ROBUST SPARSE LEARNING AND MONITORING OF HIGH-DIMENSIONAL DATA

A Dissertation
Presented to
The Academic Faculty

By

Ruizhi Zhang

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
H. Milton Stewart School of Industrial and Systems Engineering

Georgia Institute of Technology

August 2019

ROBUST SPARSE LEARNING AND MONITORING OF HIGH-DIMENSIONAL DATA

Approved by:

Dr. Yajun Mei, Advisor
School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Jianjun Shi, Advisor
School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Kamran Paynabar
School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Brani Vidakovic
School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Jie Chen
Department of Population Health Sciences, Medical College of Georgia
Augusta University

Date Approved: May 3, 2019

To my beloved parents, wife Shanshan, for their love and encouragement.

ACKNOWLEDGEMENTS

First, I would like to express my sincere gratitude to my advisor, Professor Yajun Mei, for his continuous guidance, devoted supervision, and support during my Ph.D. study and career development. Without his support, this thesis would not have been accomplished. I not only learn a lot of statistical and mathematical techniques but also learn the passion, attitude, and enthusiasm for research from him.

Second, I would like to express my deepest gratitude to my co-advisor, Professor Jianjun Shi, for his valuable advice, support, and guidance for my research and job search. Moreover, I learn a lot of experience and skills about collaborating with industrial partners and solve problems in the real world from him, which undoubtedly help me develop my future career.

In addition, I would like to express appreciation to my committee members, Professor Brani Vidakovic for his recommendation and encouragement during my job search, Professor Kamran Paynabar, and Professor Jie Chen for their productive discussion, constructive comments, and insightful advice to improve this dissertation. I am also grateful to Dr. Jian Wang and Professor Huan Xu for their valuable advice and collaboration on my research projects.

My gratitude also goes to Professor Roshan Joseph, Professor Vladimir Koltchinskii, Professor Arkadi Nemirovski, Professor Nicoleta Serban, Professor Yao Xie, Professor Tuo Zhao, who have taught me excellent and insightful courses during my Ph.D. study. A special thanks go to Professor Jeff Wu for his encouragement and valuable advice during my job search.

Moreover, I would like to thank Mr. Matthew Hagen and Mr. Xiquan Cui for their guidance and supports on my summer internship and all the colleagues in the data science team at the Home Depot for the collaborations and opportunities they have provided me.

I would like to thank all my past and present colleagues in my advisors research groups:

Dr. Hao Yan, Dr. Xiaowei Yue, Dr. Xiaolei Fang, Mr. Mostafa Reisi Gahrooei, Mr. Andi Wang, Ms. Xinran Shi, Dr. Mohammad Nabhan, Dr. Chen Zhang, Mr. Zhen Zhong, Dr. Yuan Wang, Dr. Kun Liu, Dr. Tony Yaacoub, Ms. Wanrong Zhang, Ms. Yujie Zhao. I would also like to extend a special thanks to Dr. Qingqing Liu, Dr. Rundong Du, Dr. Yuanshuo Zhao, Ms. Jing Qin, and Dr. Ran Li for their kindly and friendly help when I first moved to Atlanta.

I would also like to thank my beloved parents for teaching me the value of knowledge, and for always being with me and supporting me through my life. Most importantly, I want to thank my wife Shanshan Cao, who accompanies me throughout my undergraduate and Ph.D. journey. During the past eight years, she has always been there, supporting me and encouraging me. Without her, I would not be able to complete my Ph.D. study and achieve this important milestone in my life.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

With the rapid development of advanced sensing technology, rich and complex real-time high-dimensional streaming data are available in many systems, such as manufacturing, wireless communication, biosurveillance, and social systems. As information is accumulated over time at a fast rate by multiple sensors, it is highly desirable to develop efficient methodologies that enable to (1) extract informatic features, (2) learn the process status and detect possible changes or faults quickly, (3) implement and compute online fast, (4) be robust to outliers or model misspecification. Therefore, efficient robust and scalable schemes and algorithms, which enable real-time monitoring of high-dimensional data streams, are highly demanded.

This thesis focuses on statistical modeling to extract informative and robust features, to interpret the characteristic of the system, and to develop efficient and robust monitoring schemes that can be implemented recursively and in parallel to reduce unnecessary transition costs in the data fusion systems. The methodologies developed in the thesis are generic and can be applied to a variety of fields ranging from manufacturing processes (e.g. forging, stamping processes, semiconductor process), where functional profile data are observed sequentially, to video monitoring (e.g. Solar flare detection), where image data are collected for sequential decision making.

This thesis starts with theoretical research on change-point detection and robust M-estimation. In Chapter 1, we propose a scalable robust monitoring scheme that can detect the small but systematic change of the system efficiently and in real-time when there are some random transient outliers. We construct a new robust local detection statistic called L_α-CUSUM statistic that can reduce the effect of outliers by using the Box-Cox transformation of the likelihood function. Moreover, we propose a new concept called false-alarm breakdown point to measure the robustness of online monitoring schemes and characterize the breakdown point of our proposed schemes.

In Chapter 2, we develop some families of communication-efficient schemes for monitoring large-scale data streams. We use some shrinkage transformations such as soft-thresholding, hard-thresholding and order-thresholding on the local monitoring statistics so that to filter out unaffected data streams and save communication costs in the data fusion networks. Moreover, we conduct the detection delay analysis on our proposed schemes in both classical low-dimensional regime and modern high-dimensional regime and show that under certain conditions, our schemes are asymptotical optimal by only receiving a small proportion of data, which can reduce the transition costs.

In Chapter 3, we investigate two important properties of M-estimator, namely, robustness and tractability, in linear regression setting, when the observations are contaminated by some arbitrary outliers. By learning the landscape of the empirical risk, we show that under mild conditions when the percentage of outliers is small, many M-estimators enjoy nice robustness, which means the estimator is close to the true underlying parameter, and tractability properties, which means the estimator can be computed efficiently, even if the loss function is non-convex.

Then, in Chapter 4, we work on the applied research on nonlinear profile monitoring based on discrete Wavelet transform. We proposed the recursive CUSUM procedure that can learn the out-of-control parameters adaptively and detect unknown change efficiently. In Chapter 5, we develop a functional Poisson regression model for papers cumulative citations data. Based on our model, we can fit and learn the individual papers citation characteristic well. Our proposed model is also used for clustering different citation patterns, which can provide implications for bibliometric studies and research evaluations. Finally, we summarize our original contributions and future research plans in Chapter 6.

CHAPTER 1
ROBUST REAL-TIME MONITORING OF HIGH-DIMENSIONAL DATA STREAMS

1.1 Introduction

Robust statistics have been extensively studied in the offline context when the full data set is available for decision making and is contaminated with outliers, e.g., robust estimation (Huber, 1964; Basu, Harris, Hjort, and Jones, 1998), robust hypothesis testing (Huber, 1965; Heritier and Ronchetti, 1994), and robust regression (Yohai, 1987; Cantoni and Ronchetti, 2001). Also see the classical books, Huber and Ronchetti (2009) or Hampel, Ronchetti, Rousseeuw, and Stahel (2011), for literature review. In this chapter, we propose to develop robust methods in the context of online monitoring when one is interested in detecting sparse persistent smaller changes in high-dimensional streaming data under the contamination of transient larger outliers.

A concrete motivating example of our research is profile monitoring in a progressive forming process, see Figure 1.1 for illustration. A progressive forming process has a set of dies installed within one stamping press. The part is transferred from one die station to the next die station sequentially and each die station has a formed part processed in previous die station. During this process, the forming force measured by the tonnage sensor installed in the linkage of press is the summation of all forming forces generated in each die. The forming force is measured as a profile or functional data that consists of $2^{11} = 2048$ measurements points. As a work piece passes through the die stations, a fault in any die station might change the forming force (e.g. tonnage profiles). Figure 1.2 plots some typical patterns of the profile data under the normal condition as well as under two faulty conditions: fault #1 (the smaller change) caused by the malfunction of a part transferred

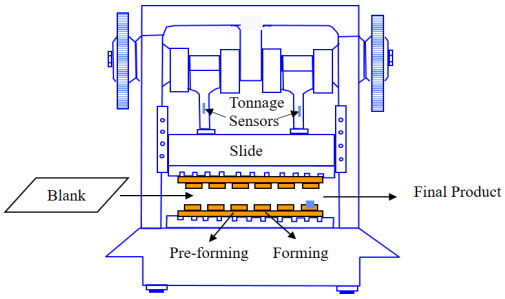

Figure 1.1: Illustration of a progressive forming process.

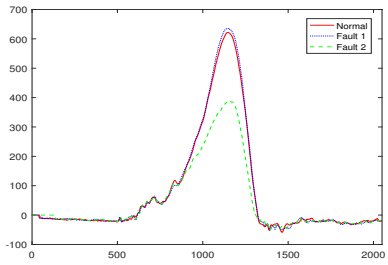

Figure 1.2: Three samples from a forming process.

in the forming station, and fault #2 (the larger change) due to missing operations in the pre-forming station. In practice, it is difficult to detect the smaller fault #1 condition since the difference between the fault #1 profile and the normal profile is sparse and small in magnitude. However, if this fault is neglected and the faulty condition remains uncovered, it will lead to persistent quality issues of formed parts, and further damage die. Meanwhile, the larger fault #2 can be observed easily due to the large difference from the normal profile. On one hand, line workers generally will be able to fix the corresponding root cause in the pre-forming station. On the other hand, the workers are generally unable to check whether it will affect the down-stream stations or not, and thus it may or may not lead the fault #1 condition. Hence, when monitoring high-dimensional data streams, it is highly desirable to develop effective methodologies to detect those smaller but persistent changes in the presence of infrequent larger changes which can be thought as outliers, and might or might not related to the smaller persistent changes.

In general, the problem of robust monitoring high-dimensional data in the presence of outliers occurs in many real-world applications such as industrial quality control, bio-surveilance, key infrastructure or internet traffic monitoring, in which sensors are deployed to constantly monitor the changing environment, see Shmueli and Burkom (2010), Tartakovsky, Polunchenko, and Sokolov (2013), and Yan, Paynabar, and Shi (2015). Unfortunately, it is highly non-trivial to develop efficient robust monitoring schemes or algorithms due to two challenges: (1) the sparsity, where only a few unknown local components or

features of data might be affected, but we do not know which local components or features are affected; and (2) the robustness, where we are interested in detecting smaller persistent changes, not the transient outliers.

