}$ of events [2]. The stopping variable τ is considered for all possible stopping with respect to the observed data stream. The event $\{\tau = n\}$, which denotes the alarm triggered after observing $\{\mathbf{X}(1), \mathbf{X}(2), \ldots, \mathbf{X}(n)\}$, is determined by the smallest σ-field $\{\mathcal{F}_n; n \geq 0\}$ generated by $\{\mathbf{X}(1), \mathbf{X}(2), \ldots, \mathbf{X}(n)\}$. Let $\mathbf{P}_\lambda, (\lambda = 1, 2, \ldots)$ be the distribution of the data stream $\{\mathbf{X}(1), \mathbf{X}(2), \ldots\}$ under which $\mathbf{X}(\lambda)$ is the first term with distribution function $F^{(1)}$ and $\mathbf{E}_\lambda$ be the expectation under $\mathbf{P}_\lambda$.

In the Minimax formulation, instead of an average expected value, which requires a distribution of the time of change, the worst case detection delay is adopted as a measure of the detection lag, conditioned on the observations before the change time λ, which is:

$$WDD(\tau) = \sup_{\lambda \geq 1} \operatorname{ess\,sup} \mathbf{E}_\lambda \{(\tau - \lambda + 1)^+ | \mathbf{X}(1), \ldots, \mathbf{X}(\lambda - 1)\} \tag{2.14}$$

Note that $WDD_\lambda(\tau) = \operatorname{ess\,sup} \mathbf{E}_\lambda \{(\tau - \lambda + 1)^+ | \mathbf{X}(1), \ldots, \mathbf{X}(\lambda - 1)\}$ is the worst case average delay under $\mathbf{P}_\lambda$, [3] and the worst case is taken over all observations before time λ.

[1] In defining the detection procedure stopping time in Eq. 2.13, we use min instead of inf as in Eq. 2.11. The difference is the minimum (maximum) of a set must be in the set while the infimum (supremum) does not have to be in the set.

[2] In mathematics, a σ-algebra over a set $\mathbf{X}$ is a nonempty collection Σ of subsets of $\mathbf{X}$ (including $\mathbf{X}$ itself) that is closed under complementation and countable unions of its members. The pair $(\mathbf{X}, \Sigma)$ is called σ-field.

[3] In defining the worst case detection delay in Eq. 2.14, we use the essential supremum of the random

Meanwhile, the probability of false alarm can be measured by the average run length (ARL) to false alarm or the false alarm rate (FAR) :

$$ARL(\tau) = \mathbf{E}_{\infty}(\tau) \tag{2.15}$$

and

$$FAR = 1/ARL = 1/\mathbf{E}_{\infty}(\tau) \tag{2.16}$$

so the objective function can be given by:

$$\inf_{\tau \in \mathcal{T}} WDD(\tau) \quad \text{subject to} \quad ARL(\tau) \geq \gamma, \tag{2.17}$$

where γ is a prescribed positive and finite constant. Lorden used Page's CUSUM test [Page, 1954] to examine the solution to Eq. 2.17 but its optimality was not fully understood until Moustakides [Moustakides, 1986] proved that for $h \geq 0$, the CUSUM stopping time

$$\tau_c(h) = \min\{n \geq 1 \mid S_n \geq h\} \tag{2.18}$$

where

$$S_n = \max_{1 \leq k \leq n} Z^k(n) \tag{2.19}$$

is optimal for all finite γ and

$$Z^k(n) = \sum_{j=k}^{n} \ln \frac{f^{(1)}(\mathbf{X}(j))}{f^{(0)}(\mathbf{X}(j))} \tag{2.20}$$

is the log-likelihood ratio (LLR) between the hypotheses H_{∞} and H_k.

Eq. 2.19 can be regarded as a maximum likelihood ratio of the unknown change point and for computational purposes, S_n can be written in a recursive form

$$S_n = \left(S_{n-1} + \ln \frac{f^{(1)}(\mathbf{X}(n))}{f^{(0)}(\mathbf{X}(n))} \right)^{+} \tag{2.21}$$

variable. The essential supremum of a random variable is related to the notion of supremun but often deals with statements which are not valid everywhere. As defined in [Poor and Hadjiliadis, 2009], it is the least upper bound of the set of constants that bound the random variable with probability one.

with $S_0 = 0$.

Another Minimax formulation was proposed by Pollak [Pollak, 1985]. In this formulation, worst case average detection delay (WADD) conditioned on the stopping time τ happens after the change time λ, is used as the measure of detection delay. That is

$$WADD(\tau) = \sup_{1 \leq \lambda < \infty} \mathbf{E}_\lambda(\tau - \lambda \mid \tau \geq \lambda) \tag{2.22}$$

Both formulations have the same measure of FAR as Eq. 2.15 stated. so the objective function is

$$\inf_{\tau \in \mathcal{T}} WADD(\tau) \quad \text{subject to} \quad ARL(\tau) \geq \gamma, \tag{2.23}$$

The Shiryaev-Roberts (SR) detection procedure is optimal with respect to the Minimax expected detection delay, and the stopping time is given as

$$\tau_{sr}(h) = \min\{n \geq 1 \mid R_n \geq h\} \tag{2.24}$$

where

$$R_n = \sum_{m=1}^{n} \frac{f^{(1)}(\mathbf{X}(m), \mathbf{X}(m+1), \ldots, \mathbf{X}(n))}{f^{(0)}(\mathbf{X}(m), \mathbf{X}(m+1), \ldots, \mathbf{X}(n))} = \sum_{m=1}^{n} \prod_{n=m}^{n} \frac{f^{(1)}(\mathbf{X}(n))}{f^{(0)}(\mathbf{X}(n))} \tag{2.25}$$

Or it can be written recursively as

$$R_n = (1 + R_{n-1}) \frac{f^{(1)}(\mathbf{X}(n))}{f^{(0)}(\mathbf{X}(n))} \tag{2.26}$$

with with $R_0 = 0$.

As we mentioned before, there are two performance indices: the detection delay and the false alarm rate. Both indices are related to the choice of threshold in a detection procedure.

Let

$$I = \int \ln \frac{f^{(1)}(x)}{f^{(0)}(x)} f^{(1)}(x) dx \tag{2.27}$$

be the Kullback-Leibler divergence between the distributions $f^{(1)}(x)$ and $f^{(0)}(x)$. For both Minimax formulations, the detection delays

$$D(\tau_h^c) \sim \frac{h}{I} \quad \text{and} \quad WADD(\tau_h^{sr}) \sim \frac{h}{I} \tag{2.28}$$

are linear to the threshold.

Choosing $h = \ln \gamma$ guarantees

$$FAR(\tau_h^c) \leq 1/\gamma \quad \text{and} \quad FAR(\tau_h^{sr}) \leq 1/\gamma \tag{2.29}$$

See [Lorden, 1971], [Pollak, 1985], [Tartakovsky et al., 2006b] and [Tartakovsky and Veeravalli, 2008].

2.4 Nonparametric Quickest Detection

From a practical point of view, it is most interesting if both the pre-change and post-change distributions are unknown but it is also most challenging in quickest detection research. In some applications the pre-change distribution can be estimated but is difficult to be accurately expressed in an explicit form which makes it impossible to be applied to those classical (parametric) detection procedures previously mentioned. To address this problem, distribution free methods have been proposed during the research on nonparametric quickest detection procedures.

Gordon and Pollak [Gordon and Pollak, 1994] proposed a rank and sign based likelihood ratio sequential detection approach. Consider a sequence of independent samples $X(1), X(2), \ldots, X(n)$, recall the Shiryaev-Roberts procedure, we rewrite Eq. 2.25 as

$$R_n = \sum_{i=1}^{n} \frac{f^{(1)}(X(i), X(i+1), \ldots, X(n))}{f^{(0)}(X(i), X(i+1), \ldots, X(n))} \tag{2.30}$$

Define

$$\rho(i,n) = \sum_{j=1}^{n} I_{\{|X_j| \leq |X_i|\}} \quad \text{and} \quad \sigma_i = I_{\{X(i)>0\}} \tag{2.31}$$

where I is the indicator function. $\rho(i,n)$ denotes the rank of the absolute value of the i-th observation among the first n samples and σ_i gives the sign of the i-th observation. Let

$$Y(i) = (\rho(i,n), \sigma_i) \tag{2.32}$$

so we get $Y(1), Y(2), \ldots, Y(n)$ as first n observations' signs and ranks of absolute values.

By replacing the likelihood ratio using nonparametric likelihood ratio based on signs and ranked absolute values, we can get the nonparametric analogs to the SR statistics [Gordon and Pollak, 1994] by

$$R_n = \sum_{i=1}^{n} \frac{p^{(1)}(Y(i), Y(i+1), \ldots, Y(n))}{p^{(0)}(Y(i), Y(i+1), \ldots, Y(n))} \tag{2.33}$$

where $p^{(0)}$ and $p^{(1)}$ represent the pre-change and post-change distributions of the signs and ranks of absolute values. By choosing appropriate parameters and using the result of Savage [Savage, 1956], both $p^{(1)}(Y(i), Y(i+1), \ldots, Y(n))$ and $p^{(0)}(Y(i), Y(i+1), \ldots, Y(n))$ have explicit forms so Eq. 2.33 is well defined.

However, in online quickest detection applications, those observations to be ranked might not be available at the time when the detection procedure is taken place so this method only works well in an off-line environment because it requires a large number of observations to be ranked.

Tartakovsky [Tartakovsky et al., 2006b] suggested replacing the log-likelihood ratio with specified score function g_i in a CUSUM test to monitor specified parameters such as mean or variance of unknown distributions since the log-likelihood ratio in Eq. 2.19 cannot be calculated. Score functions can be chosen in many ways depending on what we plan to detect. For example, in the case of detecting changes in mean values, one could use the score function below

$$g_{i,m} = X_i(m) - \mu_i - \varepsilon\theta_i \tag{2.34}$$

where μ_i is the estimated pre-change mean, θ_i is the estimated post-change mean, $\varepsilon \in (0, 1)$ is the tuning parameter. And without loss of generality, we let $\mu_i < \theta_i$. Sometimes θ_i is difficult to accurately estimate in an online environment, it is better to treat it as a positive constant. That is, $\varepsilon\theta_i = c_i$. Thus the equivalent recursive representation can be written as

$$S_n^g = \left(S_{n-1}^g + X_i(n) - \mu_i - c_i\right)^+ \tag{2.35}$$

with $S_0 = 0$. And the detection procedure is

$$\tau_g(h) = \min\{n \geq 1 \mid S_n^g \geq h\} \tag{2.36}$$

The performance indices, average detection delay and false alarm rate, can be approximated as

$$ADD(\tau_h^c) \sim \frac{h}{q_i} \quad \text{and} \quad FAR(\tau_h^c) \leq C_1 e^{-C_2 h} \tag{2.37}$$

where q_i, C_1 and C_2 are constants. These constants are difficult to compute so the ADD and FAR need to be evaluated experimentally.

This method is efficient in some applications including detecting intrusions such as DoS attacks. However, it assumes the monitored parameters have detectable differences between the pre- and post-change distributions which might not always be true. Also, the performance degrades when the estimated parameters (θ_i in detecting mean or σ_i in detecting variance) vary significantly during the process.

2.5 Decentralized Quickest Detection

Decentralized quickest detection problem draws increasing interest recently due to the fast development of distributed systems. With the constraints on communication bandwidth usage and power consumption, centralized schemes are no longer suitable for the detection task in a distributed environment where a distributed L-sensor (monitoring channel) system observe a L-component stochastic process. There are two main detecting scenarios [Poor and Hadjiliadis, 2009], [Tartakovsky and Veeravalli, 2008] for decentralized quickest detection. In the first scenario, each sensor sends a sequence of compressed or quantized observations to a fusion center, where a detection procedure is carried out to determine the true hypothesis. In the second scenario, detection procedure is performed at each sensor and all local decisions are sent to the fusion center for combining.

2.5.1 Decentralized Detection with Quantized Observations

There are a number of information structures for the decentralized configuration depending on how feedback and sensed information is used at the sensors. In this section, we only con-

sider the simplest information structure where only current observation $\mathbf{X}_i(n)$ is available at time n. Since all observations by the i-th sensor are assumed i.i.d., it is natural to use certain stationary quantizer ϕ_i, which does not depend on n to quantize the observations. So we can write the quantization as

$$\mathbf{B}_i(n) = \phi_i(\mathbf{X}_i(n)). \tag{2.38}$$

Let $p_i^{(j)}$ denote the probability mass function induced on $\mathbf{B}_i(n)$ when the observation $\mathbf{X}_i(n)$ is distributed as $f_i^{(j)}$, $j = 0, 1$. With the induced probability mass functions, the log-likelihood ratio between the hypotheses $H_{i,k}$ and $H_{i,\infty}$ is given by

$$Z_i^q(k,n) = \sum_{j=k}^{n} \ln \frac{p^{(1)}(\mathbf{B}(j))}{p^{(0)}(\mathbf{B}(j))}, \tag{2.39}$$

and the CUSUM and SR procedures can be executed. The quantized version of detection statistics are given by

$$S^q(n) = \left(S^q(n-1) + \sum_{i=1}^{L} \ln \frac{p_i^{(1)}(\mathbf{B}_i(n))}{p_i^{(0)}(\mathbf{B}_i(n))} \right)^+ \tag{2.40}$$

with $S^q(0) = 0$, and

$$R^q(n) = [(1 + R^q(n-1)] \exp \left\{ \sum_{i=1}^{L} \frac{p_i^{(1)}(\mathbf{B}_i(n))}{p_i^{(0)}(\mathbf{B}_i(n))} \right\} \tag{2.41}$$

with $R^q(0) = 0$.

The stopping times of the CUSUM and SR detection procedures at the fusion center are, respectively, given by

$$\tau_c^q = \inf\{n \geq q \mid S^q(n) \geq h\} \tag{2.42}$$

and

$$\tau_{sr}^q = \inf\{n \geq 1 \mid R^q(n) \geq h\} \tag{2.43}$$

Let I_i^q denote the K-L divergence between the induced $p_i^{(1)}$ and $p_i^{(0)}$ in the i-th sensor,

and let $I_{tot}^q = \sum_{i=1}^{L} I_i^q$ be the total K-L divergence through all L sensors. It is asymptotically optimum (as $\gamma \to \infty$) for all sensors at time n if the quantization process maximizes I_{tot}^q and

$$ADD(\tau_c^q) \sim ADD(\tau_{sr}^q) \sim \frac{\log \gamma}{I_{tot}^q}, \text{ as } \gamma \to \infty. \tag{2.44}$$

2.5.2 Decentralized Detection with Local Decisions

In this section, we introduce several detection schemes that perform local detection in the sensors and transmit only the binary local decisions to the fusion center for the global decision.

We follow Veeravalli and Tartakovsky and others [Veeravalli, 1999, Veeravalli, 2001], [Tartakovsky and Veeravalli, 2003,Tartakovsky and Kim, 2006,Tartakovsky and Polunchenko, 2008, Tartakovsky and Veeravalli, 2008].

2.5.2.1 Minimax Setting

In the Minimax setting, we can use CUSUM (or SR) tests at the sensors. Recall CUSUM statistic in the i-th sensor

$$S_i(n) = \left(S_i(n-1) + \ln \frac{f_i^{(1)}(\mathbf{X}_i(n))}{f_i^{(0)}(\mathbf{X}_i(n))} \right)^+ \tag{2.45}$$

with $S_i(0) = 0$. And let the binary decisions

$$U_i(n) = \begin{cases} 1 & S_i(n) \geq \omega_i h \\ 0 & \text{otherwise} \end{cases} \tag{2.46}$$

where ω_i is the weight and h is the threshold. The local stopping time in the i-th sensor is

$$\tau_i(h) = \inf\{n \geq 1 \mid S_i(n) \geq \omega_i h\} \tag{2.47}$$

There are three popular fusion rules. The first fusion rule is defined as

$$\tau_{all}(h) = \min\{n \geq 1 \mid \min_{1 \leq i \leq N} (S_i(n)/\omega_i) \geq h\} \tag{2.48}$$

In this fusion scheme, binary local decisions are sent to the center and the global decision

in favor of the hypothesis of change is reached at the first time when $U_i(n) = 1$ for all sensors. Mei [Mei, 2005] shows that τ_{all} is globally asymptotically optimal if $h = \ln \gamma$ and $\gamma \to \infty$.

The second fusion rule is defined as

$$\tau_{max} = \max_{1 \le i \le N} \tau_i \tag{2.49}$$

which means the global decision in favor of the hypothesis of change is reached after all local decisions support the hypothesis of change.

The third fusion rule is defined as

$$\tau_{min} = \min_{1 \le i \le N} \tau_i \tag{2.50}$$

which means the global decision of change is declared at the first time any local decision is in favor of the change.

2.5.2.2 Bayesian Setting

In the Bayesian setting, we consider the SR statistic in the i-th sensor

$$R_i(n) = [(1 + R_i(n-1)] \exp \left\{ \frac{f_i^{(1)}(\mathbf{X}_i(n))}{f_i^{(0)}(\mathbf{X}_i(n))} \right\} \tag{2.51}$$

with $R_i(0) = 0$. And the local stopping time in the i-th sensor is

$$\hat{\tau}_i(h) = \inf\{n \ge 1 \mid R_i(n) \ge \omega_i h\}. \tag{2.52}$$

Similarly, we can define three fusion procedures as in the Minimax setting, $\hat{\tau}_{all}$, $\hat{\tau}_{max}$, and $\hat{\tau}_{min}$. Unlike in the Minimax setting where only τ_{all} has the first order asymptotic performance, in the Bayesian setting $\hat{\tau}_{all}$ and $\hat{\tau}_{max}$ are both globally asymptotically optimum [Tartakovsky and Veeravalli, 2008].

Chapter 3

Nonparametric Quickest Detection

In this chapter, a nonparametric method based on Quantile-Quantile (Q-Q) plot is proposed here. This method requires no prior assumptions on the nature of the underlying distributions that generate the data stream except for keeping the i.i.d. assumption. In our nonparametric approach, we are not interested in what the real underlying distributions are but how big the difference is between the distributions.

3.1 Quantile-Quantile Plot

The Quantile-Quantile plot, or Q-Q plot, is a technique for comparing the distributions inferred directly from two data sets generated from two distributions. It draws the quantiles of the first data set against the quantiles of the second data set. For any probability distribution function F, the associated quantile function Q is essentially the inverse of F [Gilchrist, 2000], which is defined as

$$Q(t) = F^{-1}(t) = \inf\{x : F(x) \geq t\}, 0 < t < 1. \tag{3.1}$$

Suppose $\mathbf{x}$ and $\mathbf{y}$ are two batches of observations which do not need to have the same size. As shown in Figure 3.1, there are two cumulative distribution functions CDF_x and CDF_y. At any ordinate value t there are two corresponding quantiles $Q_x(t)$ and $Q_y(t)$. A Q-Q plot of $\mathbf{x}$ and $\mathbf{y}$ is actually a plot of $Q_y(t)$ *vs.* $Q_x(t)$ for an increasing t. "Q-Q plot tends to emphasize the comparative structure in the tails and to blur the distinctions in

the middle where the density are high" [Wilk and Gnanadesikan, 1968]. The reason is the quantile ($Q(t)$) is a rapidly changing function of t when the density is low (in the tails) and a slowly changing function of t when the density is high (in the middle).

The Q-Q plot can be generated following the procedures below:

1. choose the smaller sample size of $\mathbf{x}$ and $\mathbf{y}$ as the number of quantiles,
2. sort and percentilize $\mathbf{x}$ and $\mathbf{y}$,
3. plot x-quantiles *vs.* y-quantiles,
4. draw the reference line based on the 25% and 75% quantiles.

Since in our work we are only interested in finding if two distributions are identical, we can always set the reference line as $y = x$. Examples of Q-Q plot with respect to different combinations of two distributions are given in Figures 3.2 to 3.5. The 45° reference line is also plotted for comparison purpose. If the two batches of data come from a population with the same distribution family, the points should fall approximately along a straight line, as shown in Figure 3.2 where both distributions are Gaussian but with different means and variances. When two streams are identically distributed, the Q-Q plot is roughly a straight line with a slope 1, as shown in Figure 3.3. Figure 3.4 and Figure 3.5 show the plots of standard Gaussian vs. Uniform and Gamma distributions, respectively. From these figures we can see that the greater the plot departures from the reference line, the greater the chance that the two data sets come from populations with different distributions.

3.1.1 Q-Q Distance

In order to quantify the difference between two distributions, that is, the distance between the Q-Q plot and the reference line, we define the *Q-Q distance*. Consider two data sequences v_0 and v_1 of size s, respectively. On their Q-Q plot, for the j-th point, $(Q_{v_1}(\frac{j}{s}), Q_{v_0}(\frac{j}{s}))$, the distance from this point to the 45° reference line is $\frac{\sqrt{2}}{2} \cdot |Q_{v_1}(\frac{j}{s}) - Q_{v_0}(\frac{j}{s})|$, as shown in Figure 3.6. The Q-Q distance between these two finite sequences can thus be defined as

$$d_{qq}(v_0, v_1) = \frac{1}{s} \cdot \sum_{j=1}^{s} \frac{\sqrt{2}}{2} \left| Q_{v_1}(\frac{j}{s}) - Q_{v_0}(\frac{j}{s}) \right| \tag{3.2}$$

where Q_{v_i}, $i \in \{0, 1\}$ is the quantile of data stream v_i.

In the following, we provide the proof that d_{qq} is indeed a distance metric because it

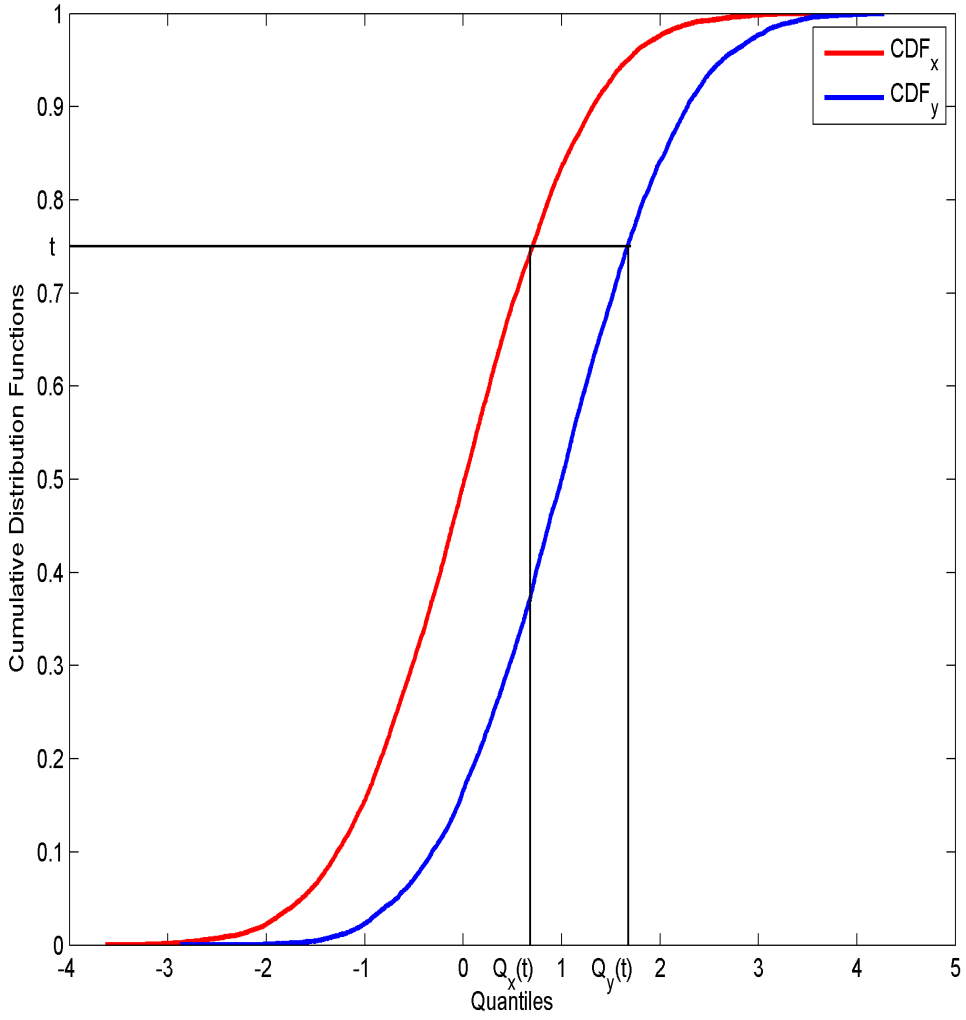

Figure 3.1: Illustration for Q-Q plot: Cumulative distribution functions with quantiles

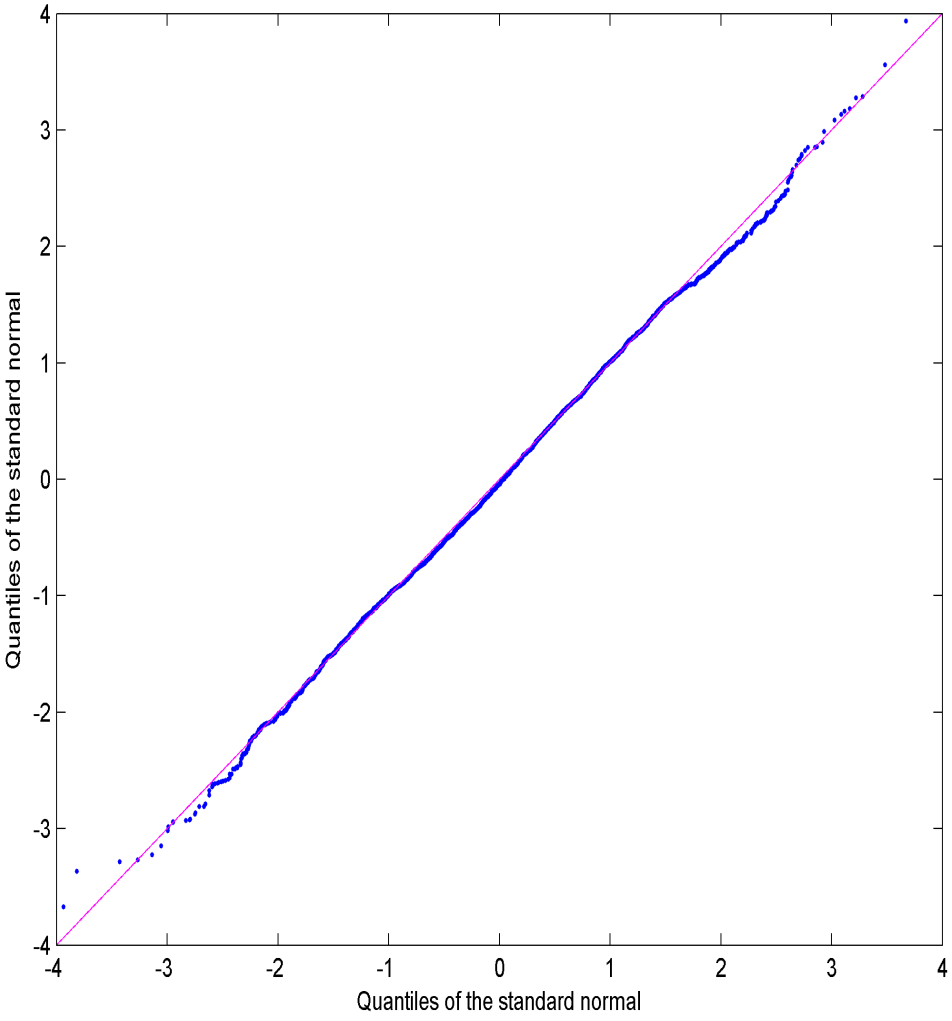

Figure 3.2: Q-Q plot: Standard normal, $N(0,1)$, *vs.* Standard Normal $N(0,1)$

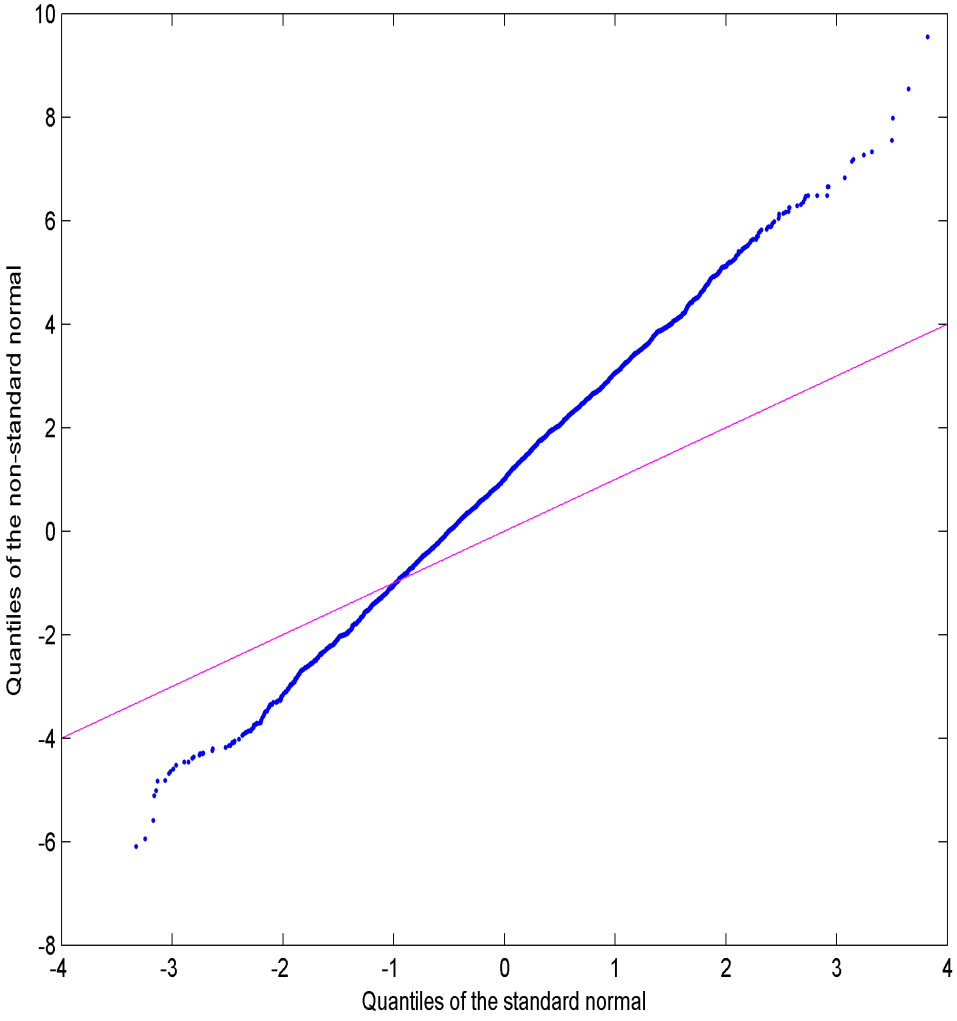

Figure 3.3: Q-Q plot: Standard normal, $N(0,1)$, *vs.* Non-Standard normal, $N(0,2)$