In the sequential change-point literature for high-dimensional data, while the sparsity issue has been investigated, no research has been done on the robustness issue. To be more specific, the sparsity has been first addressed by Xie and Siegmund (2013) using a semi-Bayesian approach, and later by Wang and Mei (2015) using shrinkage-estimation-based schemes. Chan (2017) developed asymptotic optimality theory for large-scale independent Gaussian data streams. Unfortunately all these methods are sensitive to outliers since they are based on the likelihood function of specific parametric models (e.g.. Gaussian) of the observations. Meanwhile, regarding the robustness issue, research is available for monitoring one-dimensional streaming data: rank-based method in Gordon and Pollak (1994) and Gordon and Pollak (1995), kernel-based method in Desobry, Davy, and Doncarli (2005), or least-favorable-distribution method in Unnikrishnan, Veeravalli, and Meyn (2011). Unfortunately it is unclear how to extend these existing robust methods from one-dimension to high-dimension when we also need to deal with the sparsity issue in which there is uncertainty on the subset of affected local components or features.

In this chapter, we develop efficient robust real-time monitoring schemes that are able to robustly detect smaller persistent changes in the presence of transient outliers when online monitoring of high-dimensional steaming data. From the methodology viewpoint, our proposed schemes are semi-parametric, and extend two contemporary concepts to the context of online monitoring of high-dimensional data streams: (i) L_q-likelihood in Ferrari and Yang (2010) and Qin and Priebe (2017) for robustness, and (ii) the sum-shrinkage technique in Liu, Zhang, and Mei (2019) for sparsity. These allow us to develop statistical efficient and computationally simple schemes that can be implemented recursively over time for robust real-time monitoring of high-dimensional data streams. Moreover, we also extend the concept of breakdown in the offline robust statistics (Hampel, 1968) to the

sequential change-point detection context, and conduct the false alarm breakdown point analysis, which turns out to be useful for tuning parameters in our proposed schemes.

Our research makes four contributions in the statistics field by combining robust statistics with sequential change-point detection for high-dimensional streaming data. First, our proposed method is robust with respect to infrequent outliers as well as the uncertainty of affected components of the data. Second, our proposed method can be implemented recursively and distributed via parallel computing, and thus is suitable for real-time monitoring over long time period for high-dimensional data. Third, inspired by the concept of breakdown point (Hampel, 1968) in the offline robust statistics, we propose a novel concept of false alarm breakdown point to quantify the robustness of any online monitoring schemes, and show that our proposed scheme is indeed has much larger false alarm breakdown point than the classical CUSUM-based schemes. Finally, from the mathematical viewpoint, we use Chebyshev's inequality to derive non-asymptotic low bounds on the average run length of false alarm for our proposed method. The non-asymptotic results hold regardless of dimensionality, and allow us to provide a deep insight on the effect of high-dimensionality in the context of change-point detection under the modern asymptotic regime when the dimension or the number of data streams goes to ∞.

The remainder of this chapter is organized as follows. In Section 1.2, we start with the modern assumptions and present our proposed scheme in three steps. Then we provide the theoretical properties of our proposed scheme in Section 1.3. In Section 1.4, we introduce the concept of false alarm breakdown point and propose the general method to choose the robust tuning parameter α. Simulation and case study results are presented in Section 1.5 and Section 1.6 respectively. The proofs of our main theorems are postponed to Section 1.7.

1.2 Our proposed scheme

Suppose we are monitoring a sequence of high-dimension streaming data, $\{\mathbf{Y}_n\}$, over time step $n = 1, 2, \cdots$, where the data might be corrupted with transient outliers. We want to raise an alarm as quickly as possible if there is a persistent distribution change on the data, but we prefer to take observations without any actions if there are no persistent distribution changes or if there are only transient outliers.

In this section, we will present the description of our proposed scheme, and then develop its asymptotic properties in next section, with the focus on the effect of the high-dimensionality in the context of change-point detection. At the high-level, our proposed scheme includes three components: (i) modeling extracted features, (ii) monitoring each local feature individually in parallel, and then (iii) combines local detection statistics together to make an online global-level decision. For the purpose of easy understanding, we split the presentation of our proposed scheme into three subsections, and each subsection focuses on each component of the proposed scheme.

1.2.1 Data and model

In many real-world applications such as profile monitoring in Figure 1.2, each raw data is independent over time, but local coordinates of each high-dimensional data can be dependent. In such a case, a standard technique is to extract independent features from the historical in-control data using principal component analysis (PCA), wavelets, tensor-decomposition, etc., and then monitor the feature coefficients instead of raw data themselves, see Jin and Shi (1999), Chang and Yadama (2010), Yan, Paynabar, and Shi (2015), Liu, Mei, and Shi (2015), and Paynabar, Zou, and Qiu (2016). In the context of off-line estimation or prediction, one can focus on a few important features for the purpose of dimension reduction. However, a new challenge in the monitoring context is that we do not know which features might be affected by the change, and thus one often needs to monitor

a relatively large number of features, see Wang, Mei, and Paynabar (2018) and Zhang, Mei, and Shi (2018).

For each high-dimensional raw data $\mathbf{Y}_n$, denote the corresponding K-dimensional feature coefficients as $\mathbf{X}_n = (X_{1,n}, \cdots, X_{K,n})^T$. We assume that the local features are independent, and we have sufficient historical in-control data to model the pre-change cumulative density function (cdf) F_k of the k^{th} feature $X_{k,n}$'s. Without loss of generality, we assume that the $X_{k,n}$'s have the identical distribution, say, with the same probability density function (pdf) $f_{\theta_0} =$ pdf of $N(0,1)$, under the in-control state, as we can consider the transformation $\Phi^{-1}(F_k(\cdot))$, where Φ is the cdf of the standard normal distribution, to standardize or normalized the in-control data if needed, see Efron (2012). Furthermore, as in our motivating example of profile monitoring in Figure 1.1, we further assume the $X_{k,n}$'s will have pdf g when the raw data involves larger transient changes or outliers, and will have pdf f_θ when the raw data involves a smaller persistent change, where the unknown post-change parameter $\theta \geq \theta_1$ for some known value $\theta_1 > 0$.

Mathematically, recall the Tukey-Huber's gross error model of the two-component mixture densities

$$h_\theta(x) = (1-\epsilon) f_\theta(x) + \epsilon g(x), \tag{1.1}$$

where $\epsilon \in [0,1)$ is referred to as the contamination/outlier ratio and g is the (unknown) outlier distributions. Then we model the $X_{k,n}$'s as the following change-point Tukey-Huber's gross error model: for some unknown change time $\nu = 1, 2, \cdots$, all $X_{k,n}$'s are independent and identically distributed (i.i.d.) with $h_{\theta_0}(x)$ in (1.1) when $n \leq \nu - 1$, but m out of K local streams $X_{k,n}$'s have another distribution $h_\theta(x)$ in (1.1) when $n \geq \nu$, where the post-change parameter $\theta \geq \theta_1$, and $\theta_1 - \theta_0$ is the smallest meaningful magnitude of the change, which is pre-specified.

In the sequential change-point problem, at each and every time step, we need to test the

null hypothesis

$$H_0 : \nu = \infty \quad \text{(i.e., no persistent change occurs)}$$

against a composite alternative hypothesis

$$H_1 : \nu = 1, 2, \cdots \quad \text{(i.e., a persistent change occurs at some finite time).}$$

The statistical procedure in the sequential change-point problem is often defined as a stopping time T that represents the time when we raise an alarm to declare that a change has occurred. Here T is an integer-valued random variable, and the decision $\{T = t\}$ is based only on the observations in the first t time steps. Denote by $\mathbf{P}_{\theta_0}^{(\infty)}$ and $\mathbf{E}_{\theta_0}^{(\infty)}$ the probability measure and expectation when the data $X_{k,n}$'s are i.i.d. with density h_{θ_0}, and denote by $\mathbf{P}_{\theta}^{(\nu)}$ and $\mathbf{E}_{\theta}^{(\nu)}$ the same when the change occurs at time ν and m out of K streams $X_{k,n}$'s have the post-change distribution h_θ. Under the standard minimax formulation for online change-point detection (Lorden, 1971), the performance of a stopping time T is evaluated by the average run length to false alarm (ARLFA), $\mathbf{E}_{\theta_0}^{(\infty)}(T)$ and the worst-case detection delay

$$D_{\epsilon,\theta}(T) = \sup_{\nu \geq 1} \operatorname{ess\,sup} \mathbf{E}_{\theta}^{(\nu)} \left((T - \nu + 1)^+ \big| \mathcal{F}_{\nu-1} \right). \tag{1.2}$$

Here $\mathcal{F}_{\nu-1} = (X_{1,[1,\nu-1]}, \ldots, X_{K,[1,\nu-1]})$ denotes past global information at time ν, $X_{k,[1,\nu-1]} = (X_{k,1}, \ldots, X_{k,\nu-1})$ is past local information for the k-th feature.

An efficient detection procedure T should have small detection delay $D_{\epsilon,\theta}(T)$ subject to the false alarm constraint

$$\mathbf{E}_{\theta_0}^{(\infty)}(T) \geq \gamma \tag{1.3}$$

for some pre-specified large constant $\gamma > 0$.

We should acknowledge that this is the standard formulation for monitoring of one- or

low- dimensional data, and many classical procedures have been developed such as Page's CUSUM procedure (Page, 1954), Shiryaev-Roberts procedure (Shiryaev, 1963; Roberts, 1966), window-limited procedures (Lai, 1995) and scan statistics (Glaz, Naus, Wallenstein, Wallenstein, and Naus, 2001). Also some fundamental optimality results for one-dimensional data were established in Shiryaev (1963), Lorden (1971), Pollak (1985), Pollak (1987), Moustakides (1986), Ritov (1990), and Lai (1995), etc. For a review, see the books such as Basseville and Nikiforov (1993), Poor and Hadjiliadis (2009), and Tartakovsky, Nikiforov, and Basseville (2014). Note that here we do not aim to develop optimality theorem for monitoring of high-dimensional data, which is still an open problem in a general setting. Our main objective is to develop an efficient and robust scheme, and then to investigate its statistical properties, which shed the new light of the effect of the dimensionality K on the high-dimensional change-point detection problem.

1.2.2 Robust local statistics

To develop real-time robust monitoring schemes, we propose to borrow the parallel computing technique to monitor each local feature individually, and then use the sum-shrinkage technique to combine the local monitoring statistics together to make a global decision. For that purpose, it is crucial to have an efficient local monitoring statistic that is robust to outliers. To do so, for the k^{th} local feature, we propose to define a new local L_α-CUSUM statistic:

$$W_{\alpha,k,n} \;=\; \max\Big(W_{\alpha,k,n-1} + \frac{[f_{\theta_1}(X_{k,n})]^\alpha - [f_{\theta_0}(X_{k,n})]^\alpha}{\alpha},\; 0\Big), \tag{1.4}$$

for $n \geq 1$, and $W_{\alpha,k,0} = 0$. Here $\alpha \geq 0$ is a tuning parameter that can control the tradeoff between statistical efficiency and robustness under the gross error model in (1.1) and its suitable choice will be discussed later.