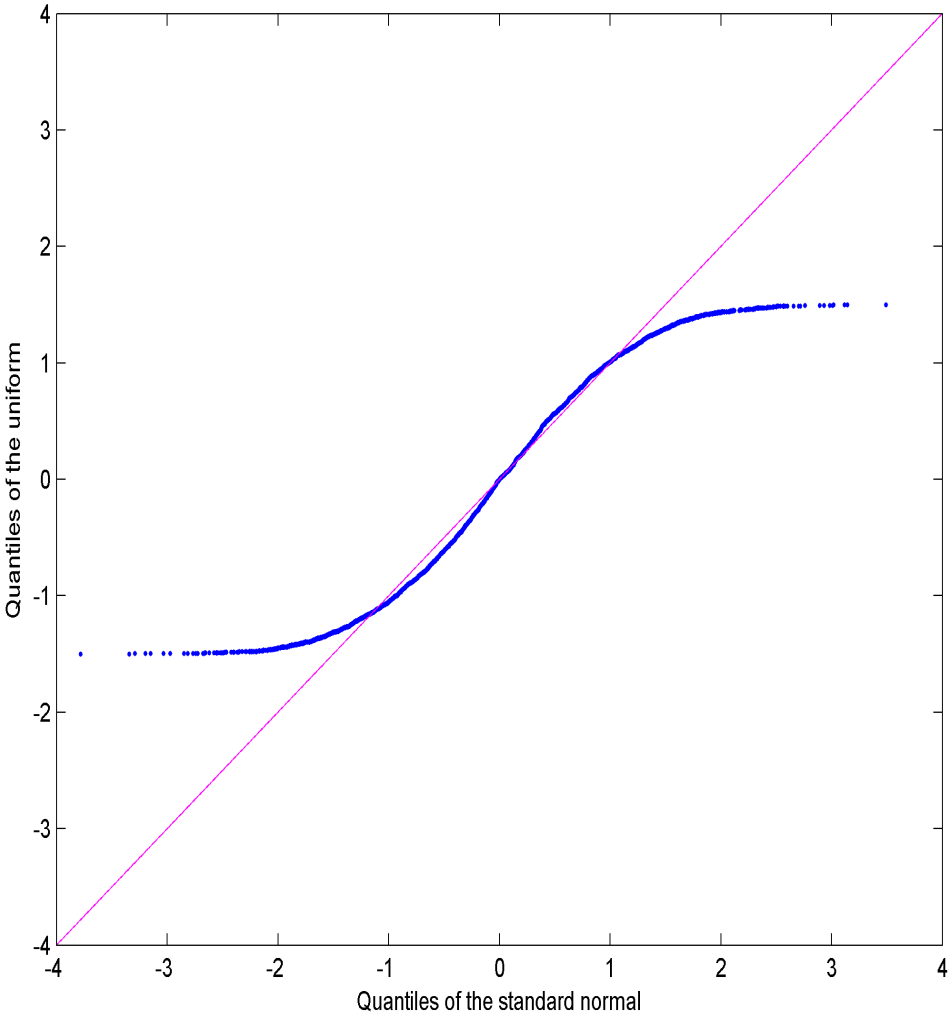

Figure 3.4: Q-Q plot: Standard normal $N(0,1)$, *vs.* Uniform, $U(-1.5, 1.5)$

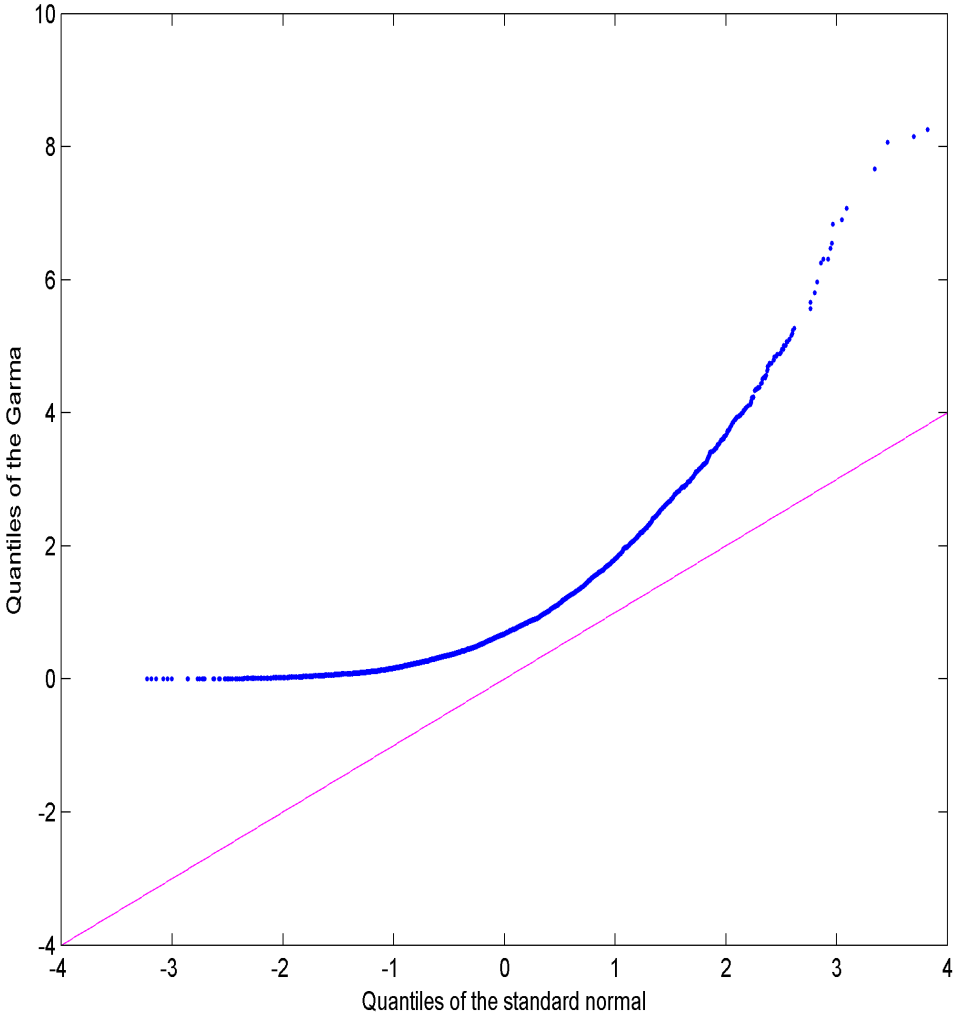

Figure 3.5: Q-Q plot: Standard normal $N(0,1)$, *vs.* $Gamma(1,1)$

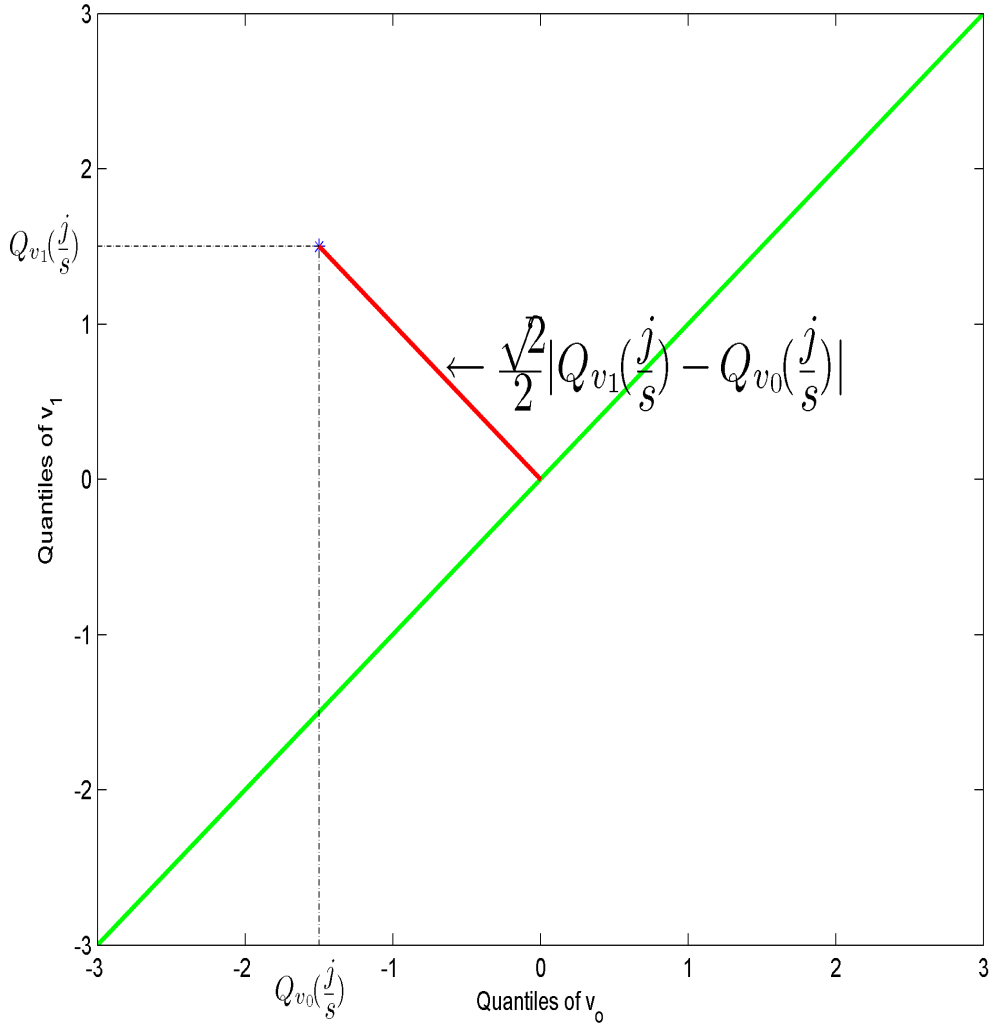

Figure 3.6: Demonstration of the Q-Q distance

satisfies the three conditions, i.e.,

- It is nonnegative: $d_{qq}(v_0, v_1) \geq 0$ and the equality holds if and only if $v_0 = v_1$.
- It is symmetric. $d_{qq}(v_0, v_1) = d_{qq}(v_1, v_0)$.
- It follows the triangle inequality. $d_{qq}(v_0, v_2) \leq d_{qq}(v_0, v_1) + d_{qq}(v_1, v_2)$

The first two conditions are easy to verify by the definition in Equation 3.2. Let us look at the condition of triangle inequality.

$$\begin{aligned} d_{qq}(v_0, v_1) + d_{qq}(v_1, v_2) &= \frac{1}{s} \cdot \sum_{j=1}^{s} \frac{\sqrt{2}}{2} \left| Q_{v_1}(\frac{j}{s}) - Q_{v_0}(\frac{j}{s}) \right| + \frac{1}{s} \cdot \sum_{j=1}^{s} \frac{\sqrt{2}}{2} \left| Q_{v_2}(\frac{j}{s}) - Q_{v_1}(\frac{j}{s}) \right| \\ &= \frac{1}{s} \cdot \sum_{j=1}^{s} \frac{\sqrt{2}}{2} \left\{ \left| Q_{v_1}(\frac{j}{s}) - Q_{v_0}(\frac{j}{s}) \right| + \left| Q_{v_2}(\frac{j}{s}) - Q_{v_1}(\frac{j}{s}) \right| \right\} \\ &\geq \frac{1}{s} \cdot \sum_{j=1}^{s} \frac{\sqrt{2}}{2} \left\{ \left| Q_{v_1}(\frac{j}{s}) - Q_{v_0}(\frac{j}{s}) + Q_{v_2}(\frac{j}{s}) - Q_{v_1}(\frac{j}{s}) \right| \right\} \\ &= d_{qq}(v_0, v_2) \end{aligned} \tag{3.3}$$

There are also other ways to define the Q-Q distance. For example, using the L2-norm instead of the L1-norm:

$$\hat{d}_{qq}(v_0, v_1) = \frac{\sqrt{2}}{2s} \cdot \sqrt{\sum_{j=1}^{s} \left(Q_{v_1}(\frac{j}{s}) - Q_{v_0}(\frac{j}{s}) \right)^2} \tag{3.4}$$

The L2-norm distance can be interpreted as a least square approach measuring the error between two quantile functions. Similarly, the Lp-norm can also be defined. Both d_{qq} and $\hat{d}_{qq}$ have almost the same performance in detecting changes. In the rest of this dissertation, we only use the L1-norm distance as the distance measure.

3.1.2 The Detection Algorithm

In this section we describe our Q-Q distance based nonparametric detection algorithm for quickest detection in data stream. Let f_1 and f_0 denote the densities when a change

happens and when there is no change. When both f_1 and f_0 are unknown, the LLR

$$Z^k(n) = \sum_{j=k}^{n} \ln \frac{f^{(1)}(\mathbf{X}(j))}{f^{(0)}(\mathbf{X}(j))}$$

and so the CUSUM's statistic

$$S_n = \max_{1 \le k \le n} Z^k(n)$$

are also unknown. We want to replace the detection statistic with an appropriate distance function which remains close to zero in no change situation and starts drifting upward until it crosses a threshold after a change happens. The Q-Q distance previously defined is perfect for this role. We will provide theoretical proof to justify this statement in the next chapter.

As a result, the stopping time of the detection procedure can be defined as

$$\tau_{qq}(h) = \min\{n \ge 1 \mid d_{qq} \ge h\} \tag{3.5}$$

where n represents time and h is the prescribed threshold. The proposed algorithm is described in Algorithm 1.

Algorithm 1 Q-Q Distance Based Quickest Detection

```
WindowOne ← first s_0 observations
WindowTwo ← first s_1 observations
while There are incoming observations do
    Slide WindowTwo forward by 1 sample
    Draw Q-Q plot of WindowOne and WindowTwo
    Calculate the distance d_qq between the plot and the 45° reference line
    if d_qq > h then
        Alarm triggered
        Break
    end if
end while
```

The Q-Q distance based nonparametric detection algorithm transforms the problem from sequential detection into the problem of comparing two static sequences. The detection procedure involves creating two windows of fixed sizes, where Window One is stationary serving as a reference to the pre-change distribution, Window Two is a moving window

always containing the latest observations. At any moment a new observation appears on the data stream, Window Two is slided one step forward and the Q-Q distance of these two windows is then updated. Whenever the distance reaches the threshold we fire an alarm. Figure 3.7 shows the detection statistic for a particular run of a simulated change.

Window One in most cases is fixed so it always maintains the same reference as the pre-change distribution. It also can be a moving window. For example, when the pre-change distribution is slowly time-varying and the task is to detect dramatic change from recent observations. In that case, Window One has to be moving forward to catch up with recent distribution. Since the estimation of pre-change distribution is relatively easy even in a nonparametric detection application, one can remove outliers, if there is any, from Window One to keep the reference accurate.

Another crucial parameter in the algorithm is the size of Window Two. Large size window can detect small changes but tends to have large detection delay and small size window can have fast detection but the false alarm rate may be higher. In the next chapter, we will give a lower bound for window size selection.

3.2 Summary

In this chapter, we have examined the problem of nonparametric quickest detection where the pre-change and post-change distributions are not known exactly. By defining and using a novel distance measure we were able to construct a detection procedure which merely relies on its observations.

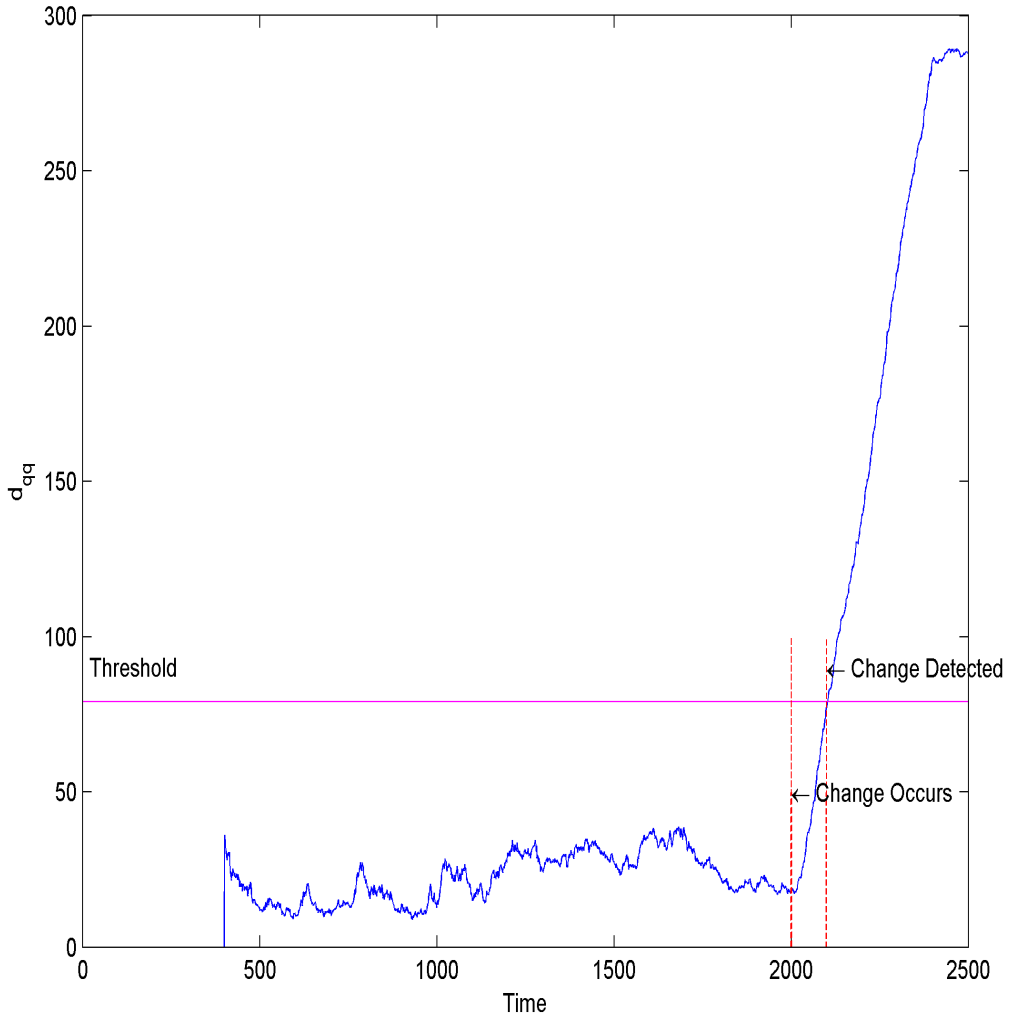

Figure 3.7: An illustration of the behavior of the detection statistic for one particular run of a simulated $N(0,1)$ to $U(0,1)$ change

Chapter 4

Performance Guarantee

In Chapter 3, we considered a nonparametric quickest detection procedure based on Q-Q distance. This procedure is essentially a two-sample comparison. Our goal is to design an algorithm which decides whether these two distributions where the two samples are drawn, are the same. In other words, we try to find the distributional difference between two finite data sequences. In this chapter, we theoretically analyze of the Q-Q distance and provide performance guarantee for the Q-Q distance based detection algorithm.

Suppose two datasets v_0 and v_1 with size m, are generated by two probability distributions $F^{(0)}$ and $F^{(1)}$. Based on the samples from v_0 and v_1, can we decide whether the two datasets were generated by the same distribution, i.e., $F^{(0)} = F^{(1)}$? Or, is it the case that $F^{(0)} \neq F^{(1)}$?

Usually, in a parametric framework, the distance between $F^{(0)}$ and $F^{(1)}$ can be calculated using the Kullback-Leibler (K-L) divergence (distance)

$$D_{kl}(f^{(0)} \| f^{(1)}) = \sum_j f^{(0)}(j) \log \frac{f^{(0)}(j)}{f^{(1)}(j)} \tag{4.1}$$

where $f^{(0)}$ and $f^{(1)}$ are the corresponding probability density functions, or the total variation which is defined by

$$\eta = \sum_x \left| F^{(0)}(x) - F^{(1)}(x) \right|. \tag{4.2}$$

However, without the knowledge of the distributions, both the K-L distance and total variation are nowhere to be found. A distance measure solely based on the finite data

samples has to be defined and this distance measure must truly quantify the discrepancy between those two underlying distributions.

We have defined the Q-Q distance, the distributional distance between two finite data sequences, as

$$d_{qq}(v_0, v_1) = \frac{1}{m} \cdot \sum_{j=1}^{m} \frac{\sqrt{2}}{2} \left| Q_{v_1}(\frac{j}{m}) - Q_{v_0}(\frac{j}{m}) \right| \tag{4.3}$$

where Q_{v_i}, $i \in \{0, 1\}$ is the quantile of data stream v_i.

The analytical study of the proposed distance measure is two-folded. On one hand, we expect d_{qq} to be very small when v_0 and v_1 are from the same distribution and if they are from two different distributions, d_{qq} should be larger than some positive number with high probability. On the other hand, we intend to determine certain lower bound on the sample size such that there is a large enough number of elements in the sample to ensure the detectable discrepancy.

Das and Resnick [Das and Resnick, 2008] applied Fell topology on closed sets to show that Q-Q plot converges to the 45° straight line in probability when the two distributions are the same but they did not provide the convergence rate as the number of samples grow.

We first introduce some technical preliminaries.

4.1 Preliminaries

Our basic tool for sample based estimation of the Q-Q distance between probability distributions is based on the Dvoretzky-Kiefer-Wolfowitz Inequality [Dvoretzky et al., 1956], [Massart, 1990] but we will also introduce a looser bound based on Vapnik-Chervonenkis Theory [Vapnik and Chervonenkis, 1971], [Vapnik, 1998]. Let us first establish some basic definitions.

Definition 4.1.1. A distribution function F is a monotone non-decreasing and right continuous function on $\mathbb{R}$, with

$$\lim_{x \to -\infty} F(x) = 0$$

and

$$\lim_{x \to +\infty} F(x) = 1.$$

Definition 4.1.2. For any distribution function $F(x)$, the quantile function $Q(t)$ is the inverse of F, which is

$$Q(t) = F^{-1}(t) = \inf\{x : F(x) \geq t\}, 0 < t < 1. \tag{4.4}$$

Definition 4.1.3. Let $X_1, X_2, \ldots, X_m$, $m \geq 1$, be i.i.d. real random variables with distribution F. The empirical distribution function F_m is a step function defined as

$$F_m(x) = \frac{1}{m} \sum_{i=1}^{m} I(X_i \leq x). \tag{4.5}$$

where I is the indicator function.