The motivation of our L_α-CUSUM statistic in (1.4) is as follows. Recall that when

locally monitoring the single k^{th} data stream $X_{k,n}$ with a possible local distribution change from f_{θ_0} to f_{θ_1}, the generalized likelihood ratio test becomes the classical CUSUM statistic $W^*_{k,n}$, which has a recursive form:

$$W^*_{k,n} = \max\left(W^*_{k,n-1} + \log\frac{f_{\theta_1}(X_{k,n})}{f_{\theta_0}(X_{k,n})},\ 0\right). \tag{1.5}$$

The CUSUM statistic enjoys nice optimality properties when all models are fully correctly specified (Moustakides, 1986), but unfortunately it is very sensitive to the outliers as in all other likelihood based methods in offline statistics. One recent idea in offline robust statistics is to replace the log-likelihood statistic $\log f(X)$ by L_α-likelihood function $([f(X)]^\alpha - 1)/\alpha$ for some $\alpha > 0$, see Ferrari and Yang (2010) and Qin and Priebe (2017). At the high-level, L_α-likelihood function is bounded below by $-1/\alpha$ when $f(X) \to 0$ for outliers, and thus become more much robust to outliers as compared to the log-likelihood statistics. Moreover, as $\alpha \to 0$, the L_α-likelihood function converges to the log-likelihood statistic, and thus it keeps statistical efficiencies when α is small. Here we apply this idea to develop L_α-CUSUM statistics that turns out to be robust to outliers. More rigorous robust properties will be discussed later in Section 1.4.

1.2.3 Efficient global monitoring statistics

With local L_α-CUSUM statistics $W_{\alpha,k,n}$ in (1.4) for each local feature, it is important to fuse these local statistics together smartly so as to address the sparsity issue. Here, we propose to combine these local statistics together and raise a global-level alarm at time

$$N_\alpha(b,d) = \inf\left\{n : \sum_{k=1}^{K} \max\{0, W_{\alpha,k,n} - d\} \geq b\right\}, \tag{1.6}$$

for some pre-specified constants $b, d > 0$ whose appropriate choices will be discussed later.

Note that our proposed scheme $N_\alpha(b,d)$ in (1.6) uses the soft-thresholding transformation, $h(W) = \max\{0, W - d\}$, to filter out those non-changing local features, and keep

only those local features that might provide information about the changing event. This will allow us to improve the detection power in the sparisty scenario when only a few local features are involved in the change, also see Liu, Zhang, and Mei (2019) for more discussions.

It is useful to compare our proposed scheme $N_\alpha(b,d)$ in (1.6) with other existing methods from the spatial-temporal detection viewpoint. In the literature, many existing change-point schemes are developed by looking at the time domain first, and then searching the spatial domain over different features for possible feature changes, see Xie and Siegmund (2013) and Wang and Mei (2015). Unfortunately, such approach is often computationally expensive and cannot be implemented online for real-time monitoring due to lack of recursive forms. Here our proposed method (1.6) switches the order of spatial and temporal domains by parallel searching for local changes for each and every possible local changes, yielding computationally simple schemes that can be implemented recursively for real-time monitoring.

We should also mention that besides the soft-thresholding transformation, there are other approaches to combine the local detection statistics together to make a global alarm. Two popular approaches in the literature are the "MAX" and the "SUM" schemes, see Tartakovsky and Veeravalli (2008) and Mei, 2010:

$$N_{\alpha,\max}(b) = \inf\left\{n \geq 1 : \max_{1\leq k\leq K} W_{\alpha,k,n} \geq b\right\}, \tag{1.7}$$

$$N_{\alpha,\text{sum}}(b) = \inf\left\{n \geq 1 : \sum_{k=1}^{K} W_{\alpha,k,n} \geq b\right\}. \tag{1.8}$$

Unfortunately, the "MAX" and "SUM" approaches are generally statistically inefficient unless in extreme cases of very few or many affected local data streams.

Note that there are three tuning parameters, α, d and b in our proposed scheme $N_\alpha(b,d)$ in (1.6) and L_α-CUSUM statistic $W_{\alpha,k,n}$ in (1.4), and it is useful to discuss what are the "optimal" choices of these turning parameters. The most challenging one is the optimal

choice of α, which is related to the robustness from the gross error models in (1.1), and will be discussed in Section 1.4 through developing a new concept of false alarm breakdown point. Meanwhile, the "optimal" choice of the shrinkage parameter d mainly depends on the spatial sparsity of the change on the K local features, or the number m of affected local feature coefficents, which will be discussed in the next section when we derive the asymptotic properties of our proposed scheme $N_\alpha(b, d)$ in (1.6). Finally, for given α and d, the choice of the threshold b is straightforward, as it can be chosen to satisfy the false alarm constraint in (1.3).

1.3 Theoretical properties

In this section, we investigate the statistical properties of our proposed scheme $N_\alpha(b, d)$ in (1.6) in the modern asymptotic setting when the dimension K goes to ∞, which shed light on the suitable choice of tuning parameters when monitoring high-dimensional data streams. It is important to note that the definition of our proposed scheme $N_\alpha(b, d)$ in (1.6) does not involve the contamination ratio ϵ or the probability density distribution of outlier g, but its statistical properties will depend on ϵ or g in the gross error model in (1.1). Hence, in this section and only in this section, we assume that ϵ and g are given, as our focus is to investigate the statistical properties of our proposed schemes.

For that purpose, let us first introduce two technical assumptions on the L_α-likelihood ratio statistic $Y = ([f_{\theta_1}(X)]^\alpha - [f_{\theta_0}(X)]^\alpha)/\alpha$ when X is distributed according to h_{θ_0} or h_{θ_1} under the gross error model in (1.1). Note that when $\alpha = 0$, the variable Y should be treated as the log-likelihood ratio $\log(f_{\theta_1}(X)/f_{\theta_0}(X))$.

The first assumption on Y is related to the detection delay properties of our proposed schemes:

Assumption 1.3.1. *Given $\theta \geq \theta_1, \epsilon \geq 0$ and $\alpha \geq 0$, assume*

$$\begin{aligned} I_\theta(\epsilon, \alpha) &= \mathbf{E}_{h_\theta}\Big[\frac{[f_{\theta_1}(X)]^\alpha - [f_{\theta_0}(X)]^\alpha}{\alpha}\Big] \\ &= (1-\epsilon)\mathbf{E}_{f_\theta}\Big[\frac{[f_{\theta_1}(X)]^\alpha - [f_{\theta_0}(X)]^\alpha}{\alpha}\Big] + \epsilon \mathbf{E}_g\Big[\frac{[f_{\theta_1}(X)]^\alpha - [f_{\theta_0}(X)]^\alpha}{\alpha}\Big] \end{aligned} \tag{1.9}$$

is positive, where $\mathbf{E}_{h_\theta}, \mathbf{E}_{f_\theta}$ and $\mathbf{E}_g$ denote the expectations when the density function of X is h_θ, f_θ and g, respectively.

We should mention that this assumption is very wild for small $\epsilon, \alpha > 0$. To see this, when $\epsilon = \alpha = 0$ and $\theta = \theta_1$, $I_\theta(\epsilon, \alpha)$ in the assumption becomes the well-known Kullback-Leibler information number

$$I_{\theta=\theta_1}(\epsilon = 0, \alpha = 0) = \mathbf{E}_{f_{\theta_1}} \log(f_{\theta_1}(\mathbf{X})/f_{\theta_0}(\mathbf{X})) = I(f_{\theta_1}, f_{\theta_0}), \tag{1.10}$$

which is always positive unless $f_{\theta_0} = f_{\theta_1}$. Since all functions are continuous with respect to α and ϵ, it is reasonable to assume that $I_\theta(\epsilon, \alpha)$ are also positive for small $\epsilon, \alpha > 0$. Indeed, if f_θ belongs to a one-parameter exponential family

$$f_\theta(x) = \exp(\theta x - b(\theta)), \tag{1.11}$$

where $b(\theta)$ is strictly convex on $\mathbb{R}$, then it is straightforward to show that $I_\theta(\epsilon = 0, \alpha = 0)$ would be an increasing function of θ. This implies $I_\theta(\epsilon = 0, \alpha = 0) \geq I_{\theta=\theta_1}(\epsilon = 0, \alpha = 0) = I(f_{\theta_1}, f_{\theta_0}) > 0$ for all $\theta \geq \theta_1$. Thus, $I_\theta(\epsilon, \alpha) > 0$ for small $\epsilon, \alpha > 0$, and Assumption 1.3.1 holds.

The second assumption on Y is related to the false alarm rate of our proposed schemes, and involves some basic probability knowledge on the moment generating function (MGF). For a random variable Y with pdf $s(y)$, recall that the MGF is given by $\varphi(\lambda) = \mathbf{E}(e^{\lambda Y}) = \int e^{\lambda y} s(y) dy$ when well-defined. A nice property of MGF is that $\varphi(\lambda)$ is a convex function of λ with $\varphi(0) = 1$. An important corollary is that there often exists another non-zero

constant λ^* such that $\varphi(\lambda^*) = 1$, and $\lambda^* > 0$ if and only if $\mathbf{E}(Y) < 0$, see Lemma 3.2.1 in the Appendix. Our second assumption essentially says that this is the case under the pre-change hypothesis, and is rigorously stated as follows.

Assumption 1.3.2. *Given $\epsilon \geq 0$ and $\alpha \geq 0$, assume there exists a number $\lambda(\epsilon, \alpha) > 0$ such that*

$$\begin{aligned} 1 &= \mathbf{E}_{h_{\theta_0}} \exp\left\{\lambda(\epsilon,\alpha)\frac{[f_{\theta_1}(X)]^\alpha - [f_{\theta_0}(X)]^\alpha}{\alpha}\right\} \\ &= (1-\epsilon)\mathbf{E}_{f_{\theta_0}} \exp\left\{\lambda(\epsilon,\alpha)\frac{[f_{\theta_1}(X)]^\alpha - [f_{\theta_0}(X)]^\alpha}{\alpha}\right\} + \\ &\quad \epsilon\mathbf{E}_g \exp\left\{\lambda(\epsilon,\alpha)\frac{[f_{\theta_1}(X)]^\alpha - [f_{\theta_0}(X)]^\alpha}{\alpha}\right\}. \end{aligned} \tag{1.12}$$

We should mention that Assumption 1.3.2 is reasonable at least when ϵ and α are small. To see this, note that when $\alpha = 0$ and $\epsilon = 0$, for $Y = \log(f_{\theta_1}(X)/f_{\theta_0}(X))$, we have $\mathbf{E}_{f_{\theta_0}}(e^Y) = 1$ and thus $\lambda(\epsilon = 0, \alpha = 0) = 1$ in Assumption 1.3.2. Therefore, $\lambda(\epsilon, \alpha)$ should be in the neighborhood of 1 and thus are positive when ϵ and α are small.

With Assumptions 1.3.1 and 1.3.2, we are able to present the properties of our proposed scheme $N_\alpha(b, d)$ in (1.6) in the following subsections. Subsection 1.3.1 discusses the false alarm properties, whereas subsection 1.3.2 investigates the detection delay properties including the robustness regarding on the number of affected local data streams.

1.3.1 False alarm analysis

In this subsection, we analyze the global false alarm rate of our proposed scheme $N_\alpha(b, d)$ in (1.6) for online monitoring K independent features under the gross error model in (1.1), no matter how large K is. The classical techniques in sequential change-point detection for one-dimensional data are based on the change of measure arguments and then use renewal theory to conduct overshoot analysis under the asymptotic setting as the global threshold b goes to ∞. Unfortunately such renewal-theory-based analysis often yields poor approximations when the dimension K is moderately large, since the overshoot constant generally

increases exponentially as a function of the dimension K. Moreover, they cannot be extended to the modern asymptotic regime when the number K of local data streams goes to ∞. In other words, these classical techniques are unable to provide deep insight on the effects of the dimension K.

Here we present an alternative approach that is based on Chebyshev's inequality and can provide useful information bounds on the global false alarm rate regardless of how large the number K of features is.

Theorem 1.3.1. *Given that Assumption 1.3.2 holds for $\epsilon \geq 0$ and $\alpha \geq 0$, i.e., $\lambda(\epsilon, \alpha) > 0$. If $\lambda(\epsilon, \alpha)b > K \exp\{-\lambda(\epsilon, \alpha)d\}$, then the average run length to false alarm of our proposed scheme $N_\alpha(b, d)$ in (1.6) satisfies*

$$\mathbf{E}_\epsilon^{(\infty)}[N_\alpha(b, d)] \geq \frac{1}{4} \exp\left(\left[\sqrt{\lambda(\epsilon, \alpha)b} - \sqrt{K \exp\{-\lambda(\epsilon, \alpha)d\}} \right]^2 \right). \tag{1.13}$$

The detailed proof of Theorem 1.3.1 will be postponed in Section 1.7, and here let us add some comments to better understand the theorem. First, our rigorous, non-asymptotic result in (1.13 holds no matter how large the number K of features is. This allows us to investigate the modern asymptotic regime when the dimension K goes to ∞.

Second, the assumption of $\lambda(\epsilon, \alpha)b > K \exp\{-\lambda(\epsilon, \alpha)d\}$ essentially says that the global threshold b of our proposed scheme $N_\alpha(b, d)$ in (1.6) should be large enough if one wants to control the global false alarm rate when online monitoring large-scale streams. In particular, in order to satisfy the false alarm constraint γ in (1.3), it is natural to set the right-hand side of (1.13) to γ. This yields a conservative choice of b that satisfies $\sqrt{\lambda(\epsilon, \alpha)b} = \sqrt{K \exp\{-\lambda(\epsilon, \alpha)d\}} + \sqrt{\log(4\gamma)}$. Such a choice of b will automatically satisfy the key assumption of $\lambda(\epsilon, \alpha)b > K \exp\{-\lambda(\epsilon, \alpha)d\}$ in the theorem.

Third, when $\epsilon = \alpha = 0$, we have $\lambda(\epsilon = 0, \alpha = 0) = 1$, and our lower bound (1.13) is similar, though slightly looser, as compared to those results in equation (3.17) of Liu, Zhang, and Mei (2019), whose arguments are heuristic under a more refined assumption

on some tail distributions (see $G(x)$ defined in (2.39) below). Here we provide a rigorous mathematical statement in Theorem 1.3.1 with fewer assumptions, though the price we pay is that the corresponding lower bound is a little loose.

Finally, it turns out that our lower bound (1.13) provides the correct first-order term of the classical CUSUM procedure when online monitoring $K = 1$ data stream under the idealized model. In that case, we have $\epsilon = \alpha = d = 0$, and the classical CUSUM procedure is the special case of our procedure $N_{\alpha=0}(b, d = 0)$. Since $\lambda(\epsilon = 0, \alpha = 0) = 1$, our lower bound (1.13) shows that for any $b > 1$,

$$\liminf_{b\to\infty} \frac{\log \mathbf{E}^{(\infty)}_{\epsilon=0}[N_{\alpha=0}(b, d = 0)]}{b} \geq 1. \tag{1.14}$$

Meanwhile, as the classical CUSUM procedure, it is well-known from the classical renewal-theory-based techniques that $\lim_{b\to\infty} \frac{\log \mathbf{E}^{(\infty)}_{\epsilon=0}[N_{\alpha=0}(b,d=0)]}{b} = 1$, see Lorden (1971). Hence, our lower bound (1.13) provides the correct first-order term for $\log \mathbf{E}^{(\infty)}_{\epsilon}[N_{\alpha}(b, d)]$ under the one-dimensional case as $b \to \infty$. As a result, we feel our lower bound in (1.13) is not bad in the modern asymptotic regime when the dimension K goes to ∞.

1.3.2 Detection delay analysis

In this subsection, we provide the detection delays of our proposed scheme $N_\alpha(b, d)$ in (1.6) under the gross error model h_θ in (1.1) when m out of K features are affected by the occurring event for some given $1 \leq m \leq K$. In particular, note our proposed scheme $N_\alpha(b, d)$ in (1.6) only use the information of the pre-change parameter θ_0, the minimal magnitude of the change parameter θ_1 and tuning parameters α, b, d, we will investigate its detection delay properties when the true post-change parameter θ is not less than θ_1. The following theorem presents the detection delay properties, and the proof will be postponed in Section 1.7.

Theorem 1.3.2. *Suppose Assumption 1.3.1 of $I_\theta(\epsilon, \alpha) > 0$ in (1.9) holds , and assume m*

out of K features are affected. If $b/m + d$ goes to ∞, then the detection delay of $N_\alpha(b, d)$ satisfies

$$D_{\epsilon,\theta}(N_\alpha(b,d)) \leq (1+o(1))\frac{1}{I_\theta(\epsilon,\alpha)}\left(\frac{b}{m}+d\right), \tag{1.15}$$

where the $o(1)$ term does not depend on the dimension K, and might depend on m and α as well as the distributions h_θ.

Theorem 1.3.2 characterizes the detection delay of our proposed scheme $N_\alpha(b, d)$ in (1.6), which is constructed by using the density function of f_{θ_0} and f_{θ_1}, under the gross error model when the true post-change parameter $\theta \geq \theta_1$. As we can see, the upper bound of the detection delay depends on the value of $I_\theta(\epsilon, \alpha)$, which might have different properties depending on whether $\alpha > 0$ (Our proposed L_α-CUSUM) or $\alpha = 0$ (Classical CUSUM).

As a concrete example, assume f_θ is the pdf of the normal distribution $N(\theta, 1)$, $\theta_0 = 0, \theta_1 = 1$, we can get

$$I_\theta(\epsilon=0,\alpha) = \begin{cases} \frac{1}{\alpha\sqrt{1+\alpha}}(\frac{1}{\sqrt{2\pi}})^\alpha\left(e^{-\frac{\alpha(\theta-1)^2}{2(1+\alpha)}} - e^{-\frac{\alpha\theta^2}{2(1+\alpha)}}\right), & \text{if } \alpha > 0 \\ \theta - 1/2, & \text{if } \alpha = 0. \end{cases}$$

In this case, when $\alpha = 0$, $I_\theta(\epsilon = 0, \alpha = 0)$ is a monotonic increasing function of θ, which implies the detection delay of the scheme $N_{\alpha=0}(b, d)$ for $\theta \geq \theta_1$ is maximized when $\theta = \theta_1$ (the designed minimal magnitude of the change). However, such property may no longer hold when $\alpha > 0$. Figure 1.3 plots the curve $I_\theta(0, \alpha)$ as a function of θ for two different choices of $\alpha = 0.21$ and 0.51. Both functions $I_\theta(0, \alpha)$ are highly nonlinear: they first increase and then decrease. This implies for robust change-point detection in the present of transient outliers, it will be difficult to detect both smaller changes and very larger changes: the former is consistent with the classical result with $\alpha = 0$, and the latter is a new phenomena as the larger change might be regarded as outliers. This is the price we paid for robust detection in the present of transient outliers. This phenomena is also

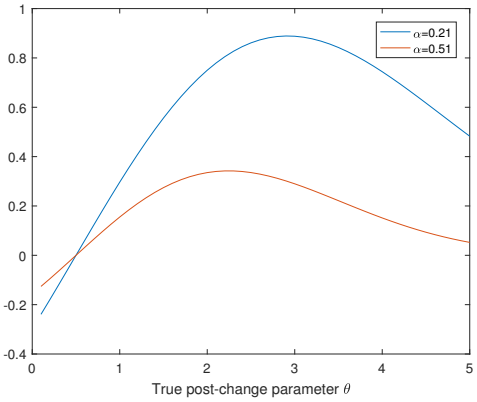

Figure 1.3: The value of $I_\theta(0, \alpha)$ with two choices of $\alpha = 0.21$ and $\alpha = 0.51$.

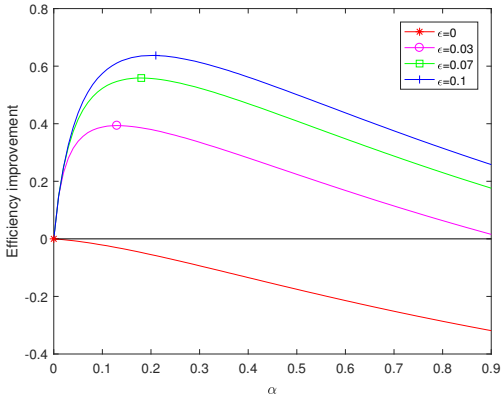

Figure 1.4: Search for the optimal α

observed when monitor the dependent data streams under the hidden Markov models (Fuh and Mei, 2015).

So far Theorems 1.3.1 and 1.3.2 investigate the statistical properties of our proposed scheme $N_\alpha(b, d)$ in (1.6) without considering the false alarm constraint γ in (1.3). Let us now investigate the detection delay properties of our proposed scheme $N_\alpha(b, d)$ in (1.6) under the gross error model in (1.1), subject to the false alarm constraint γ in (1.3). The following corollary characterizes such detection delay properties under the asymptotic regime when the false alarm constraint $\gamma = \gamma(K) \to \infty$ as the dimension $K \to \infty$ whereas the number m of affected features $m = m(K)$ may or may not go to ∞. It also includes the suitable choices of the soft-threshold parameter d and the global detection threshold b.