Definition 4.1.4. If we re-order $X_1, X_2, \ldots, X_m$ such that

$$X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(m)},$$

the empirical quantile function $Q_m(t)$ can then be defined as

$$Q_m(t) = \begin{cases} X_{(j)} & \text{if} \quad \frac{j-1}{m} < t \leq \frac{j}{m}, \quad j = 1, 2, \ldots, m \\ X_{(1)} & \text{if} \quad t = 0. \end{cases} \tag{4.6}$$

and obviously $Q_m(t) = F_m^{-1}(t)$.

The following theorem, known as *Convergence in Quantile*, states the equivalence of convergence in distribution and convergence in quantile.

Theorem 4.1.5. [Shorack, 2000] *Let Q be the quantile function associated with a distribution function F, and Q_m be the empirical quantile function associated with the empirical distribution function F_m. F_m converges to F in probability, that is $F_m \to F$, if and only if $Q_m \to Q$.*

The proof of this theorem is attached in the Appendix.

Theorem 4.1.6. (Glivenko-Cantelli)

$$\sup_{x \in \mathbb{R}} |F_m(x) - F(x)| \xrightarrow{a.s.} 0 \tag{4.7}$$

From Theorem 4.1.5 we know that convergence in distribution is equal to convergence in quantile. By the Glivenko-Cantelli theorem, which determines the asymptotic behavior of the empirical distribution function as the number of i.i.d. observations grows, the empirical distribution function is almost surely convergent to the true distribution for every x, so the empirical quantile function will also almost surely converge [1] to the true quantile function.

4.2 Statistical Guarantees

So far we have accomplished the first task by showing that d_{qq} will be small if and only if $F^{(0)} = F^{(1)}$ given a large enough sample size m. Next we want to study the relationship between the samples size and the detection accuracy. That is, how close is $F^{(0)}$ to $F^{(1)}$ and how close is $Q^{(0)}$ to $Q^{(1)}$?

One of the methods to analyze the sample based Q-Q distance between probability distributions is the Vapnik-Chervonenkis theory. Here we follow the work by Wasserman [Wasserman, 2005].

Definition 4.2.1. Consider a probability space $(\Omega, \mathcal{A}, P)$. Let $X_1, X_2, \ldots, X_m \sim P$, $m \geq 1$, be i.i.d. real samples from a probability measure P. For $A \in \mathcal{A}$, the empirical probability P_m is defined by

$$P_m(A) = \frac{1}{m} \sum_{i=1}^{m} I(X_i \in A). \tag{4.8}$$

Given a finite set $R = \{x_1, x_2, \ldots, x_m\}$. Let S be a subset of R. We say $\mathcal{A}$ picks out S if

$$A \bigcap R = S \quad \text{for some} \quad A \in \mathcal{A}.$$

Let $N_{\mathcal{A}}(R)$ be the number of subsets of R picked out by $\mathcal{A}$ as

$$N_{\mathcal{A}}(R) = \sharp \left\{ A \bigcap R : A \in \mathcal{A} \right\}. \tag{4.9}$$

Definition 4.2.2. If $N_{\mathcal{A}}(R) = 2^m$, where n is the number of elements in R, we say that R is *shattered* by $\mathcal{A}$. Let $\mathcal{R}_m$ denote all finite sets with m elements. The *shatter coefficient*

[1] In probability theory, convergence *almost surely* (a.s.) means the convergence happens with probability one while convergence *in probability* is a weaker convergence meaning as the sequence progresses the probability of no convergence becomes smaller and smaller towards zero.

is defined by

$$s(\mathcal{A}, m) = \max_{R \in \mathcal{R}_m} N_{\mathcal{A}}(R) \tag{4.10}$$

Definition 4.2.3. The Vapnik-Chervonenkis dimension of $\mathcal{A}$ is defined by

$$VC(\mathcal{A}) = \max \{m : s(\mathcal{A}, m) = 2^m\} \tag{4.11}$$

Theorem 4.2.4. (Vapnik-Chervonenkis Theory) *Suppose $\mathcal{A}$ has finite VC dimension d. For all $n \geq d$,*

$$s(\mathcal{A}, m) \leq m^d + 1 \tag{4.12}$$

and

$$P\left(\sup_{A \in \mathcal{A}} |P_m(A) - P(A)| > \epsilon\right) \leq 8(m^d + 1)e^{-m\epsilon^2/32} \tag{4.13}$$

In this work, we wish to bound $P\left(\sup_{x \in \mathbb{R}} |F_m(x) - F(x)| > \epsilon\right)$. Let $\mathcal{A} = \{(-\infty, x); x \in \mathbb{R}\}$ and $A = (-\infty, x]$. Since $P((-\infty, x)) = F(x)$ and $P_m((-\infty, x)) = F_m(x)$, and $VC(\mathcal{A}) = 1$, we can get

$$P\left(\sup_{x \in \mathbb{R}} |F_m(x) - F(x)| > \epsilon\right) = P\left(\sup_{A \in \mathcal{A}} |P_m(A) - P(A)| > \epsilon\right) \leq 8(m+1)e^{-m\epsilon^2/32} \tag{4.14}$$

In fact, there is a tighter bound by applying the Dvoretzky-Kiefer-Wolfowitz (DKW) inequality to the difference in distributions.

Theorem 4.2.5. (Dvoretzky-Kiefer-Wolfowitz Inequality) *For any $\epsilon > 0$,*

$$P\left(\sup_{x \in \mathbb{R}} |F_m(x) - F(x)| > \epsilon\right) \leq 2e^{-2m\epsilon^2} \tag{4.15}$$

The DKW Inequality predicts how quickly an empirical distribution, F_m, will converge to the true distribution F from which the empirical samples are drawn. This strengthens the Glivenko-Cantelli theorem by quantifying the rate of convergence.

Now, We intend to determine certain lower bound on the sample size such that there is a large enough number of elements in the sample to ensure the detectable discrepancy.

That is, we would like a bound of the form

$$P\left(|Q_n(t) - Q(t)| > \epsilon\right) \leq \text{something small}$$

and the following theorem will do the work.

Theorem 4.2.6. [Shao, 2003] *Let $F(x_t + \epsilon) > t$ for any $\epsilon > 0$ and $m \geq 1$.*

$$P\left(|Q_m(t) - Q(t)| > \epsilon\right) \leq 4e^{-2m\delta_\epsilon^2} \tag{4.16}$$

where $\delta_\epsilon = \min(F(x_t + \epsilon) - t,\ t - F(x_t - \epsilon))$.

The proof of this theorem is attached in the Appendix.

Given $FAR = \gamma$,

$$4e^{-2m\delta_\epsilon^2} \leq \gamma \Leftrightarrow \ln 4 - 2m\delta_\epsilon^2 \leq \ln \gamma \Leftrightarrow \frac{\ln 4 - \ln \gamma}{2\delta_\epsilon^2} \leq m \tag{4.17}$$

When ϵ is small, we consider $\delta_\epsilon \cong \epsilon$.

Theorem 4.2.6 gives a lower bound on the size of the second window in our Q-Q distance based quickest detection algorithm, while keeping the false alarm rate at a given level. Let us see an example:

Suppose we need to ensure that the distance between our empirical quantile function Q_n and the true quantile function Q on the real line is less than or equal to $\epsilon = 0.1$, with a false alarm rate no more than 0.05. According to Eq. 4.17,

$$\frac{\ln 4 - \ln 0.05}{2(0.1)^2} \leq m \Rightarrow m \geq 219.1 \tag{4.18}$$

a samples size of $m = 220$ would be large enough to detect if the distance between two quantile functions is larger than 0.1 while keeping the false alarm rate below 5%.

4.3 Limitations of Q-Q Distance

We have proved the convergence of our Q-Q distance and provided a tight bound of the window size for our detection algorithm. However, the Q-Q distance has two limitations

which need to be improved in the future.

The first limitation is the computational cost due to the embedded sorting algorithm, especially when the window size is big. Nevertheless, it may be improved by developing fast sorting method. For Q-Q distance based binary detection, which we will show in the next chapter, the sorting can actually be avoided by summing the quantized stream.

The second limitation is the incapability when the properties or parameters defining the stream are fast time-varying, such as in detecting or predicting the turns of stock exchange index. One possible solution is to combine Q-Q distance with some density estimation method to timely update the latest distribution in the moving reference window. This needs additional future research.

Chapter 5

Decentralized Quickest Detection

Consider a distributed sensing environment with L geographically deployed sensors. Let $\mathbf{X}(n)$, $n \geq 1$, again be the L-dimensional data collected that is chosen for monitoring, but the component $\mathbf{X}_i(n)$ represent the i-th channel of the data stream. At an unknown point in time, λ ($\lambda \geq 1$), a change happens and the distributions of all channels change accordingly. All channels are assumed to be independent and the observations in each channel are independent and identically distributed (i.i.d.).

While a great deal of research has been done in applying quickest detection techniques to distributed systems, to the best of our knowledge, there has been no decentralized nonparametric quickest detection procedures successfully implemented. For example, the score function based nonparametric detection method [Tartakovsky et al., 2006b] works well in centralized detection but would fail in binary detection when the sum of the mean values of the induced distributions is larger than 1. More detailed explanation is provided in Section 5.1. Here we show how Q-Q distance based nonparametric detection can be deployed in a decentralized fashion using two schemes, binary detection and decision fusion.

5.1 Quantized Quickest Detection

We have examined the binary quickest detection for parametric detection algorithms. As discussed in Section 2.5.1, binary quickest detection looks for an optimal quantizer such that the original observation data can be coded as binary streams, which are then transmitted

as the compressed representation of data to a processing center for detection. However, currently there is no quantized nonparametric detection procedure. Rank based and score function based detection procedures are unable to be implemented for quantized stream. The difficulty lies in the way that quantization of the stream will not change the fact that both the pre-change and post-change distributions are still unknown. In this section, we present the binary detection implementation for Q-Q distance based nonparametric algorithm, which, to our best knowledge, has never been implemented.

In parametric binary detection procedures with known distributions, the optimal quantizer is the one that maximizes the K-L information distance of the induced distributions. However, in nonparametric detection the computation of the quantization threshold is not feasible. Instead, we use a fixed stationary sensor quantizer, e.g., the estimated pre-change mean, as the threshold.

5.1.1 The Detection Algorithm

Let m_i be the estimated mean of the observations under distribution $f_i^{(0)}$ which we use as the quantization threshold. We first quantize $\mathbf{X}_i(n)$ as

$$\mathbf{B}_i(n) = \begin{cases} 1 & \mathbf{X}_i(n) \geq m_i \\ 0 & \mathbf{X}_i(n) < m_i \end{cases} \tag{5.1}$$

resulting in a Bernoulli sequence, where $\{\mathbf{B}_i(1), \mathbf{B}_i(2), \ldots, \mathbf{B}_i(k-1)\}$ has the induced probability mass function $p_i^{(0)}$ and $\{\mathbf{B}_i(k), \mathbf{B}_i(k+1), \ldots\}$ has $p_i^{(1)}$. If we draw the Q-Q plot of the two segments, all drawings fall in only four possible spots, (0,0), (1,1), (0,1) and (1,0). It is easy to see that the greater the difference between $p^{(0)}$ and $p^{(1)}$, the more points fall in (0,1) or (1,0), and the bigger the distance d_{qq}. See Fig. 5.1 for a comparison of Q-Q plots generated using the original observations and the quantized Bernoulli streams with the pre-change distribution being a standard Gaussian and the post-change distribution a uniform from 0 to 3.

This binary detection procedure sometimes requires training for particular situations or unknown anomalies. For example, when the induced distributions $p_i^{(0)}$ and $p_i^{(1)}$ are the same or close, as shown in Fig. 5.2 with the before-change distribution a zero-mean

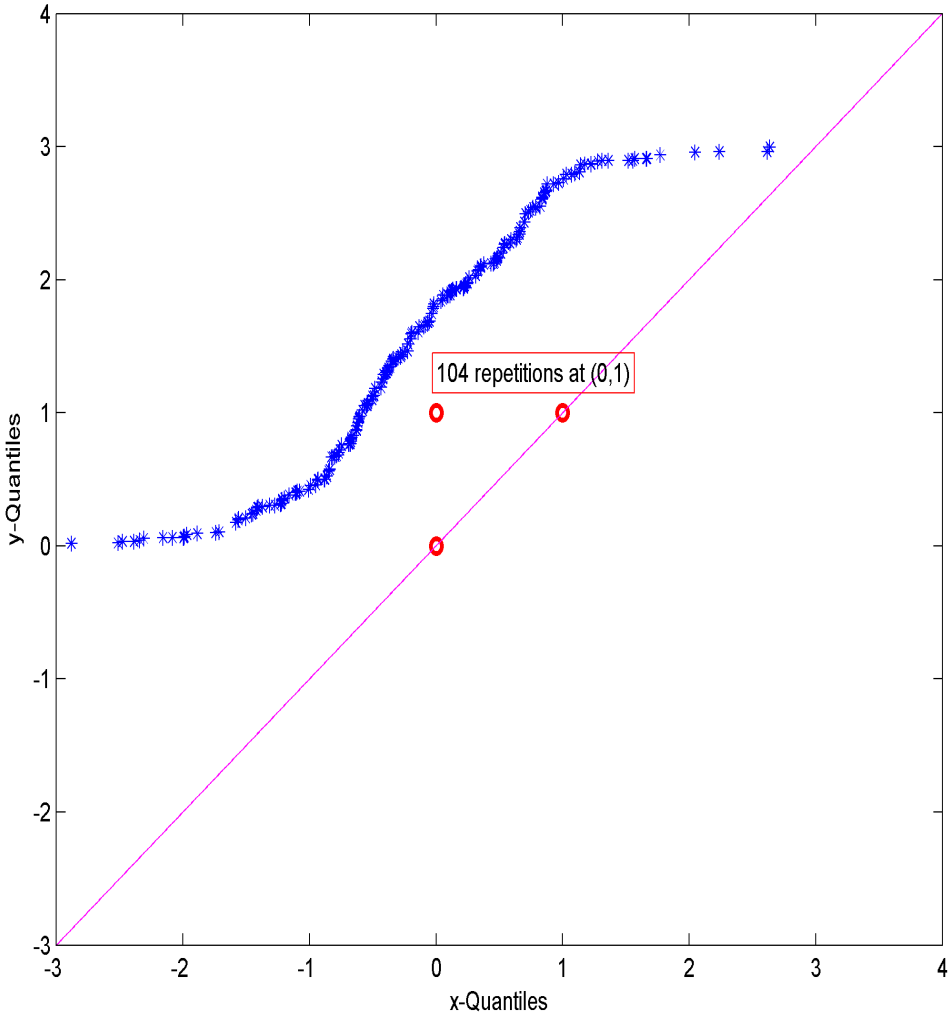

Figure 5.1: Q-Q plots of two sequences before and after quantization

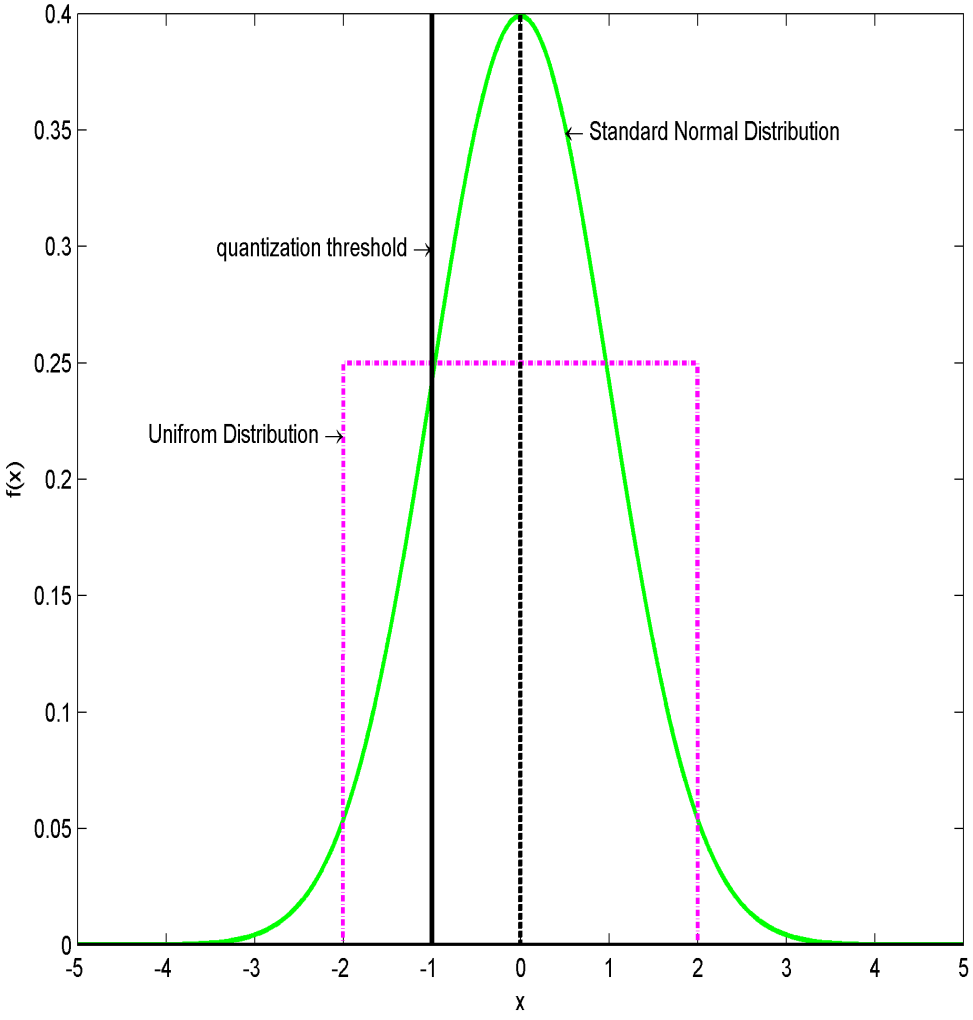

Figure 5.2: The choice of quantization threshold

Gaussian and the after-change distribution a zero-mean Uniform, then using the mean as the threshold to quantize the stream cannot really differentiate the segments. So another quantization threshold which can amplify the distance, -1 in this case, could be used.

The stopping time of the binary detection procedure at the processing center can then be defined as

$$\tau_{qb}(h) = \min\{n \geq 1 \mid \sum_{i=1}^{L} d_{qq,i} \geq h\} \tag{5.2}$$

and $d_{qq,i}$ is the Q-Q distance in the i-th channel.

The binary detection algorithm is described in Algorithm 2 and is slightly different from Algorithm 1 .

Algorithm 2 Q-Q Distance Based Binary Quickest Detection

for $i = 1$ to L **do**
  Quantize the stream
  WindowOne $\leftarrow$ first s_0 quantized observations in the i-th channel
  WindowTwo $\leftarrow$ first s_1 quantized observations in the i-th channel
  while There are incoming observations **do**
    Slide WindowTwo forward by 1 sample
    Count 1's in WindowOne and WindowTwo, respectively
    Calculate the distance $d_{qq,i}$ using the counts of 1's
  end while
end for
if $\sum d_{qq,i} > h$ **then**
  Alarm triggered
  Break
end if

5.1.2 Remarks

In Q-Q distance based quantized quickest detection, depending on the application, the original data stream can be quantized into multiple discrete values, not just only binary stream. In this dissertation we only use the simplest binary data stream to show the concepts and detection process.

As we mentioned earlier, in binary detection, if the score function based procedure is used, then Equation 2.35 would be rewritten as

$$S_n^g = \left(S_{n-1}^g + \mathbf{B}_i(n) - p^{(0)} - p^{(1)}\right)^+. \tag{5.3}$$

If the sum of induced distributions $p^{(0)}$ and $p^{(1)}$ is larger than 1, the output of Eq. 5.3 will remain at 0, causing the procedure to fail for detecting any changes.

5.2 Decision-based Fusion

Another decentralized detection scenario is to perform quickest detection locally at each sensor and only send local decisions to the processing center for decision fusion. When parametric detection procedures, such as CUSUM tests, are performed at local sensors, three popular fusion rules [Tartakovsky and Veeravalli, 2008] have been adopted to generate the global decision. Suppose $\tau_i = \min\{n \geq 1 \mid S_{n,i} \geq \omega_i h\}$ is the local stopping time of the i-th sensor, where ω_i is the weight of the threshold. At the fusion center, the first fusion rule is to choose the stopping time as

$$\tau_{max} = \max_{1 \leq i \leq N} \tau_i \tag{5.4}$$

which takes the largest τ_i as the global decision (stopping time), indicating that a change is declared only when all the local sensors have voted for a change.

The second rule is to choose τ_{min},

$$\tau_{min} = \min_{1 \leq i \leq N} \tau_i \tag{5.5}$$

such that the global decision of change is declared at the first time any local decision is in favor of the change.

The stopping time τ_{all} in the third fusion rule is defined as

$$\tau_{all} = \min\{n \geq 1 \mid \min_{1 \leq i \leq N} (S_{n,i}/\omega_i) \geq h\} \tag{5.6}$$

By this rule, a change is declared at the first time when all sensors send 1's to the fusion center.

The fusion rules above are widely used in asymptotic performance analysis [Mei, 2005], [Moustakides, 2006], [Tartakovsky and Veeravalli, 2008] but these rules actually do not consider the highly possible sensor/channel failures that may lead to incorrect local decisions

being transmitted to the fusion center. For example, a malfunctioning sensor with stuck-at faults can continuously report "no change" to the fusion center irrespective of the fact that a change has occurred. This faulty sensor would cause a complete failure of detection when the τ_{all} or τ_{max} rule is adopted. It is desired that the errors or uncertainty in some sensors can be corrected by other sensors. Here we use majority voting [Lam and Suen, 1997] to aggregate the local decisions generated by the Q-Q distance-based procedure in real time. Majority voting is one of the simplest fusion methods for decision fusion tasks and is as effective as the other more complicated schemes [Lam and Suen, 1997]. Assume each sensor reports a local decision

$$\psi_{n,i} = \begin{cases} 1 & \text{if } d_{qq,i} \geq \omega_i h \\ 0 & \text{otherwise} \end{cases} \tag{5.7}$$

to the center at every time interval n. The stopping time at the center is defined as

$$\tau_v = \min\{n \geq 1 \mid \sum_{i=1}^{L} \psi_{n,i} > N/2\} \tag{5.8}$$

which means a change is declared the first time when more than half of the sensors agree.

Chapter 6

Experimental Evaluation

This chapter presents the experiment results of the algorithms proposed in this dissertation. We begin with results from the Q-Q plot based nonparametric quickest detection in Section 6.1. This section demonstrates the overall detectability of our algorithm. Comparisons between the proposed algorithm and the classical parametric and other nonparametric detection algorithms are made. In Section 6.2, we show the performance of our binary detection algorithm, as well as the decision fusion. Then, in Section 6.3, we investigate the computational cost of our proposed algorithm by comparing its actual computation time with other existing detection algorithms. Finally, in Section 6.4, we apply both our Q-Q distance based detection algorithm and the score function based algorithm to the real intrusion detection application. These four sections serve as the successful evidence of our Q-Q distance based algorithms.

6.1 Nonparametric Quickest Detection

We compare the Q-Q distance based detection procedure with the benchmark CUSUM detection (parametric) and the score function based (nonparametric) detection schemes. We present the results of Monte Carlo simulations with 10^4 replications which are sufficient for estimating ADD and FAR. The plot of ADD *vs.* -log(FAR) is used to show the performance of different detection procedures. The lower the FAR requirement, the longer the delay. In addition, a shorter performance curve would indicate the detection failure at

low false alarm rates. We perform two sets of experiments to show the effectiveness of the Q-Q distance-based nonparametric detection scheme in both single channel environment and the multi-channel distributed environment.

6.1.1 Single Channel Detection

We compare the detection capability of the three approaches in identifying changes among three different combinations of distributions, including Gaussian changed to uniform distribution, Gaussian changed to another Gaussian with a different variance, and Gaussian changed to another Gaussian with a different mean. The degree of changes is the smallest in the last combination.

Change detection - from Gaussian to Uniform. In this experiment, the distribution changes from a standard Normal distribution to a Uniform distribution between 0 and 1. Fig. 6.1 shows the operating characteristics. With known information of the distributions, the CUSUM test yields the best detection performance with the shortest delay and lowest FAR. Although the Q-Q distance-based method shows higher ADD compared to the score function based procedure, its performance curve is longer, indicating that it survives lower FAR better than the score function based scheme.

Small change detection - from Gaussian to Gaussian of different variance. In this experiment, the pre-change distribution remains as the standard Gaussian with variance, 1, but the post-change distribution becomes a Gaussian as well just with a slightly higher variance, 1.5. Figure 6.2 shows that the CUSUM test still performs the best providing the shortest delay and the lowest false alarm rate. However, the Q-Q distance-based procedure outperforms the score function based procedure across the entire range of false alarm rate. In addition, the performance curve from the score function based scheme is a lot shorter than the other two schemes indicating the procedure fails at low FAR.

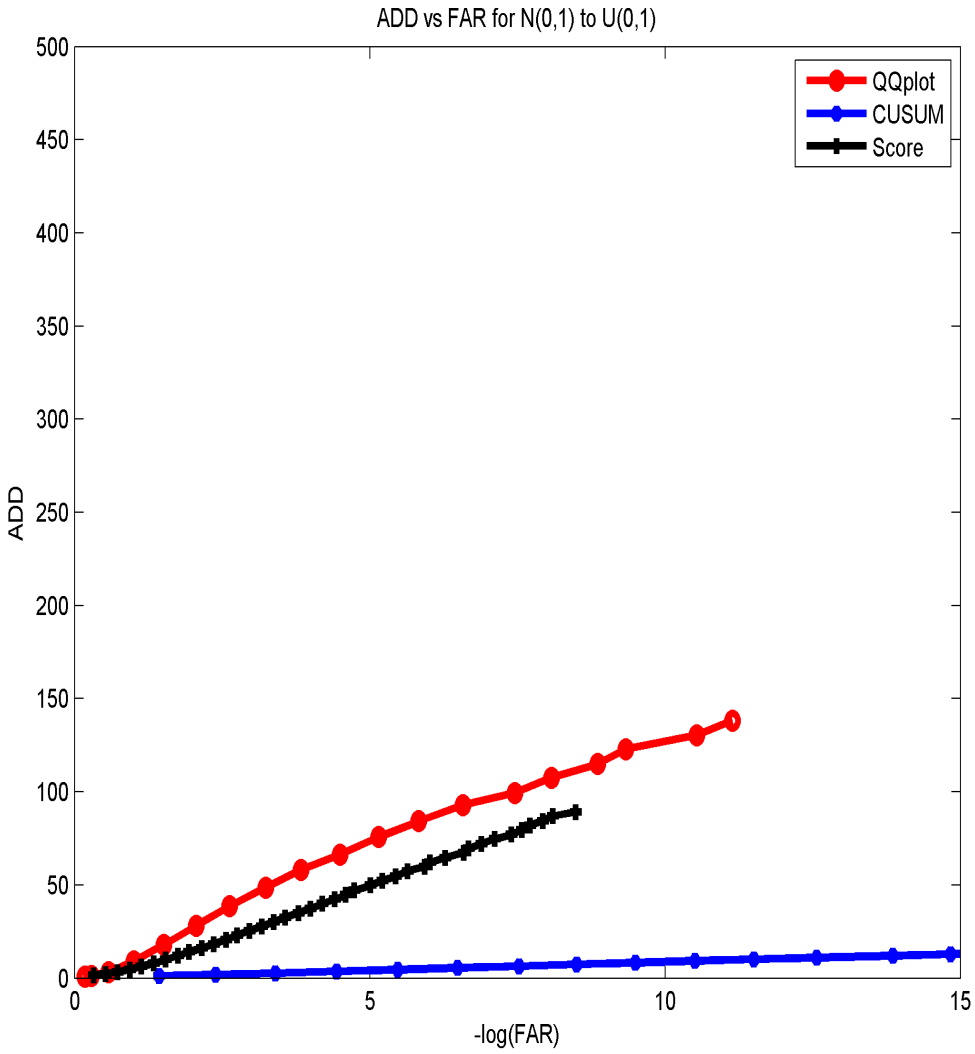

Figure 6.1: ADD vs. FAR (Change from Normal(0,1) to Uniform(0,1))

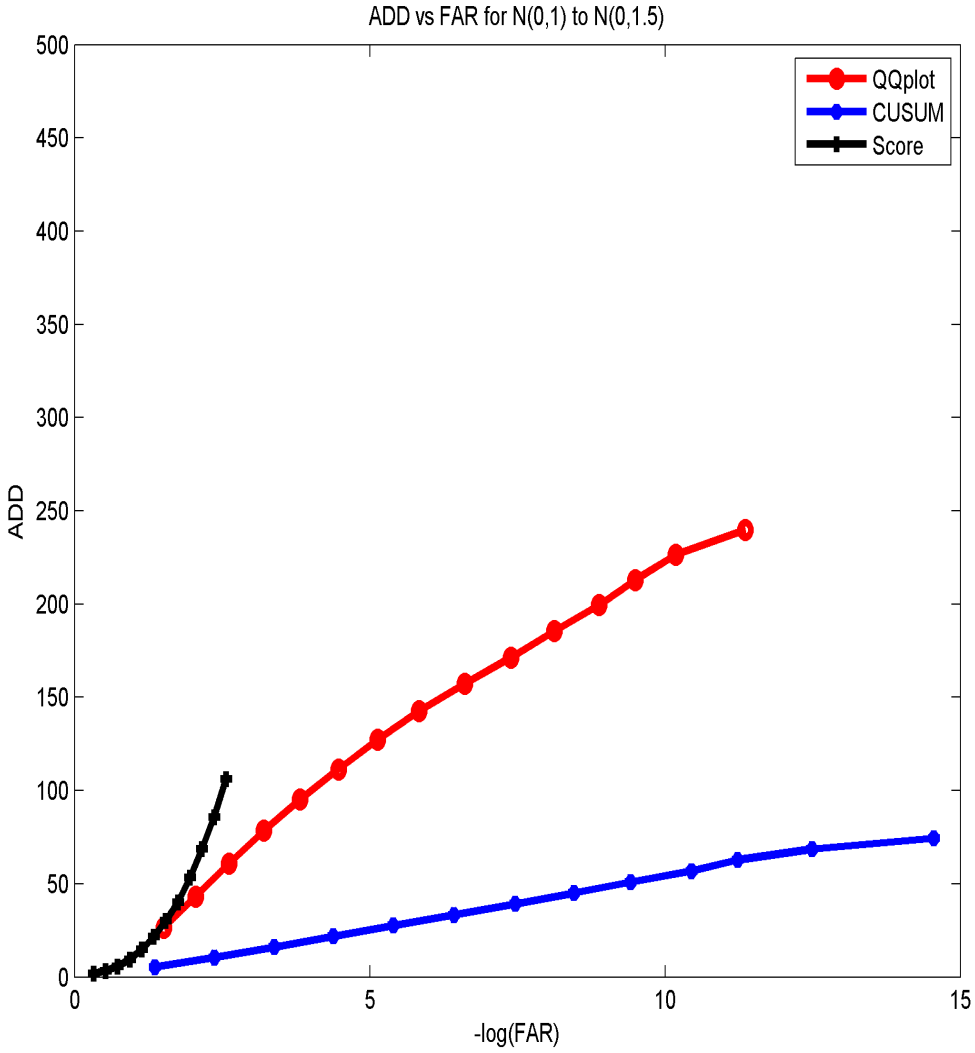

Figure 6.2: ADD vs. FAR (Small Change from Normal(0,1) to Normal(0,1.5))

Small change detection - from Gaussian to Gaussian of different mean. In this experiment, the distributions still change from Gaussian to Gaussian but with even smaller scale, from zero-mean (pre-change) to a mean of 0.5, 0.3, and 0.2 (post-change), respectively. The score function based procedure presents almost the same performance as the CUSUM test when the false alarm rate is relatively high but fails with low FARs, indicating its inability of detecting small changes. As we observe from Figs. 6.3-6.5, as the degree of changes in distributions becomes smaller and smaller, the Q-Q distance-based

detection procedure still succeeds in the detection alongside the CUSUM test although both experience larger delays, showing the merit of the proposed detection scheme in the identification of small changes. When the change is even smaller, for example, small change from Normal(0,1) to Normal(0.1,1), both the Q-Q distance-based and the CUSUM procedures will fail to detect the change.

6.1.2 Detection with Different Window Sizes

In this section, we compare the impact of window sizes. We use three window sizes of 200, 400, and 600 observations to detect a change from $N(0,1)$ to $N(0,1.5)$. As we can see in Figure 6.6, larger window size has bigger detection delay and lower false alarm rate while smaller window size has short delay but higher false alarm rate.

6.2 Decentralized Nonparametric Quickest Detection

For decentralized detection, we conduct two experiments that evaluate binary detection and decision-based fusion. We simulate 6 independent channels which have distribution changes at exact the same time. From channel 1 to channel 6, the changes are Uniform(0,1) to Uniform(0,5), Uniform(-1,1) to Uniform(-1,4), Normal(0,1) to Normal(2,1), Normal(0,1) to Uniform(0,2), Poisson distributions with mean of 8 to 10, and Normal(0,1) to Normal(0.5,1).

6.2.1 Binary Quickest Detection

In the first experiment, we compare the performance of binary detection using the Q-Q distance-based nonparametric scheme and the benchmark parametric CUSUM detection scheme. Fig. 6.7 shows the operating characteristics. With complete information of the original and induced distributions, the CUSUM scheme outperforms the Q-Q distance-based scheme which is within expectation. However, the Q-Q distance-based scheme still shows acceptable performance in a sense of small detection delay and low false alarm rate. To the best of our knowledge, this is the first success implementation of nonparametric detection procedure using decentralized binary detection.