Corollary 1.3.1. *Under the assumptions of Theorems 1.3.1 and 1.3.2, for a given $\alpha \geq 0$ and given $d \geq 0$, a choice of global detection threshold*

$$b_\gamma = \frac{1}{\lambda(\epsilon, \alpha)} \left(\sqrt{\log(4\gamma)} + \sqrt{K \exp\{-\lambda(\epsilon, \alpha) d\}} \right)^2, \tag{1.16}$$

will guarantee that our proposed scheme $N_\alpha(b, d)$ satisfies the global false alarm constraint γ in (1.3). Moreover, in the asymptotic regime when the false alarm constraint $\gamma = \gamma(K) \to \infty$ and $m = m(K) << \min(\log \gamma, K)$ as the dimension $K \to \infty$, with

$b = b_\gamma$ in (1.16), a first-order optimal choice of the soft-thresholding parameter d that minimizes the upper bound of detection delay in (1.15) is

$$d_{opt} = \frac{1}{\lambda(\epsilon, \alpha)} \left\{ \log \frac{K}{m} + \log \frac{\log \gamma}{m} \right\}, \tag{1.17}$$

and the detection delay of the corresponding optimized scheme $N_\alpha(b_\gamma, d_{opt})$ in (1.6) satisfies

$$D_{\epsilon,\theta}(N_\alpha(b_\gamma, d_{opt})) \leq \frac{1 + o(1)}{\lambda(\epsilon, \alpha) I_\theta(\epsilon, \alpha)} \left\{ \frac{\log \gamma}{m} + \log \frac{\log \gamma}{m} + \log \frac{K}{m} \right\}. \tag{1.18}$$

Note that on the right-hand side of (1.18), the dominant order is $\max(\frac{\log \gamma}{m}, \log \frac{K}{m})$, and the second term of $\log \frac{\log \gamma}{m}$ might be negligible. However, we decide to keep it in Corollary 1.3.1, since this term will help us to compare with some classical results. As research is rather limited in the sequential change-point detection literature in the modern asymptotic regime when the number K of data streams goes to ∞. If we compare the optimal soft-thresholding parameter d_{opt} in (1.17) with the minimum detection delay in (1.18), the effects of the dimension K are the same, but the effects of the false alarm constraint γ are different. Thus, different asymptotic scenarios may arise depending on the asymptotic orders of $\log \frac{K}{m}$, $\log \frac{\log \gamma}{m}$ and $\frac{\log \gamma}{m}$, and below we consider several extreme cases.

First, let us consider the extreme case when $\log \frac{K}{m} << \log \frac{\log \gamma}{m}$, i.e., $K << \log \gamma$. This is consistent with the classical asymptotic regime when K is fixed and the false alarm constraint γ goes to ∞. In this case, for our proposed scheme, the minimum detection delay in (1.18) is of order $\frac{\log \gamma}{m}$. To be more concrete for the idealized model with $\epsilon = 0, \alpha = 0$, $\lambda(\epsilon = 0, \alpha = 0) = 1$, if the true post-change parameter $\theta = \theta_1$, then $I_{\theta=\theta_1}(\epsilon = 0, \alpha = 0) = I(f_{\theta_1}, f_{\theta_0})$, which is the Kullback-Leibler divergence. Hence based on the Corollary 1.3.1, the delay of $N_{\alpha=0}(b_\gamma, d_{opt})$ would be bounded above by $\frac{1+o(1)}{I(f_{\theta_1}, f_{\theta_0})} \frac{\log \gamma}{m}$. Meanwhile, under the idealized model, for any scheme T satisfying the false alarm constraint γ in (1.3), it is well-known that $D_{\epsilon=0}(T) \geq \frac{1+o(1)}{I(f_{\theta_1}, f_{\theta_0})} \frac{\log \gamma}{m}$ as γ goes to ∞, see Mei (2010). This suggests

that our proposed scheme with $\alpha = 0$ attains the classical asymptotic lower bound under the idealized model with $\epsilon = 0$ and the true post-change parameter $\theta = \theta_1$, in the classic asymptotic regime of $K << \log \gamma$.

Second, let us consider another extreme case when $\log \frac{K}{m} >> \frac{\log \gamma}{m}$, or equivalently, when $\log \gamma << m \log \frac{K}{m}$. This may occur when the number m of affected data streams is fixed and $\log \gamma = o(\log K)$, i.e., the false alarm constraint γ is relatively small as compared to K. In this case, both the optimal soft-thresholding parameter d_{opt} in (1.17) and the minimum detection delay in (1.18) are of order $\log \frac{K}{m}$, and the impact of the false alarm constraint γ is negligible. In other words, our proposed scheme need to take at most $O(\log K)$ observations to detect the sparse post-change scenario when only m out of K data streams are affected. This is consistent with the modern asymptotic regime results in the off-line high-dimensional sparse estimation that $O(\log K)$ observations can fully recover the K-dimensional sparse signal, see Candes and Tao (2007).

Third, the other extreme case is when both $\log \frac{K}{m}$ and $\log \frac{\log \gamma}{m}$ have the same order. This can occur if $m = K^{1-\beta}$ and $\log \gamma = K^{\zeta}$ for some $0 < \beta, \zeta < 1$, which was first investigated in Chan (2017) under the idealized model for Gaussian data. It is interesting to compare our results with those in Chan (2017). Under the idealized model with $\epsilon = 0$, the optimal choice of $\alpha = 0$, and thus our results in Corollary 1.3.1 showed that the the detection delay of our proposed scheme is of order $K^{\zeta+\beta-1} + (\xi + 2\beta - 1) \log K$, which is actually of order $\log K$ if $\frac{1-\zeta}{2} < \beta < 1 - \zeta$ but of order $K^{\zeta+\beta-1}$ if $\zeta + \beta > 1$. These two cases are exactly the assumptions in Theorems 1 and 4 of Chan (2017). While the assumption of $m << \min(\log \gamma, K)$ in Corollary 1.3.1 corresponds to $\zeta + \beta > 1$, in which our detection delay bound is identical to the optimal detection bound in Chan (2017), it is not difficult to see that the proof of Corollary 3.4.1 can be extended to the case of $\frac{1-\zeta}{2} < \beta < 1 - \zeta$, in which our results are only slightly weaker than that of Chan (2017) in the sense that the order is the same but our constant coefficient is larger. The latter is understandable because Chan (2017) used the Guassian assumptions extensively to conduct

a more careful detection delay analysis than our results in (1.15), and his results are refiner for Gaussian data under the idealized model. Meanwhile, our results are more general as they are applicable to any distributions and the gross error models. More importantly, our results give an simpler and more intuitive explanation on those assumptions in the theorems of Chan (2017), and provide a deeper insight of online monitoring large-scale data streams under general settings.

Fourth, from the detection delay point of view, Corollary 1.3.1 seems to suggest that an ideal choice of α is to maximize $\lambda(\epsilon, \alpha) I_\theta(\epsilon, \alpha)$ for each and every $\theta \geq \theta_1$, which is impossible. Here we follow the standard change-point or statistical process control (SPC) literature to tune the α value on the boundary $\theta = \theta_1$ as it is often easier to detect smaller changes than larger changes. In this case, we can define an optimal choice of α as the one that maximizes $\lambda(\epsilon, \alpha) I_{\theta_1}(\epsilon, \alpha)$. For the purpose of better illustration, we treat $\alpha = 0$ as the baseline since it corresponds to the classical CUSUM scheme that is optimal under the idealized model. Then relation (1.18) inspires us to define the asymptotic efficiency improvement of the proposed scheme $N_\alpha(b, d)$ with $\alpha \geq 0$ as compared to the baseline scheme $N_{\alpha=0}(b, d)$ as

$$e(\epsilon, \alpha) = \frac{\lambda(\epsilon, \alpha) I_{\theta_1}(\epsilon, \alpha)}{\lambda(\epsilon, \alpha = 0) I_{\theta_1}(\epsilon, \alpha = 0)} - 1 \tag{1.19}$$

Hence, the oracle optimal choice of α can be defined by maximizing the efficiency improvement $e(\epsilon, \alpha)$. That is

$$\alpha_{oracle}(\epsilon) = \arg\max_{\alpha \geq 0} [\lambda(\epsilon, \alpha) I_{\theta_1}(\epsilon, \alpha)] = \arg\max_{\alpha \geq 0} [e(\epsilon, \alpha)] \tag{1.20}$$

It is non-trivial to derive the theoretical properties of α_{oracle} as a function of ϵ, as it will depend on the relationships between $f_{\theta_0}, f_{\theta_1}$ and the contamination density g. But the good news is that the numerical values of α_{oracle} can be found fairly easy. The main tool is the Monte Carlo integration and grid search, and our key idea to simplify computational

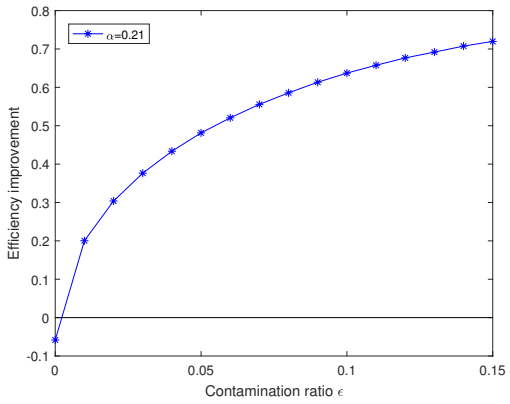

Figure 1.5: Efficiency improvement when $\alpha = 0.21$

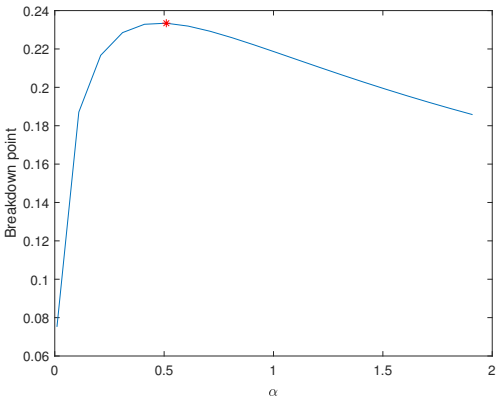

Figure 1.6: Search for the optimal α by maximizing false alarm breakdown point

complexity is to run Monte Carlo simulation *once* to compute $\lambda(\epsilon, \alpha)$ in (1.9) and $I_{\theta_1}(\epsilon, \alpha)$ in (1.12) simultaneously for many possible combinations of (ϵ, α).

As an illustration, we consider a concrete example when f_{θ_0} is the pdf of $N(0, 1)$, f_{θ_1} is the pdf of $N(1, 1)$, g is the pdf of $N(0, 3^2)$. Figure 1.4 plots $e(\epsilon, \alpha)$ as a function of the tuning parameter α for several fixed ϵ. From Figure 1.4, it is clear that when $\epsilon = 0$, the $e(\epsilon = 0, \alpha)$ curve (red curve) is linearly decreasing as a function of $\alpha \geq 0$, and thus the optimal choice of α is 0 for $\epsilon = 0$. This is consistent with the optimality properties of the CUSUM statistic under the idealized model without outliers. Meanwhile, for any other contamination rate $\epsilon > 0$, the $e(\epsilon, \alpha)$ curve is first increasing and then decreasing as α increases. Thus the optimal choice of α_{oracle} is often positive when $\epsilon > 0$. For instance, when $\epsilon = 0.1$, Figure 1.4 (blue curve) shows that $\alpha_{oracle}(\epsilon = 0.1) \approx 0.21$, and $e(\epsilon = 0.1, \alpha = 0.21) \approx 0.63$. This suggests that our proposed L_α-CUSUM based scheme with $\alpha = 0.21$ will be 63% more efficient than the baseline CUSUM based scheme under the gross error model when there are 10% outliers. Figure 1.5 shows the efficiency improvement of our proposed L_α-CUSUM based scheme with $\alpha = 0.21$ under different contamination ratio ϵ from 0 to 0.15. From the plot, we can see that as compared to the classical CUSUM based method, our proposed L_α-CUSUM based scheme with $\alpha = 0.21$

will gain $40\% \sim 70\%$ more efficiency when the contamination ratio $\epsilon \in [2\%, 15\%]$, and the price we pay is to lose 5% efficiency under the idealized model with $\epsilon = 0$.

Note the oracle optimal choice of $\alpha_{oracle}(\epsilon)$ in (1.20) requires the full information of the outliers ϵ and g, which may be unknown in practice. In the next section, we will investigate the robustness property of our proposed scheme and provide a practical way to choose α, which does not rely on any information of outliers.

1.4 Breakdown point analysis

In the classical offline robust statistics, the breakdown point is one of the most popular measures of robustness of statistical procedures. At a high-level, in the context of finite samples, the breakdown point is the smallest percentage of contaminations that may cause an estimator or statistical test to be really poor. For instance, when estimating parameters of a distribution, the breakdown point of the sample mean is 0 since a single outlier can completely change the value of the sample mean, whereas the breakdown point of the sample median is $1/2$. This suggests that the sample median is more robust than the sample mean.

Since the pioneering work of Hampel (1968) for the asymptotic definition of breakdown point, much research has been done to investigate the breakdown point for different robust estimators or hypothesis testings in the offline statistics, see Krasker and Welsch (1982) and Rousseeuw (1984). To the best of our knowledge, no research has been done on the breakdown point analysis under the online monitoring or change-point context.

Given the importance of the system-wise false alarm rate for online monitoring large-scale data streams in real-world applications, here we focus on the breakdown point analysis for false alarms. Intuitively, for a family of procedures $T(b)$ that is robust, if it is designed to satisfy the false alarm constraint γ in (1.3) under the idealized model with $\epsilon = 0$, then its false alarm rate should not be too bad under the gross error model with some small amount of outliers. There are two specific technical issues that require further clar-

ification. First, how bad is a "bad" false alarm rate? We propose to follow the sequential change-point detection literature to assess the false alarm rate by $\log \mathbf{E}_{\theta_0}^{(\infty)}(T(b))$ and deem the false alarm rate unacceptable if $\log \mathbf{E}_{\theta_0}^{(\infty)}(T(b))$ is much smaller than the designed level of $\log \gamma$, i.e., if $\log \mathbf{E}_{\theta_0}^{(\infty)}(T(b)) = o(\log \gamma)$. Second, what kind of the contamination function g in (1.22) should we consider in the gross error model? In the previous subsection we investigate the asymptotic properties of our proposed schemes when the contamination distribution g is given. However, this is unsuitable for breakdown point analysis. Here we propose to follow the offline robust statistics literature to consider the ϵ-contaminated distribution class in Huber (1964) that includes any arbitrary contamination functions g's.

To be more rigorous, in and only in this section, we define $\mathbf{E}_f^{(\infty)}$ as the expectation when the observations are i.i.d with pdf f, we propose to define the false alarm breakdown point of a family of schemes $T(b)$ as follows.

Definition 1.4.1. *Given a family of schemes $T(b)$ with $b = b_\gamma$ satisfying the false alarm constraint γ under the idealized model with $\epsilon = 0$, i.e., $\mathbf{E}_{f_{\theta_0}}^{(\infty)}(T(b)) = (1 + o(1))\gamma$, as $\gamma \to \infty$. The false alarm breakdown point $\epsilon^*(T)$ of $T(b)$'s is defined as*

$$\epsilon^*(T) = \inf\{\epsilon \geq 0 : \inf_{h_0' \in \hbar_{0,\epsilon}} \log(\mathbf{E}_{h_0'}^{(\infty)} T(b)) = o(\log \gamma)\}, \tag{1.21}$$

where the set $\hbar_{0,\epsilon}$ is the ϵ-contaminated distribution density class of the idealized model $f_{\theta_0}(x)$ for given $\epsilon \in [0, 1)$, and is defined as

$$\hbar_{0,\epsilon} = \{h | h = (1 - \epsilon) f_{\theta_0} + \epsilon g, g \in G\}, \tag{1.22}$$

and G denotes the class of all probability densities on the data $X_{k,n}$'s.

Now we are ready to conduct the false alarm breakdown point analysis for our proposed scheme $N_\alpha(b, d)$ in (1.6) with a given tuning parameter $\alpha \geq 0$. To do so, for the densities

$f_{\theta_0}(x)$ and $f_{\theta_1}(x)$, and for any given $\alpha \geq 0$, we define an intrinsic bound

$$M(\alpha) = \operatorname*{ess\,sup}_x \frac{[f_{\theta_1}(x)]^\alpha - [f_{\theta_0}(x)]^\alpha}{\alpha}, \tag{1.23}$$

and the density power divergence between f_{θ_0} and f_{θ_1}:

$$d_\alpha(\theta_0, \theta_1) = \int \left\{ [f_{\theta_1}(x)]^{1+\alpha} - (1 + \frac{1}{\alpha}) f_{\theta_0}(x)[f_{\theta_1}(x)]^\alpha + \frac{1}{\alpha}[f_{\theta_0}(x)]^{1+\alpha} \right\} dx. \tag{1.24}$$

Note that $d_\alpha(f_{\theta_0}, f_{\theta_1})$ was proposed in Basu, Harris, Hjort, and Jones (1998), which showed that it is always positive when f_{θ_1} and f_{θ_0} are different. Moreover, when $\alpha = 0$, $d_{\alpha=0}(\theta_0, \theta_1)$ becomes Kullback-Leibler information number $I(f_{\theta_0}, f_{\theta_1}) = \int f_{\theta_0}(x) \log \frac{f_{\theta_0}(x)}{f_{\theta_1}(x)} dx$.

With these two new notations, the following theorem derives the false alarm breakdown point of our proposed schemes $N_\alpha(b, d)$ as a function of the tuning parameter α for a fixed soft-thresholding parameter d when online monitoring a given K number of data streams.

Theorem 1.4.1. *Suppose that $f_\theta(x) = f(x - \theta)$ is a location family of density function with continuous probability density function $f(x)$, and assume $f_{\theta_0}(x) - f_{\theta_1}(x)$ takes both positive and negative values for $x \in (-\infty, +\infty)$. For $\alpha \geq 0$, and any fixed d and K, the false alarm breakdown point of our proposed scheme $N_\alpha(b, d)$ is given by*

$$\epsilon^*(N_\alpha) = \frac{d_\alpha(\theta_0, \theta_1)}{d_\alpha(\theta_0, \theta_1) + (1 + \alpha)M(\alpha)}, \tag{1.25}$$

where $M(\alpha)$ and $d_\alpha(\theta_0, \theta_1)$ are defined in (1.23) and (1.24). In particular, $\epsilon^*(N_\alpha) = 0$ if $M(\alpha) = \infty$ and $d_\alpha(\theta_0, \theta_1)$ is finite.

The proof of Theorem 1.4.1 will be presented in Section 1.7. Here let us apply the results for widely used normal distributions, i.e., when f_θ is the pdf of $N(\theta, \sigma^2)$. In this case, when $\alpha = 0$, the density power divergence $d_{\alpha=0}(\theta_0, \theta_1) = \frac{1}{2\sigma^2}(\theta_1 - \theta_0)^2$ is finite, but the bound $M(\alpha = 0)$ in (1.23) becomes $+\infty$ since it is the supremum of the log-likelihood

ratio $\log f_{\theta_1}(x) - \log f_{\theta_0}(x) = (\theta_1 - \theta_0)x - (\theta_1^2 - \theta_0^2)/2$ over $x \in (-\infty, \infty)$. Hence,

$$\epsilon^*(N_{\alpha=0}) = 0. \tag{1.26}$$

That is, the false alarm breakdown point of the baseline CUSUM-based scheme $N_{\alpha=0}$ is 0, i.e., any amount of outliers will deteriorate the false alarm rate of the classical CUSUM statistics-based schemes. This is consistent with the offline robust statistics literature that the likelihood-function based methods are very sensitive to model assumptions and are generally not robust.

Meanwhile, for any $\alpha > 0$, note that

$$\begin{aligned}\int_{-\infty}^{\infty} f_{\theta_0}(x)[f_{\theta_1}(x)]^{\alpha} dx &= \frac{1}{(\sqrt{2\pi}\sigma)^{1+\alpha}} \int_{-\infty}^{\infty} \exp\left(-\frac{(x-\theta_0)^2 + \alpha(x-\theta_1)^2}{2\sigma^2}\right) dx \\ &= \frac{1}{(\sqrt{2\pi}\sigma)^{\alpha}\sqrt{1+\alpha}} \exp\left(-\frac{\alpha(\theta_1-\theta_0)^2}{2(1+\alpha)\sigma^2}\right),\end{aligned}$$

and thus it is not difficult from (1.24) to show that,

$$d_{\alpha}(\theta_0, \theta_1) = \frac{\sqrt{1+\alpha}}{\alpha(\sqrt{2\pi}\sigma)^{\alpha}} \left(1 - \exp(-\frac{\alpha(\theta_1-\theta_0)^2}{2(1+\alpha)\sigma^2})\right). \tag{1.27}$$

Moreover, if we let $M(= 1/\sqrt{2\pi\sigma^2})$, then $|f_{\theta}(x)| \leq M$ for all x. By the definition in (1.23), we have $|M(\alpha)| \leq 2M^{\alpha}/\alpha$, which is finite for any $\alpha > 0$. This implies that for normal distributions, $\epsilon^*(N_{\alpha}) > 0$ for any $\alpha > 0$. Thus our proposed L_{α}-CUSUM based scheme with $\alpha > 0$ is much more robust than the classical CUSUM scheme.