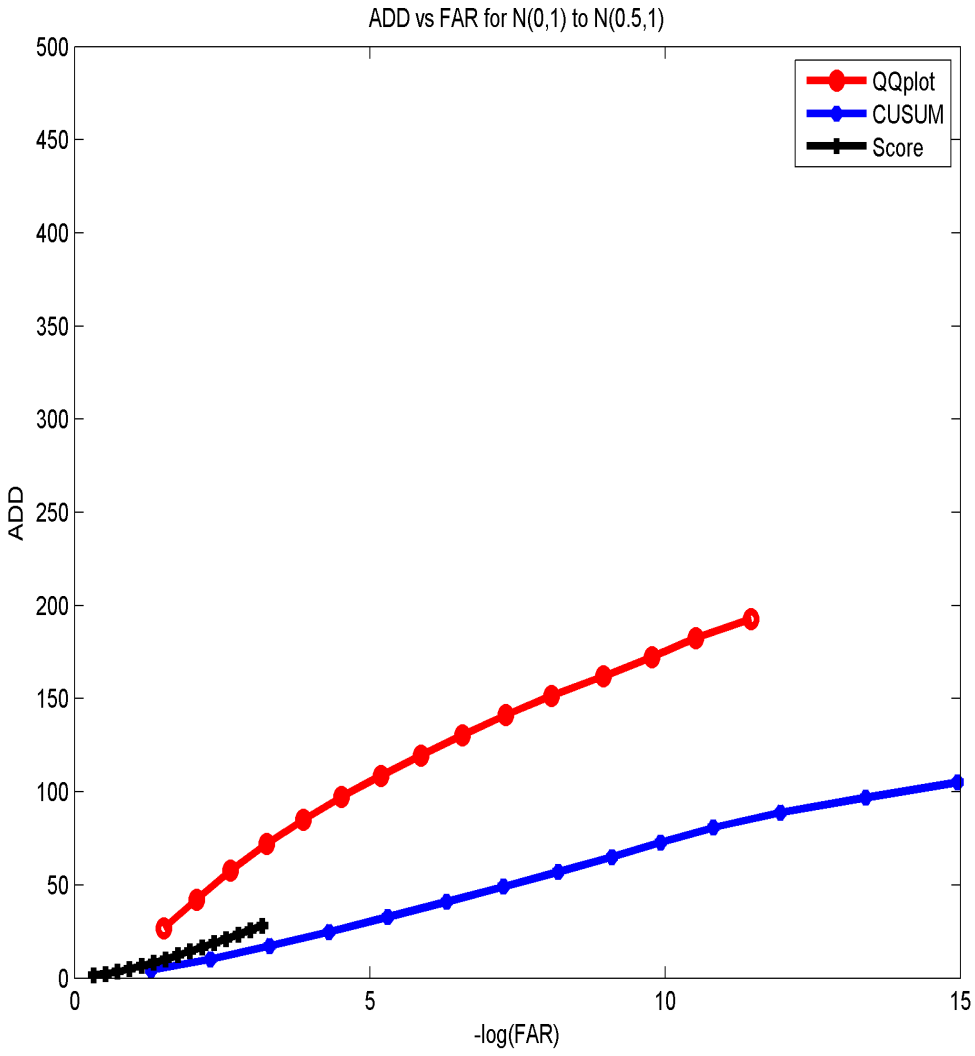

Figure 6.3: ADD vs. FAR (Small Change from Normal(0,1) to Normal(0.5,1))

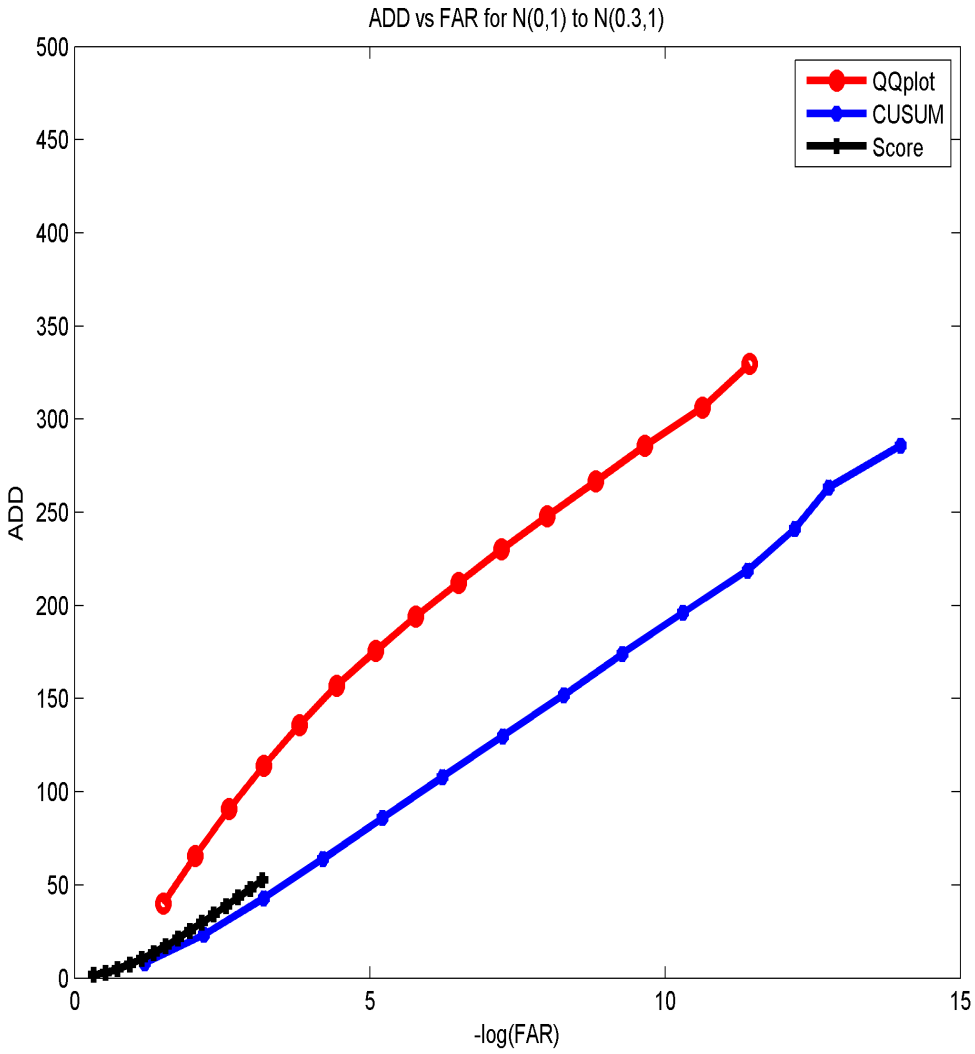

Figure 6.4: ADD vs. FAR (Small Change from Normal(0,1) to Normal(0.3,1))

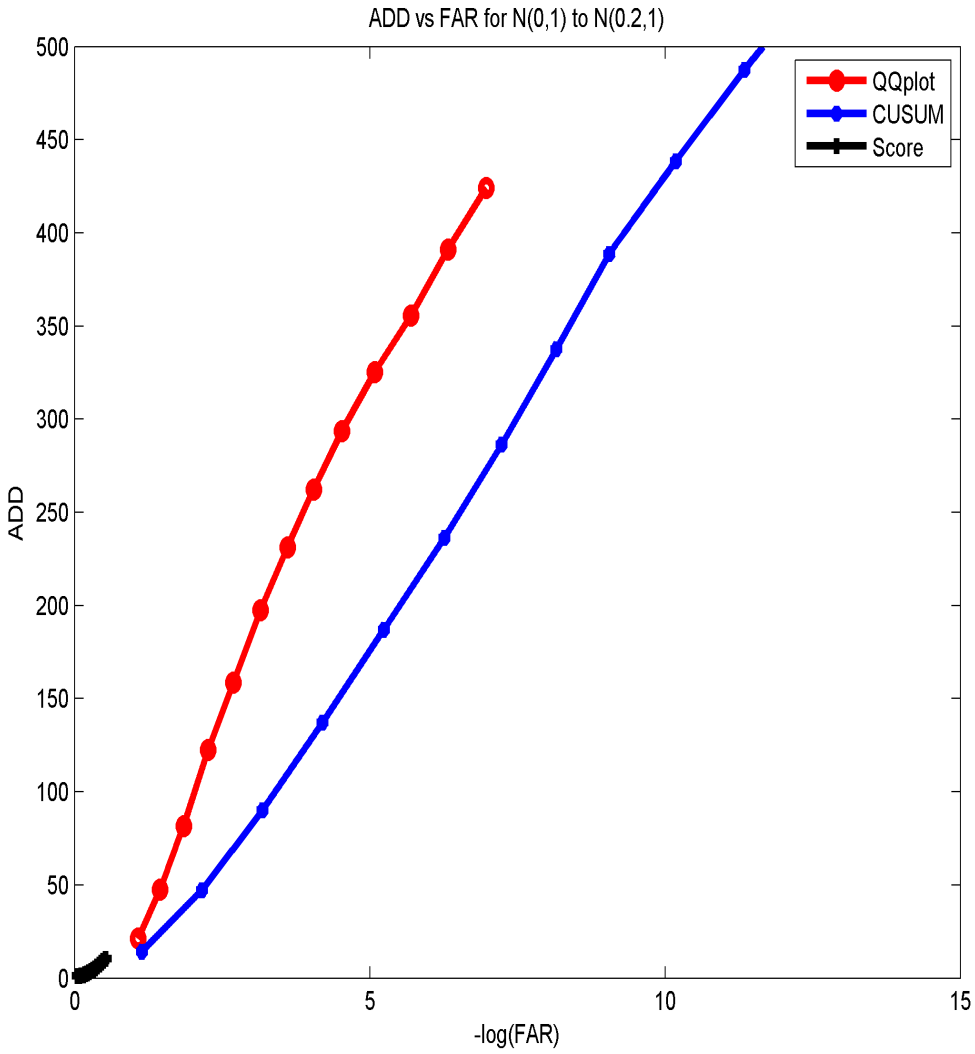

Figure 6.5: ADD vs. FAR (Small Change from Normal(0,1) to Normal(0.2,1))

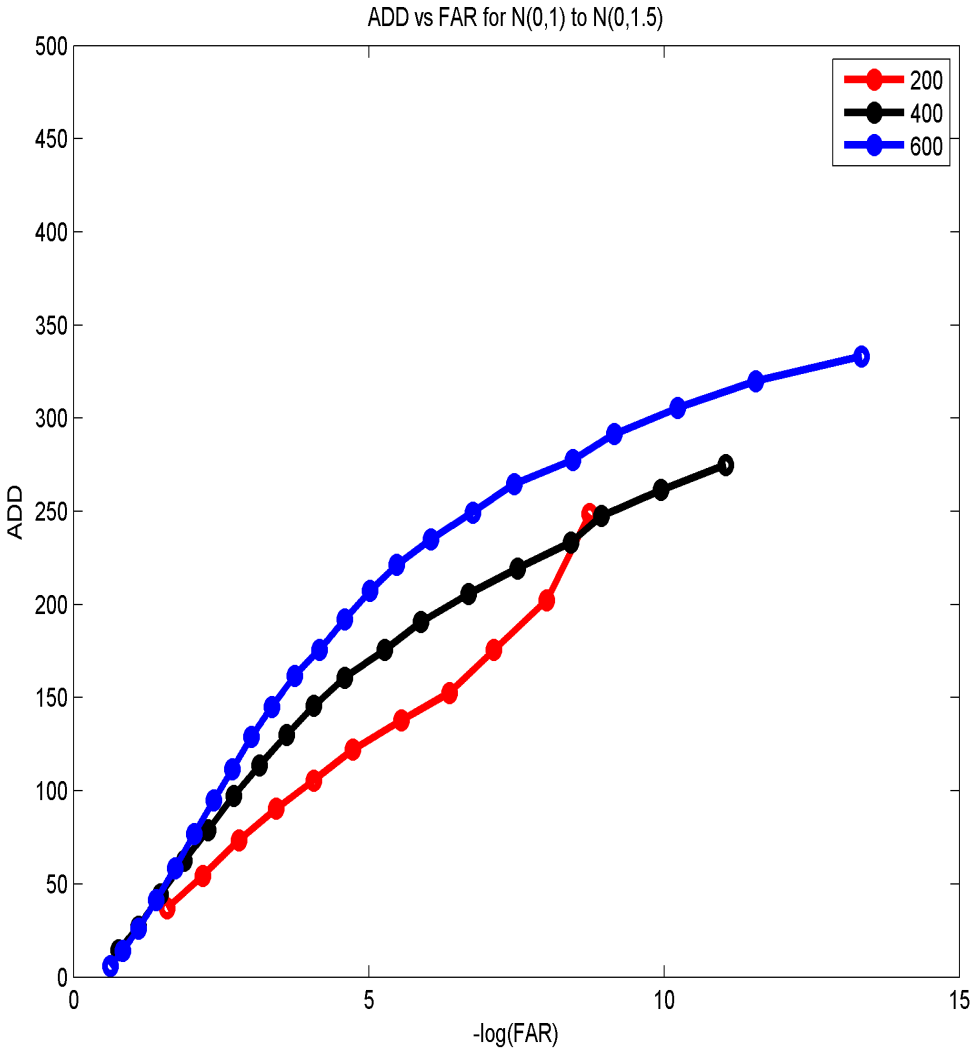

Figure 6.6: ADD vs. FAR with different window sizes

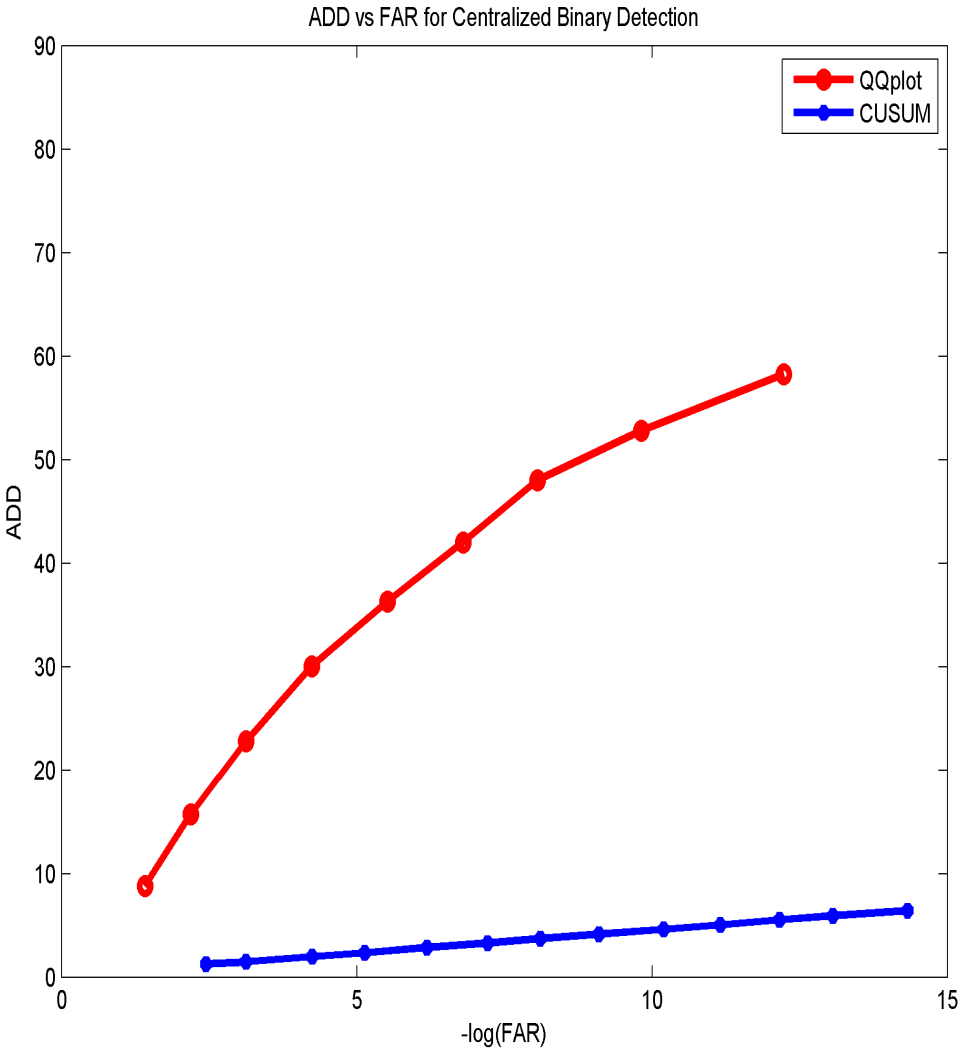

Figure 6.7: Operating characteristics of binary detection procedures

Table 6.1: Decision fusion results

Rules	τ_{min}	τ_{max}	τ_{all}	τ_v	τ_{qb}
Average Delay	-89	156	157	69	48
Variance	252.5	41.0	40.1	17.2	12.2
Minimum Delay	-2071	65	65	28	8
Maximum Delay	42	299	299	133	81

6.2.2 Decision Fusion

In the second experiment, we compare the performance of the four decision-based fusion rules, τ_{min}, τ_{max}, τ_{all}, and τ_v, as well as the result from binary detection, τ_h^{qb}, as defined in Eq. 5.2. Note that only the Q-Q distance-based detection procedure is applied. We observe from Fig. 6.8 that the τ_{min} rule yields an average stopping time before the real change time with big variance which means it produces lots of false alarms and the performance is very unstable or unpredictable. The τ_{max} and τ_{all} rules generate similar results and introduce fewer false alarms but give much longer delays than the others. τ_v from the majority voting gives the smallest detection delay with the smallest variance among the four decision-based fusion rules, showing its effectiveness in producing more accurate and reliable results. τ_h^{qb} is the average delay from the Q-Q distance-based binary detection. It yields an even smaller delay and variance compared to the τ_v rule, which is consistent with our previous discussion that although decision-based fusion provides the most effective usage of communication bandwidth and energy, it experiences a little bit degradation in performance in terms of delay and stability because only local decisions are fused. However, the τ_v-based fusion still shows very close performance to binary detection. The data of the fusion results are also tabulated in Table 6.2.

Experimental results with simulated data showed that this detection procedure is able to detect the changes with comparable performance as the benchmark CUSUM detection scheme. The majority voting decision fusion rule generates better detection decision than other traditional rules from parametric detection.

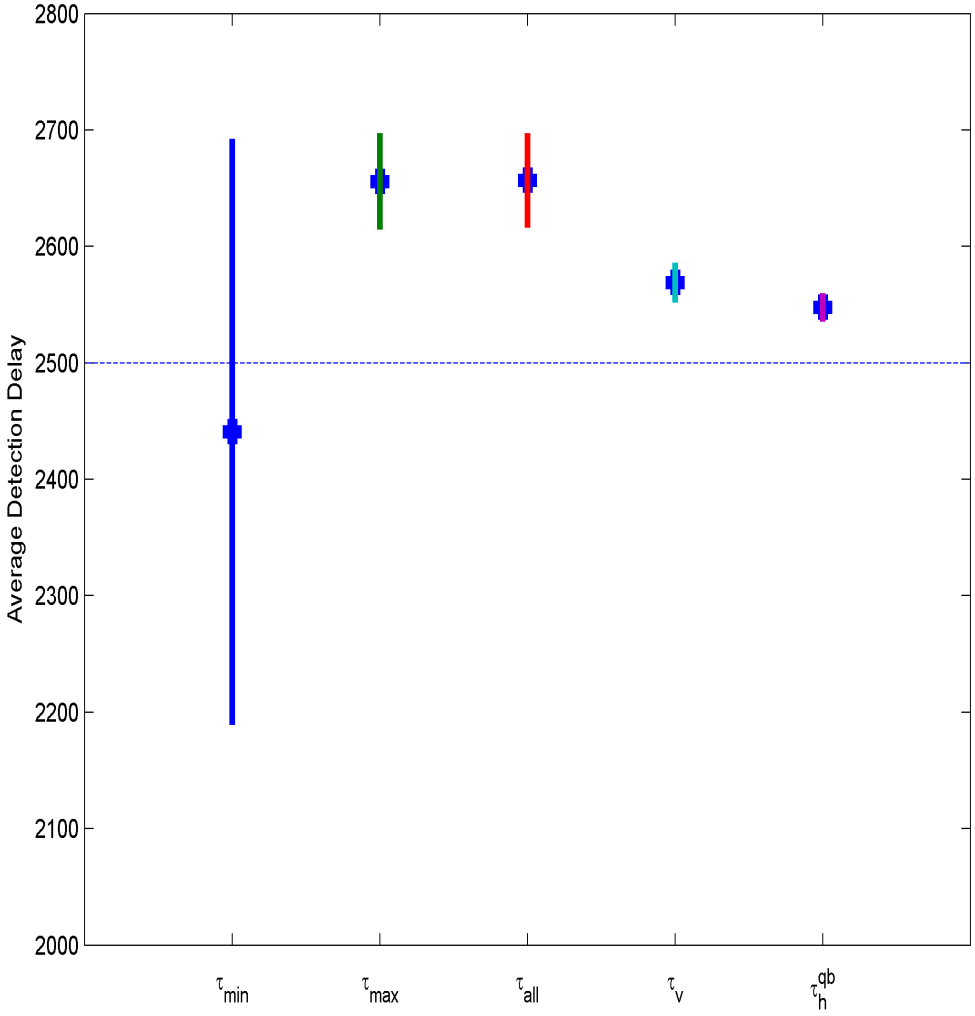

Figure 6.8: Results of decision fusion rules

Table 6.2: Running time

Rules	Q-Q	CUSUM	Score Function
Minimum running time (s)	0.1060	0.0024	1.8662e-004
Maximum running time (s)	0.1784	0.0622	6.6210e-004
Average running time (s)	1.1081	0.0025	1.9857e-004

6.3 Computational Efficiency

We now explore the computational efficiency of the proposed algorithm. The implementation for these results is in Matlab on a laptop computer with an Intel Core 2 Duo at 1.83GHz with 2GB memory. The table below shows the running time in seconds for a single observation. The Q-Q distance based procedure runs the slowest compared to the CUSUM test and the score function based procedure because there is an embedded sorting task. The computational complexity of the proposed algorithm is $O(n \log n)$ while the other two are essentially $O(n)$.

6.4 Application to Intrusion Detection

Previous experiments show the success of Q-Q distance based algorithms in detecting changes in synthetic data. Now we apply the nonparametric procedure to the real world data. We adopt the KDD Cup 99 data [KDD, 1999] to conduct the experiments. The KDD data is the data set used for the Third International Knowledge Discovery and Data Mining Tools Competition. It has 41 features extracted from the DARPA Off-line Intrusion Detection Evaluation [Lippmann et al., 2000]. Also, this data includes 38 different attack types. Each of the 38 attack types falls into one of the four attack categories. They are:

- DOS: denial of service,
- U2R: unauthorized access to root privileges,
- R2L: unauthorized access to local from a remote machine,
- Probe: surveillance and port scan activities,

and when there was no attack, the observation was labeled as NORMAL.

From the original data set, we choose a segment containing 8000 observations. The first 4000 observations are from the category of NORMAL, and the next 4000 are from DOS. We assume that the distributions of some of the features change when an attack is undertaken. Here we choose to detect the changes of two features "count" and "dst.host.count". Figures 6.9 and 6.11 are the real observations of the two features. Figures 6.10 and 6.12 show the detection results and the Q-Q distance based procedure clearly outperforms the score function based procedure.

This chapter concludes our experimental results for the nonparametric quickest detection, binary detection, decision fusion, and real application. It documents the qualitative and quantitative results of each algorithm. We now move to the final chapter of this dissertation, which is the conclusions that we draw from these experiments and some discussion for the future research.

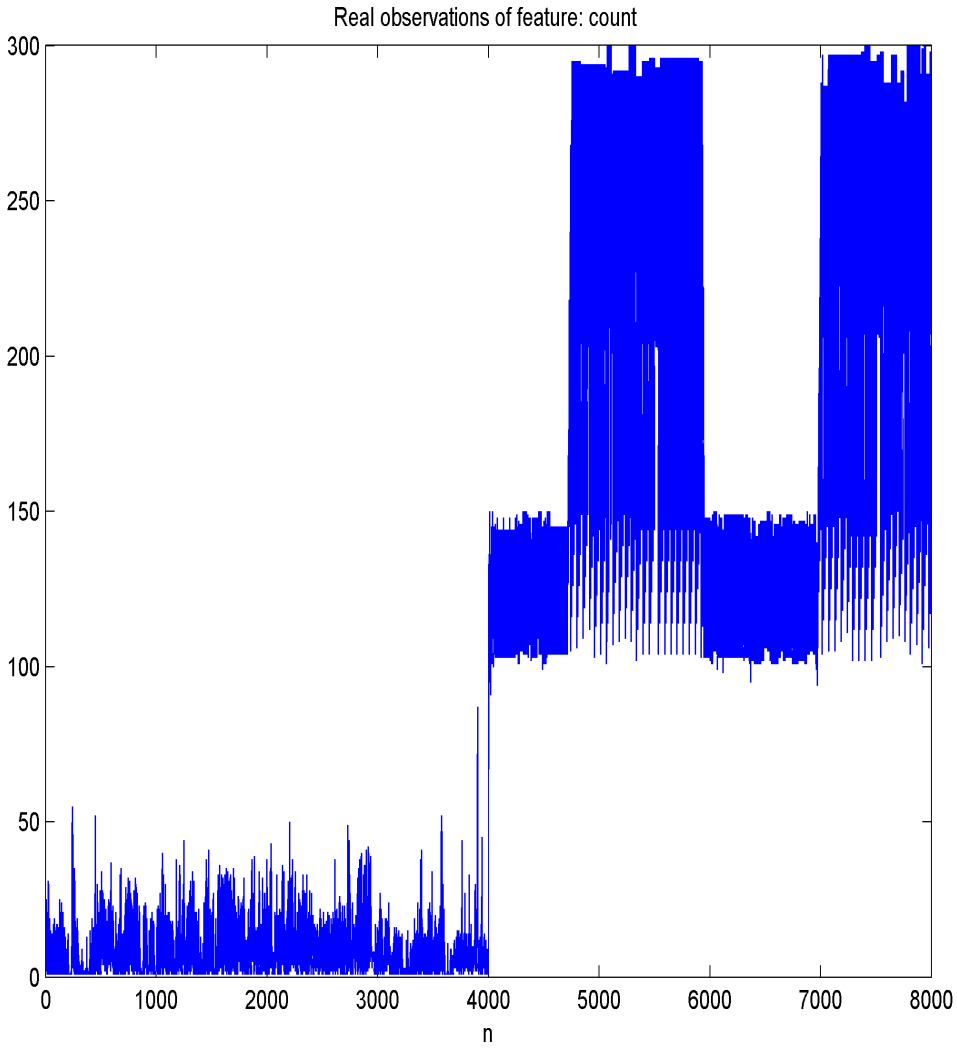

Figure 6.9: The original observations of feature: count

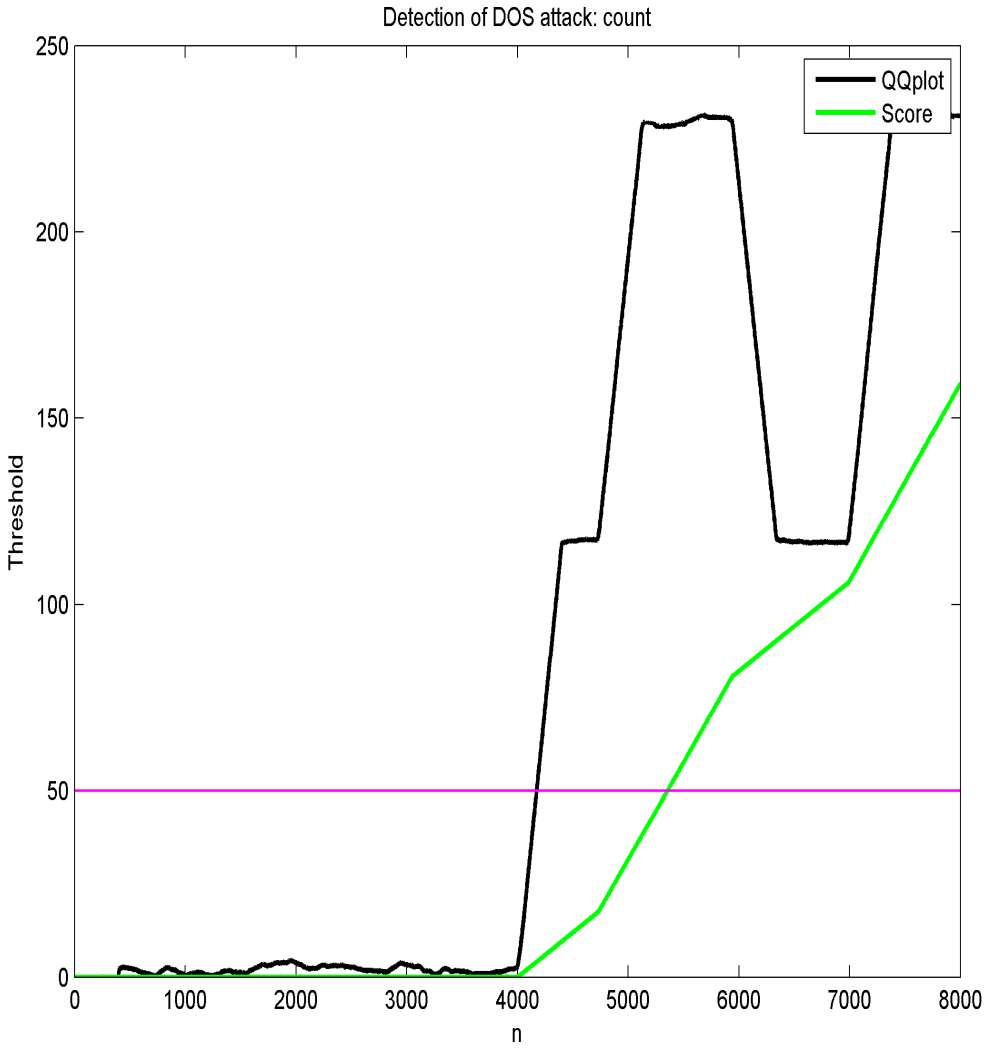

Figure 6.10: Detection of DoS attack

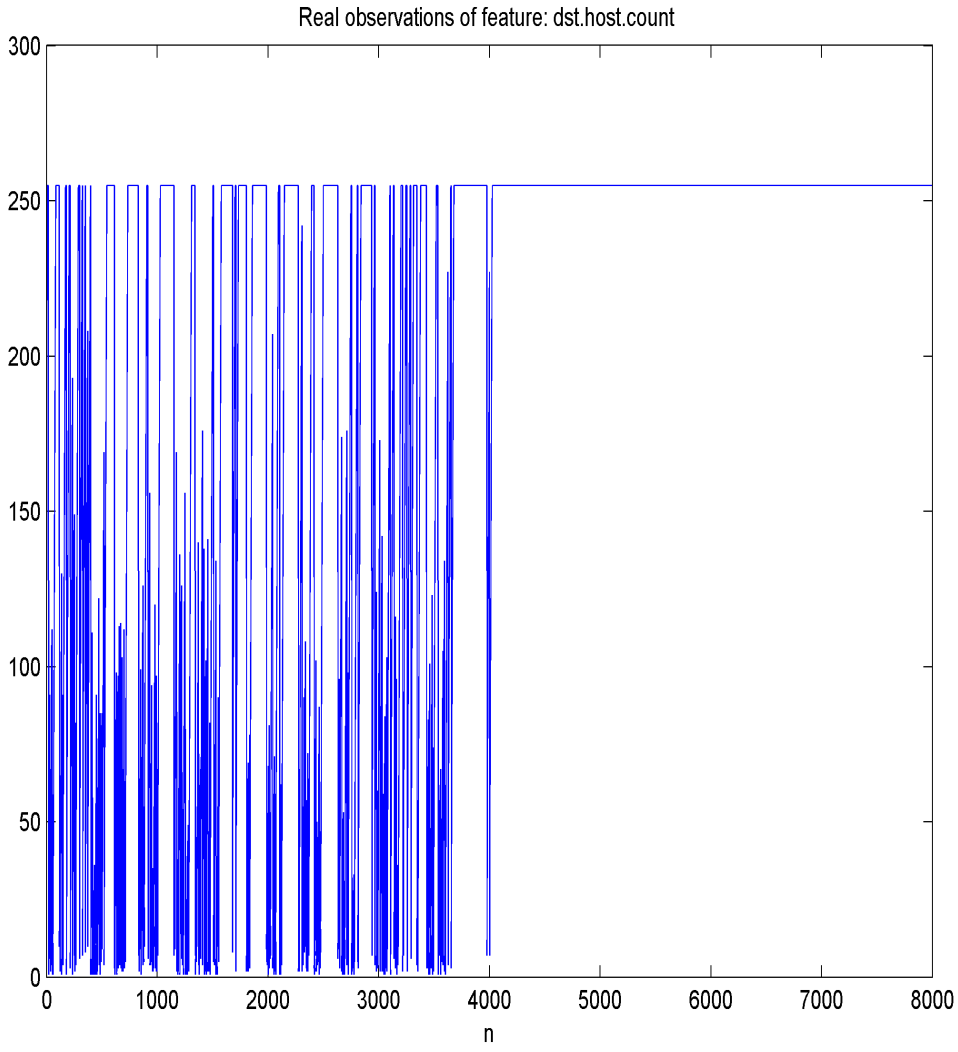

Figure 6.11: The original observations of feature: dst.host.count

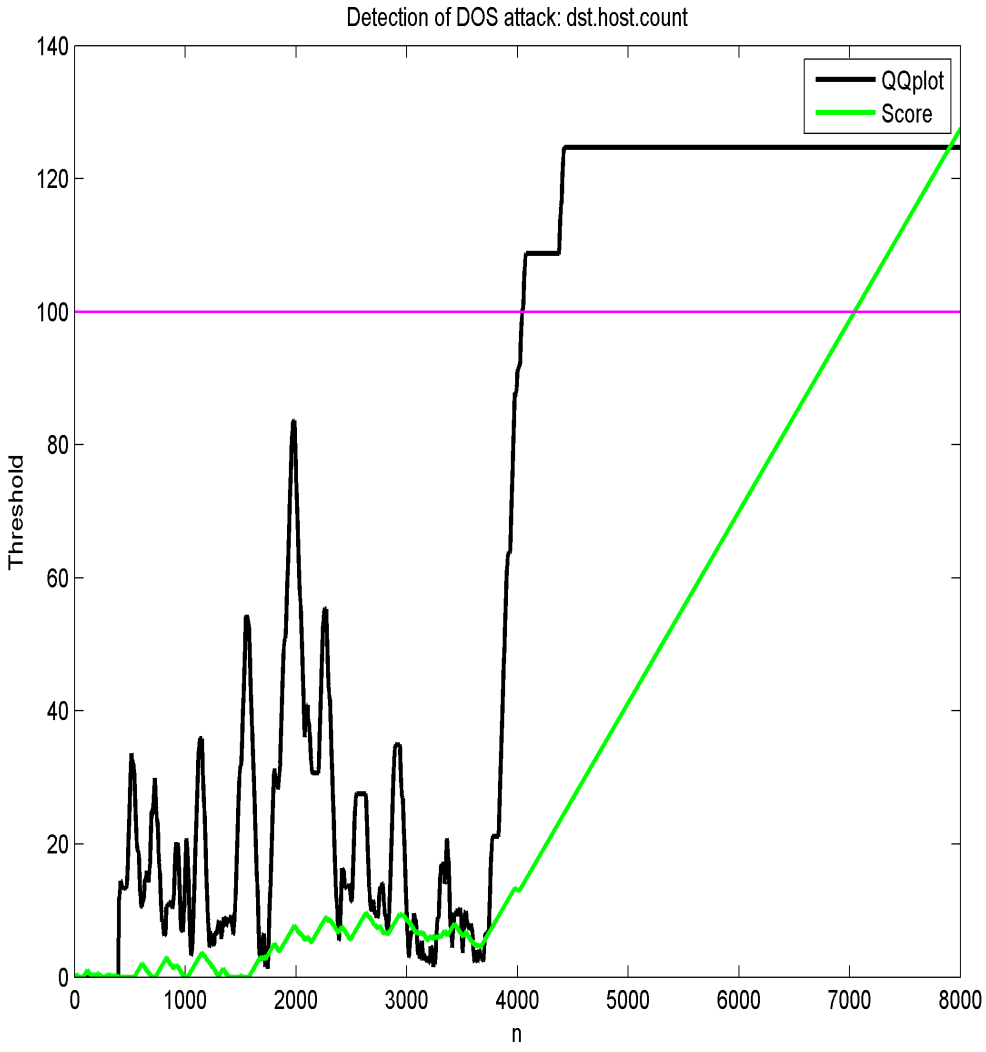

Figure 6.12: Detection of DoS attack

Chapter 7

Conclusions

In this dissertation, we have described a nonparametric quickest detection algorithm to detect changes in data streams in cases where no prior knowledge on the probability distribution is at hand. We also extend this detection scheme into a distributed detection environment where quickest detection with quantized data and decision-based fusion are presented. In the previous chapters, we have reviewed other related research in the literature, and we have presented the theoretical analysis along with experimental results to support our approach. We now conclude this dissertation with a brief summary of the contributions and a short discussion of future research.

7.1 Summary of Contributions

The primary contribution of this research as described in this chapter is the creation of the Q-Q distance and the Q-Q distance based detection algorithm. This algorithm extends the state of the art in nonparametric quickest detection and decentralized nonparametric quickest detection. In this dissertation, we offer four contributions as follows:

- **Q-Q Distance.** This contribution is a novel distance measure between distributions. The advantage of this distance measure is that it calculates the distance between distributions inferred directly from data sets generated from the distributions, thus no prior knowledge of the distributions is needed and less estimation error is introduced.