Note the false alarm breakdown point of our proposed scheme does not require any information about the contamination ratio ϵ and contamination distribution g. Therefore, we proposed to choose the optimal robustness parameter α which maximizes the false alarm breakdown point in (1.25). That is

$$\alpha_{opt} = \arg\max_{\alpha \geq 0} \frac{d_{\alpha}(\theta_0, \theta_1)}{d_{\alpha}(\theta_0, \theta_1) + (1+\alpha)M(\alpha)} \tag{1.28}$$

To be more specific, let us use the same example when $f_{\theta_0} \sim N(0,1)$ and $f_{\theta_1} \sim N(1,1)$. By (1.27), we can compute the value $d_\alpha(0,1)$ for any $\alpha \geq 0$. While we do not have analytic formula for the upper bound $M(\alpha)$ in (1.23), its numerical value can be easily found by brute-force exhaustive search over the real line $x \in (-\infty, \infty)$. Figure 1.6 shows the false alarm breakdown point of our proposed scheme $N_\alpha(b,d)$ when α varies from 0 to 2. We can see clearly the breakdown point will first increase and then decrease, which yields the optimal choice of α_{opt} as 0.51, with corresponding breakdown point as 0.233. That means our proposed scheme with the choice of $\alpha = 0.51$ could tolerate 23.3% arbitrarily bad observations in terms of keeping the designed false alarm constraint stable.

It is interesting to compare the performance of the two choices of α_{oracle} in (1.20) and α_{opt} in (1.28). By the previous subsection, when $\epsilon = 0.1$ and contamination distribution is $N(0, 3^2)$, we get $\alpha_{oracle} = 0.21$ with the efficiency improvement as 63%. If we use $\alpha_{opt} = 0.51$, we will get the corresponding efficiency improvement as 55%, which makes sense because α_{opt} uses the full information of the outliers. However, from Theorem 1.4.1 and Figure 1.6, we can get the false alarm breakdown point of our proposed scheme with the choice of $\alpha_{oracle} = 0.21$ is 0.217, which implies α_{oracle} can tolerate less arbitrarily contaminations than the choice of α_{opt}. In the next section, we will also compare the performance of the two choices of α by conducting simulation studies.

1.5 Numerical simulations

In this section we conduct extensive numerical simulation studies to illustrate the robustness and efficiency of our proposed scheme $N_\alpha(b,d)$ in (1.6).

In our simulation studies, we assume that there are $K = 100$ independent features, and at some unknown time, $m = 10$ features are affected by the occurring event. Also the change is instantaneous if a feature is affected, and we do not know which subset of features will be affected. In our simulations below, we set $f_\theta =$ pdf of $N(\theta, 1)$. Then pre-change parameter $\theta_0 = 0$, the minimal magnitude of the change $\theta_1 = 1$, and the contamination

density g = pdf of $N(0, 3^2)$. Our proposed scheme $N_\alpha(b, d)$ in (1.6) is constructed by using the density function f_{θ_0} and f_{θ_1}

We conduct four different simulation studies based on the gross error model in (1.1) with different values of the contamination rate ϵ. In the first one, we consider the case when the true post-change parameter $\theta = \theta_1 = 1$, $\epsilon = 0.1$, and the objective is to illustrate that with optimized tuning parameters, our proposed robust scheme $N_\alpha(b, d)$ in (1.6) will have better detection performance than the other comparison methods in the presence of outliers. In the second one, we consider the case when $\theta = \theta_1 = 1$, $\epsilon = 0$ to demonstrate that our proposed robust scheme in the first experiment does not lose much efficiency under the idealized model. In the third simulation study, we illustrates that the false alarm rate of our proposed robust scheme indeed is more stable as compared to those CUSUM- or likelihood-ratio- based methods as the contamination rate ϵ in (1.1) varies. In the last simulation study, we investigate the sensitivity of our proposed scheme $N_\alpha(b, d)$ when the true post-change parameter θ is greater than θ_1. The detailed simulation results under these three simulation studies are presented below.

In our first simulation study, we consider the case when $\epsilon = 0.1$, e.g., 10% of data are from the outlier distribution $N(0, 3^2)$. In this case, for our proposed robust scheme $N_\alpha(b, d)$ in (1.6), as shown in previous sections, the two optimal choices of α are $\alpha_{oracle}(\epsilon = 0.1) = 0.21$ and $\alpha_{opt} = 0.51$. By (1.17), if $\log(\gamma) << K$, then the corresponding optimal shrinkage parameters $d \approx \frac{1}{\lambda(\epsilon=0.1,\alpha=0.21)} \log \frac{K}{m} = 1.6831$, $d \approx \frac{1}{\lambda(\epsilon=0.1,\alpha=0.51)} \log \frac{K}{m} = 0.9684$ for $K = 100$ and $m = 10$, since $\lambda(\epsilon = 0.1, \alpha = 0.21) = 1.3681$ and $\lambda(\epsilon = 0.1, \alpha = 0.51) = 2.3777$. For the baseline CUSUM-based scheme, i.e., $N_{\alpha=0}(b, d)$ with $\alpha = 0$, we consider two different choices of the shrinkage parameter d: one designed for $\epsilon = 0.1$ and the other designed for $\epsilon = 0$. Since $\lambda(\epsilon = 0.1, \alpha = 0) = 0.4572$ and $\lambda(\epsilon = 0, \alpha = 0) = 1$, by (1.17), we derive two optimal d values for the baseline scheme: $d \approx \frac{1}{\lambda(\epsilon=0.1,\alpha=0)} \log \frac{K}{m} = 5.0363$. and $d \approx \frac{1}{\lambda(\epsilon=0,\alpha=0)} \log \frac{K}{m} = 2.3026$.

In summary, we will compare the following eight different schemes.

- Our proposed scheme $N_\alpha(b,d)$ in (1.6) with $\alpha_{oracle}=0.21$ and $d=1.6831$ optimized for $m=10$ and $\epsilon=0.1$;

- Our proposed scheme $N_\alpha(b,d)$ in (1.6) with $\alpha_{opt}=0.51$ and $d=0.9684$ optimized for $m=10$ and $\epsilon=0.1$;

- The baseline CUSUM-based scheme $N_{\alpha=0}(b,d)$ with $d=2.306$ optimized for $m=10$ and $\epsilon=0$;

- The baseline CUSUM-based scheme $N_{\alpha=0}(b,d)$ with $d=5.0363$ optimized for $m=10$ and $\epsilon=0.1$;

- The MAX scheme $N_{\alpha=0.21,\max}(b)$ in (1.7);

- The SUM scheme $N_{\alpha=0.21,\text{sum}}(b)$ in (1.8);

- The method $N_{XS}(b,p_0=0.1)$ in Xie and Siegmund (2013) based on generalized likelihood ratio:

$$N_{XS}(b,p_0)=\inf\left\{n\geq 1:\max_{0\leq i<n}\sum_{k=1}^{K}\log\left(1-p_0+p_0\exp\left[\left(U^+_{k,n,i}\right)^2/2\right]\right)\geq b\right\},$$

where for all $1\leq k\leq K, 0\leq i<n$,

$$U^+_{k,n,i}=\max\left(0,\frac{1}{\sqrt{n-i}}\sum_{j=i+1}^{n}X_{k,j}\right).$$

- The method $N_{Chan,1}(b)$ in Chan (2017) under the idealized model that is an extension of the SUM scheme in Mei (2010):

$$N_{Chan,1}(b)=\inf\left\{n\geq 1:\sum_{k=1}^{K}\log\left(1-p_0+0.64*p_0\exp(W^*_{k,n}/2)\right)\geq b\right\},$$

where $W^*_{k,n}$ is the CUSUM statistics in (1.5).

Table 1.1: A comparison of the detection delays of 9 schemes with $\gamma = 5000$ under the gross error model. The smallest and largest standard errors of these 9 schemes are also reported under each post-change hypothesis based on 1000 repetitions in Monte Carlo simulations.

Gross error model with $\epsilon = 0.1$										
	# affected local data streams									
	1	3	5	8	10	15	20	30	50	100
Smallest standard error	0.43	0.16	0.10	0.07	0.06	0.03	0.03	0.02	0.01	0.00
Largest standard error	1.35	0.35	0.27	0.22	0.22	0.17	0.14	0.12	0.12	0.10
Our proposed robust scheme										
$N_{\alpha=0.21}(b = 16.40, d = 1.6831)$	46.2	21.1	15.1	11.4	10.1	8.2	7.2	6.0	4.9	4.0
$N_{\alpha=0.51}(b = 9.26, d = 0.9684)$	49.3	22.6	16.2	12.2	10.9	8.9	7.8	6.5	5.3	4.2
Other methods for comparison										
$N_{\alpha=0}(b = 84.74, d = 2.3026)$	94.5	41.0	27.6	19.7	17.0	12.9	10.9	8.6	6.5	4.7
$N_{\alpha=0}(b = 41.51, d = 5.0363)$	74.7	35.1	25.1	19.1	16.9	13.7	12.0	10.1	8.3	6.6
$N_{\alpha=0.21,\max}(b = 8.16)$	31.5	21.8	19.4	17.5	16.8	15.8	15.1	14.3	13.4	12.4
$N_{\alpha=0.21,\text{sum}}(b = 70.25)$	70.9	29.7	19.8	13.8	11.6	8.7	7.0	5.3	3.7	2.2
$N_{Chan,1}(b = 22.55, p_0 = 0.1)$	74.7	35.7	25.3	19.1	16.9	13.4	11.5	9.3	7.2	5.1
$N_{Chan,2}(b = 48.7, p_0 = 0.1)$	407.3	86.4	55.5	38.4	32.9	24.2	19.8	14.9	10.3	6.2
(Standard error)	(12.1)	(0.76)	(0.53)	(0.3)	(0.25)	(0.19)	(0.15)	(0.1)	(0.07)	(0.04)
$N_{XS}(b = 130, p_0 = 0.1)$	290.6	97.5	58.3	38.4	32	22.7	17.6	12.7	8.1	4.7
(Standard error)	(5.85)	(2.21)	(1.12)	(0.68)	(0.64)	(0.41)	(0.31)	(0.22)	(0.15)	(0.08)

- The method $N_{Chan,2}(b, p_0 = 0.1)$ in Chan (2017) which is similar as $N_{XS}(b, p_0)$:

$$N_{Chan,2}(b, p_0) = \inf\left\{ n : \max_{0\le i<n} \sum_{k=1}^{K} \log\left(1 - p_0 + 2(\sqrt{2}-1)p_0 \exp\left[\frac{\left(U^{+}_{k,n,i}\right)^2}{2}\right]\right) \ge b \right\}.$$

For each of these 9 schemes $T(b)$, we first find the appropriate values of the threshold b to satisfy the false alarm constraint $\gamma \approx 5000$ under the gross error model in (1.1) with $\epsilon = 0.1$ (within the range of sampling error). Next, using the obtained global threshold value b, we simulate the detection delay when the change-point occurs at time $\nu = 1$ under several different post-change scenarios, i.e., different number of affected sensors. All Monte Carlo simulations are based on 1000 repetitions.