- **Q-Q Distance Based Quickest Detection.** This contribution is a novel nonparametric quickest detection algorithm for the detection of distributional change in data streams. The strength of this algorithm is its effectiveness of detecting changes, especially small changes.

- **Decentralized Nonparametric Quickest Detection.** This contribution is mainly a novel binary quickest detection algorithm for distributed detection. This algorithm is an extension of our nonparametric quickest detection algorithm and seems to be the first decentralized nonparametric quickest detection procedure.

- **Performance Guarantee.** This contribution is the supporting theories to our algorithms. It proves the convergence of the Q-Q distance in high probability when there is no change in the data stream and gives a lower bound on the choice of sample size.

For each of the contributions, we have presented both qualitative and quantitative results to demonstrate their effectiveness. The Q-Q distance based quickest detection algorithm is to appear in [Yang and Qi, 2010c]. We have submitted the decentralized nonparametric quickest detection algorithm [Yang and Qi, 2010a] for review. Further, a general overview paper with definition of the distance measure, the nonparametric quickest detection algorithms, and the theoretical analysis will be submitted [Yang and Qi, 2010b] for review. With this summary of the contributions, we now turn to the future directions for this research.

7.2 Directions for Future Research

The ideas and concepts in this dissertation offer a great deal of possible future research directions. We here identify the following areas as what we think the most important.

7.2.1 Automatic Threshold Selection

The first important area for future research is in automatic selection of the threshold(s). We have shown that the threshold is approximately linearly related to the detection delay, as the threshold goes to infinity. Although the detection procedure itself is not sensitive to the threshold, by choosing the appropriate threshold, the detection procedure could

produce better performance. Automatic threshold selection is not impossible but does require additional research.

7.2.2 Detection with Dependent Observations

So far all our discussions are based on the assumption that all elements in the data stream are independent of one another. However, there are situations where dependent elements need to be monitored. In parametric detection, there have been discussions such as to extend the optimality of the CUSUM to certain dependent situations, or to use some local hypothesis approach [Poor and Hadjiliadis, 2009]. We believe similar methods could be used to relax the assumption of independence among elements in nonparametric detection research.

7.3 Closing Remarks

This whole work is based on the use of Q-Q plot which I accidentally found on the Internet when I was looking for a better way to plot an empirical probability density function. Its definition suggests that it can be adapted in quickest detection to differentiate the distributions underlying the data streams but there are really few Q-Q plot applications that we can refer to. We had to dig deeper to make it work in quickest detection. Further, we have guaranteed the detection performance analytically and numerically. Obviously, our implementation of a nonparametric quickest detection framework does not solve all the problems in the quickest detection arena but hopefully it does provide a tiny step towards extending the state of the art in quickest detection.

> "I FIND THAT A GREAT PART OF THE INFORMATION I HAVE WAS ACQUIRED BY LOOKING UP SOMETHING AND FINDING SOMETHING ELSE ON THE WAY." – FRANKLIN P. ADAMS

> "ANYTHING ONE MAN CAN IMAGINE, OTHER MEN CAN MAKE REAL." – JULES VERNE

Bibliography

Bibliography

[KDD, 1999] (1999). KDD Cup. The third international knowledge discovery and data mining tools competition. http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html.

[Anderson, 1980] Anderson, J. P. (1980). Computer security threat monitoring and surveillance. Technical report, James P. Anderson Co.

[Andersson et al., 2004] Andersson, E., Bock, D., and Frisén, M. (2004). Detection of turning points in business cycles. *Journal of Business Cycle Measurement and Analysis*, 1(1):93–108.

[Andersson et al., 2006] Andersson, E., Bock, D., and Frisén, M. (2006). Some statistical aspects of methods for detection of turning points in business cycles. *Journal of Applied Statistics*, 33(3):257–278.

[Andreou and Ghyels, 2004] Andreou, E. and Ghyels, E. (2004). The impact of sampling frequency and volatility estimators on change-point tests. *Journal of Financial Econometrics*, 2(2):290–318.

[Andrews et al., 1996] Andrews, D., Lee, I., and Ploberger, W. (1996). Optimal change-point tests for normal linear regression. *Journal of Econometrics*, 70(1):9–38.

[Basseville and Nikiforov, 1993] Basseville, M. and Nikiforov, I. (1993). *Detection of Abrupt Changes: Theory and Application*. Prentice-Hall.

[Berkes et al., 2004] Berkes, I., Gombay, E., Horváth, L., and Kokoszka, P. (2004). Sequential change-point detection in GARCH(p,q) models. *Econometric Theory*, 20(6):1140–1167.

[Brännäs and Westlund, 1980] Brännäs, K. and Westlund, A. (1980). *Lecture Notes in Control and Information Sciences*, chapter On the Recursive Estimation of Stochastic and Time-varying Parameters in Econometric Systems, pages 414–422. Springer Berlin/Heidelberg.

[Cardenas et al., 2004] Cardenas, A., Baras, J., and Ramezani, V. (2004). Distributed change detection for worms, ddos and other network attacks. *Proceedings of the 2004 American Control Conference*, 2:1008–1013. Boston, MA.

[Chang, 2002] Chang, R. (2002). Defending against flooding-based distributed denial-of-service attacks: a tutorial. *IEEE Communications Magazine*, 40(10):42–51.

[Das and Resnick, 2008] Das, B. and Resnick, S. (2008). Qq plot, random sets and data from a heavy tailed distribution. *Stochastic Models*, 24(1):103–132.

[Denning, 1987] Denning, D. (1987). An intrusion detection model. *IEEE Transactions on Software Engineering*, SE-13(2):222–232.

[Dvoretzky et al., 1956] Dvoretzky, A., Kiefer, J., and Wolfowitz, J. (1956). Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *Annals of Mathematical Statistics*, 27(3):642–669.

[Gilchrist, 2000] Gilchrist, W. (2000). *Statistical Modelling with Quantile Functions*. Chapman & Hall/CRC.

[Girshick and Rubin, 1952] Girshick, M. and Rubin, H. (1952). A bayes approach to a quality control model. *The Annals of Mathematical Statistics*, 23(1):114–125.

[Gordon and Pollak, 1994] Gordon, L. and Pollak, M. (1994). An efficient sequential nonparametric scheme for detecting a change in distribution. *The Annals of Statistics*, 22(2):763–804.

[Karatzas, 2003] Karatzas, I. (2003). A note on bayesian detection of change-points with an expected miss criterion. *Statistics and Decisions*, 21(1):3–14.

[Kifer et al., 2004] Kifer, D., Ben-David, S., and Gehrke, J. (2004). Detecting change in data streams. *Proceedings of the Thirtieth International Conference on Very Large Data Bases*, 30:180–191.

[Kim et al., 2004] Kim, H., Rozovskii, B., and Tartakovsky, A. (2004). A nonparametric multichart CUSUM test for rapid detection of dos attacks in computer netwetwork. *International Journal of Computing and Information Sciences*, 2(3):149–158.

[Lai, 1995] Lai, T. (1995). Sequential change-point detection in quality control and dynamical systems (with discussion). *Journal of the Royal Statistical Society B*, 57(4):613–658.

[Lam and Suen, 1997] Lam, L. and Suen, S. (1997). Application of majority voting to pattern recognition: An analysis of its behavior and performance. *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans*, 27(5):553–568.

[Lippmann et al., 2000] Lippmann, R., Haines, J. W., Fried, D. J., Korba, J., and Das, K. (2000). The 1999 DARPA off-line intrusion detection evaluation. *Computer Networks: The International Journal of Computer and Telecommunications Networking*, 34(4):579–595.

[Lorden, 1971] Lorden, G. (1971). Procedures for reacting to a change in distribution. *Annals of Mathematical Statistics*, 42(6):1897–1908.

[Massart, 1990] Massart, P. (1990). The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *The Annals of Probability*, 18(3):1269–1283.

[Mei, 2005] Mei, Y. (2005). Information bounds and quickest change detection in decentralized decision systems. *IEEE Transactions on Information Theory*, 51(7):2669–2681.

[Moustakides, 1986] Moustakides, G. (1986). Optimal stopping times for detecting changes in distributions. *Annals of Statistics*, 14(4):1379–1387.

[Moustakides, 2006] Moustakides, G. (2006). Decentralized CUSUM change detection. *Proceedings of the* 9^{th} *International Conference on Information Fusion.* Florence, Italy.

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

[Page, 1955] Page, E. (1955). A test for a change in a parameter occurring at an unknown point. *Biometrika*, 42(3-4):523–527.

[Pelkowitz, 1987] Pelkowitz, L. (1987). The general discrete time disorder problem. *Stochastics*, 20(2):89–110.

[Pollak, 1985] Pollak, M. (1985). Optimal detection of a change in distribution. *Annals of Statistics*, 13(1):206–227.

[Poor, 1998] Poor, H. (1998). Quickest detection with exponential penalty for delay. *Annals of Statistics*, 26(6):2179–2205.

[Poor and Hadjiliadis, 2009] Poor, H. and Hadjiliadis, O. (2009). *Quickest Detection*. Cambridge University Press.

[Raghavan and Veeravalli, 2008] Raghavan, V. and Veeravalli, V. (2008). Quickest detection of a change process across a sensor array. *Proceedings of the* 11^{th} *International Conference on Information Fusion*. Cologne, Germany.

[Ritov, 1990] Ritov, Y. (1990). Decision theoretic optimality of the CUSUM procedure. *Journal of Applied Statistics*, 18(3):1464–1469.

[Roberts, 1966] Roberts, S. (1966). A comparison of some control chart procedures. *Technometrics*, 8(3):411–430.

[Savage, 1956] Savage, I. (1956). Contributions to the theory of rank order statistics–the one-sample case. *The Annals of Mathematical Statistics*, 30(4):1018–1023.

[Shao, 2003] Shao, J. (2003). *Mathematical Statistics*. New York : Springer.

[Shewhart, 1931] Shewhart, W. (1931). *Economic Control of Quality of Manufactured Product*. Princeton, NJ: Van Nostrand Reinhold Co.

[Shiryaev, 1963] Shiryaev, A. (1963). On optimum methods in quickest detection problems. *Theory of Probability and Its Applications*, 8(1):22–46.

[Shiryaev, 1978] Shiryaev, A. (1978). *Optimal Stopping Rules*. Springer-Verlag, New York.

[Shorack, 2000] Shorack, G. (2000). *Probability for Statisticians*. Springer.

[Tartakovsky and Kim, 2006] Tartakovsky, A. and Kim, H. (2006). Performance of certain decentralized distributed change detection procedures. *Proceedings of the* 9^{th} *International Conference on Information Fusion*. Florence, Italy.

[Tartakovsky and Polunchenko, 2008] Tartakovsky, A. and Polunchenko, A. (2008). Quickest changepoint detection in distributed multisensor systems under unknown parameters. *Proceedings of the* 11^{th} *International Conference on Information Fusion*. Cologne, Germany.

[Tartakovsky et al., 2006a] Tartakovsky, A., Rozovskii, B., Blazek, R., and Kim, H. (2006a). Detection of intrusions in information systems by sequential change-point methods. *Statistical Methodology*, 3(3):252–293.

[Tartakovsky et al., 2006b] Tartakovsky, A., Rozovskii, B., Blazek, R., and Kim, H. (2006b). A novel approach to detection of intrusions in computer networks via adaptive sequential and batch-sequential change-point detection methods. *IEEE Transactions on Signal Processing*, 54(9):3372–3382.

[Tartakovsky and Veeravalli, 2003] Tartakovsky, A. and Veeravalli, V. (2003). Quickest change detection in distributed sensor systems. *Proceedings of the* 6^{th} *International Conference on Information Fusion*, pages 756–763. Cairns, Australia.

[Tartakovsky and Veeravalli, 2008] Tartakovsky, A. and Veeravalli, V. (2008). Asymptotically optimal quickest change detection in distributed sensor systems. *Sequential Analysis*, 27(4):441–475.

[Teneketzis and Varaiya, 1984] Teneketzis, D. and Varaiya, P. (1984). The decentralized quickest detection problem. *IEEE Transactions on Automatic Control*, AC-29(7):641–644.

[Tenney and Sandell, 1981] Tenney, R. and Sandell, N. (1981). Detection with distributed sensors. *IEEE Transactions on Aerospace and Electronic Systems*, AES-17(4):501–510.

[Vapnik, 1998] Vapnik, V. (1998). *Statistical Learning Theory*. Wiley-Interscience.

[Vapnik and Chervonenkis, 1971] Vapnik, V. and Chervonenkis, A. (1971). On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16(2):264–280.

[Veeravalli, 1999] Veeravalli, V. (1999). Sequential decision fusion: Theory and applications. *Journal of the Franklin Institute*, 336(2):301–322.

[Veeravalli, 2001] Veeravalli, V. (2001). Decentralized quickest change detection. *IEEE Transactions on Information Theory*, 47(4):1657–1665.

[Wasserman, 2005] Wasserman, L. (2005). *All of Nonparametric Statistics*. Springer.

[Wilk and Gnanadesikan, 1968] Wilk, M. and Gnanadesikan, R. (1968). Probability plotting methods for the analysis of data. *Biometrika*, 55(1):1–17.

[Yang and Qi, 2010a] Yang, D. and Qi, H. (2010a). An effective approach to nonparametric quickest detection and its decentralized realization. *EURASIP Journal on Advances in Signal Processing*. Submitted for review.

[Yang and Qi, 2010b] Yang, D. and Qi, H. (2010b). A new effective nonparametric quickest detection approach. *IEEE Transactions on Signal Processing*. Submitted for review.

[Yang and Qi, 2010c] Yang, D. and Qi, H. (2010c). A new nonparametric quickest detection procedure. *Proceedings of the* 35^{th} *International Conference on Acoustics, Speech, and Signal Processing*. Dallas, Texas.

[Yashchin, 1997] Yashchin, E. (1997). Change-point models in industrial applications. *Nonlinear Analysis*, 30(7):3997–4006.

Appendix

Proof of Theorem 4.1.5. Suppose $F_m \to F$. Let $F(x) = t, t \in (0,1)$ and $z \equiv F^{-1}(t)$. x is a continuity point of F.

When $x < z$, we have

$$\begin{aligned} F(x) < t &\Rightarrow F_m(x) < t \quad \text{for} \quad m \geq M_x \\ &\Rightarrow F_m^{-1}(t) \geq x \quad \text{for} \quad m \geq M_x \\ &\Rightarrow \lim_{m\to\infty} F_m^{-1}(t) \geq x \end{aligned} \tag{1}$$

where M_x is some integer number. Since there are continuity points x closing in from the left to z, we have

$$\lim_{m\to\infty} F_m^{-1}(t) \geq x \Rightarrow \lim_{m\to\infty} F_m^{-1}(t) \geq z \tag{2}$$

When $x > z$, we have

$$\begin{aligned} F(x) > t &\Rightarrow F_m(x) > t \quad \text{for} \quad m \geq M_x \\ &\Rightarrow F_m^{-1}(t) \leq x \quad \text{for} \quad m \geq M_x \\ &\Rightarrow \lim_{m\to\infty} F_m^{-1}(t) \leq x \end{aligned} \tag{3}$$

Since there are continuity points x closing in from the right to z, we have

$$\lim_{m\to\infty} F_m^{-1}(t) \leq x \Rightarrow \lim_{m\to\infty} F_m^{-1}(t) \leq z \tag{4}$$

From Equation 2 and Equation 4,

$$\lim_{m\to\infty} F_m^{-1}(t) = z$$

and

$$\lim_{m\to\infty} F_m^{-1}(t) = F^{-1}(t) \tag{5}$$

That is, $Q_m(t) \to Q(t)$ for all but at most a countable number of t's. The proof of the converse is similar. □

Proof of Theorem 4.2.6. The left side of the inequality can be rewritten as follows,

$$\mathbb{P}\left(|Q_m(t) - Q(t)| > \epsilon\right) = \mathbb{P}\left(Q_m(t) > Q(t) + \epsilon\right) + \mathbb{P}\left(Q_m(t) < Q(t) - \epsilon\right) \tag{6}$$

From Equation 3.1, we can see that $F(x) \geq t$ if and only if $x \geq F^{-1}(t)$ for any c.d.f. F on $\mathbb{R}$. So we have

$$\begin{aligned}
\mathbb{P}(Q_m(t) > Q(t) + \epsilon) &= \mathbb{P}(t > F_m(Q(t) + \epsilon)) \\
&= \mathbb{P}(F(Q(t) + \epsilon) - F_m(Q(t) + \epsilon) > F(Q(t) + \epsilon) - t) \\
&\leq \mathbb{P}(\sup_{x\in\mathbb{R}} |F_n m(x) - F(x)| > \delta_\epsilon) \\
&\leq 2e^{-2m\delta_\epsilon^2}
\end{aligned} \tag{7}$$

and

$$\begin{aligned}
\mathbb{P}(Q_m(t) < Q(t) - \epsilon) &= \mathbb{P}(t < F_m(Q(t) - \epsilon)) \\
&= \mathbb{P}(F_m(Q(t) - \epsilon) - F(Q(t) - \epsilon) > t - F(Q(t) - \epsilon)) \\
&\leq \mathbb{P}(\sup_{x\in\mathbb{R}} |F_m(x) - F(x)| > \delta_\epsilon) \\
&\leq 2e^{-2m\delta_\epsilon^2}
\end{aligned} \tag{8}$$

This proves 4.16. □

Vita

Dayu Yang was born in Chengdu, China, on March 14th, 1974, the second son of Yunmin Yang and Shunying Yu. After graduating from Chongqing No.1 Middle School in 1992, he attended Chongqing University of Posts & Telecommunications where he received a Bachelor of Science degree in Communications Engineering. In 1996, he joined China Mobile and worked there for five years as a senior engineer. In 2001, He attended University of Tennessee at Knoxville where he received Master of Science degree in Electrical Engineering. He returned to University of Tennessee at Knoxville in 2006 and joined the Advanced Imaging & Collaborative Information Processing (AICIP) Lab at as a graduate research assistant where he completed his Doctor of Philosophy degree in January 2009.