Table 1.1 summarizes simulated detection delays of these nine schemes under 10 different post-change hypothesis, depending on different numbers of affected local data streams. Since our proposed scheme $N_{\alpha=0.21}(b, d = 1.6831)$ is optimized for the case when $m = 10$ out of data streams are affected under the gross error models, it is not surprising that it indeed has the smallest detection delays among all comparison methods when 10 data streams

are affected. In particular, our proposed schemes $N_\alpha(b, d)$ have much smaller detection delay than the three CUSUM-based schemes $N_{\alpha=0}(b, d = 5.0363)$, $N_{\alpha=0}(b, d = 2.3026)$ and $N_{Chan,1}(b, p_0 = 0.1)$. This illustrates that the improvement of L_α-CUSUM statistics with $\alpha = 0.21$ is significant as compared to the baseline CUSUM statistics in the presence of outliers.

Moreover, compared with the choice of $\alpha_{oracle} = 0.21$, our proposed scheme with $\alpha_{opt} = 0.51$ yields overall larger detection delays under those 10 different post-change hypothesis. This is consistent to the previous discussion that α_{oracle} would be better than α_{opt} when the contamination ratio ϵ and contamination distribution g are known. Note $\alpha_{opt} = 0.51$ does not use any information about ϵ and g but still led smaller detection delays than the two baseline CUSUM-based schemes $N_{\alpha=0}(b, d = 5.0363)$ and $N_{\alpha=0}(b, d = 2.3026)$, which suggests the usefulness of α_{opt}, especially when the contaminations are unknown.

In addition, the detection delays of the two likelihood-ratio-based methods $N_{XS}(b, p_0)$ and $N_{Chan,2}(b, p_0)$ are extremely large, especially when the number of affected data stream is small. The reason is that they do not suppose that $f_{\theta_1} = N(1, 1)$ is known and are designed to be efficient against $f_\theta = N(\theta, 1)$ for all $\theta > 0$. Hence they want to detect say $f_\theta = N(3, 1)$ quickly as well. Due to the presence of outliers, a significant proportion of the observations have values close to 3 and these two methods, $N_{XS}(b, p_0)$ and $N_{Chan,2}(b, p_0)$, will take this into the consideration and detect a possible change of distribution to $f_\theta = N(3, 1)$ having occurred. Since the detection delays of $N_{XS}(b, p_0 = 0.1)$ and $N_{Chan,2}(b, p_0 = 0.1)$ are very large, we use separate rows in Table 1.1 to show the standard deviation of their detection delays.

It is also interesting to note that the MAX-scheme $N_{\alpha=0.21,\max}(b)$ and the SUM-scheme $N_{\alpha=0.21,\text{sum}}(b)$ are designed for the case when $m = 1$ or $m = K$ features are affected, and Table 1.1 confirmed that their detection delays are indeed the smallest in their respective designed scenarios. However, when the number of affected features m is moderate, our proposed scheme $N_{\alpha=0.21}(b, d)$ will have smaller detection delay, which implies our pro-

posed scheme with soft-thresholding transformation could be more robust to the number of affected features.

Next, for our proposed robust scheme $N_\alpha(b, d)$ with two choices of $\alpha_{oracle}, \alpha_{opt}$, we want to investigate how much efficiency it will lose as compared to the other seven schemes under the idealized model with $\epsilon = 0$. We re-calculate the threshold b for each of these schemes $T(b)$, so as to satisfy the false alarm constraint $\gamma \approx 5000$ under the idealized model with $\epsilon = 0$.

Table 1.2 summarizes the results of our second simulation study on the detection delays of these 9 schemes under 10 different post-change hypothesis. Among all schemes, $N_{XS}(b, p_0)$ and $N_{Chan,2}(b, p_0)$ generally yield the competing smallest detection delay. However, we want to emphasize that both schemes are computationally expensive. Specifically, even if we use a time window of size k as in Chan (2017) to speed up the implementation of $N_{XS}(b, p_0)$ and $N_{Chan,2}(b, p_0)$, at each time n, $O(Kk^2)$ computations are needed to get the global monitoring statistics, whereas our proposed scheme only require $O(K)$ computations to get the global monitoring statistics. For instance, for a given global threshold b around 4.25 , it took about 130 minutes on average to finish 1000 Monte Carlo simulation runs in our laptop. If we did not know $b \approx 4.25$ and wanted to search for 10 different values of b's by bisection method based on 1000 Monte Carlo runs for each b, it would have taken about $10 * 130 = 1300$ computer minutes for the case of $\gamma = 5000$. Meanwhile, due to the nice recursive formula, our proposed schemes can be implemented in real-time. For instance, it took about 15 minutes to find such threshold b from a range of values for our proposed schemes based on 1000 Monte Carlo runs (the time is shorter if our initial guess range of b is closer) and all of these simulations are conducted on a Windows 10 Laptop with Intel i5-6200U CPU 2.30 GHz.

In addition, under the idealized model with $\epsilon = 0$, the corresponding $\alpha_{oracle} = 0$, which suggest that the baseline CUSUM scheme $N_{\alpha=0}(b, d = 2.3026)$ should have good performance when $m = 10$ data streams are affected. Moreover, in corollary 3.4.1, we

Table 1.2: A comparison of the detection delays of 9 schemes with $\gamma = 5000$ under the idealized model. The smallest and largest standard errors of these 9 schemes are also reported under each post-change hypothesis based on 1000 repetitions in Monte Carlo simulations.

Gross error model with $\epsilon = 0$										
	# affected local data streams									
	1	3	5	8	10	15	20	30	50	100
Smallest standard error	0.29	0.12	0.08	0.05	0.04	0.03	0.03	0.02	0.01	0.00
Largest standard error	0.58	0.20	0.12	0.07	0.06	0.05	0.03	0.03	0.02	0.01
Our proposed robust scheme										
$N_{\alpha=0.21}(b = 11.69, d = 1.6831)$	33.5	15.6	11.5	8.9	8.0	6.7	5.9	5.0	4.2	3.4
$N_{\alpha=0.51}(b = 7.63, d = 0.9684)$	39.4	18.1	13.3	10.2	9.2	7.6	6.6	5.7	4.7	4.0
Comparison of other methods										
$N_{\alpha=0}(b = 21.52, d = 2.3026)$	33.6	15.2	11.0	8.4	7.5	6.1	5.3	4.5	3.7	3.0
$N_{\alpha=0}(b = 7.35, d = 5.0363)$	22.4	13.8	11.1	9.3	8.6	7.6	7.0	6.3	5.5	4.8
$N_{\alpha=0.21,\max}(b = 7.14)$	24.4	17.1	15.4	14.1	13.6	12.8	12.2	11.6	10.9	10.2
$N_{\alpha=0.21,\text{sum}}(b = 58.81)$	56.0	23.2	15.5	10.8	9.1	6.8	5.6	4.2	3.0	2.0
$N_{chan,1}(b = 3.44, p_0 = 0.1)$	26.7	14.2	10.9	8.6	7.8	6.3	5.5	4.5	3.4	2.3
$N_{chan,2}(b = 4.25, p_0 = 0.1)$	26.3	13.1	9.7	7.2	6.3	4.8	3.9	2.9	2.0	1.1
$N_{XS}(b = 19.5, p_0 = 0.1)$	30.9	13.2	9.2	7.2	5.7	4.7	3.5	2.5	1.8	1.0

show the detection delay of our proposed scheme nearly achieves the optimal detection lower bound in Chan (2017), which can be validated from the numerical results in Table 1.2 since it compares well with the best possible method.

Another interesting observation from Table 1.2 is that the detection delay of our proposed robust scheme $N_{\alpha=0.21}(b, d = 1.6831)$ is comparable with that of $N_{\alpha=0}(b, d = 2.3026)$, and it just takes 6.3% more time steps to raise a correct global alarm under the idealized model when $m = 10$ data streams are affected. Recall that in Table 1.1, $N_{\alpha=0}(b, d = 2.3026)$ takes 68.3% more time steps than $N_{\alpha=0.21}(b, d = 1.6831)$ to raise a global alarm under the gross error model with $\epsilon = 0.1$. In other words, our proposed robust scheme $N_{\alpha=0.21}(b, d = 1.6831)$ sacrifices about 6.3% efficiency under the idealized model with $\epsilon = 0$, but can gain 68.3% efficiency under the gross error model with proportion of outliers $\epsilon = 0.1$.

In the third experiment, we want to investigate the impact of contamination rate ϵ on the false alarms, and illustrate the robustness of our proposed L_α-CUSUM statistics with respect to ϵ. Since the MAX-scheme $N_{\alpha=0.21,\max}(b = 7.14)$ and the SUM-scheme

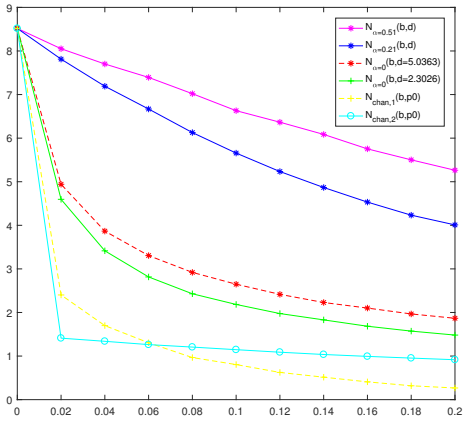

Figure 1.7: Each line represents the average run length to false alarm, $\log \mathbf{E}_{\theta_0}^{(\infty)}(T)$, of a selected scheme as a function $\epsilon \in (0, 0.2)$.

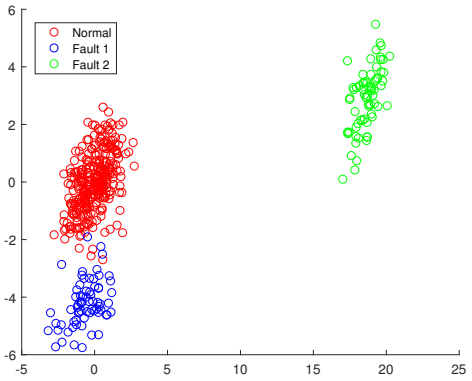

Figure 1.8: Projection of all samples on two wavelet coefficients

$N_{\alpha=0.21,\text{sum}}(b = 58.81)$ are based on local L_α-CUSUM statistics, their robustness properties to the outliers are similar to our proposed scheme $N_{\alpha=0.21}(b = 11.69, d = 1.6831)$ and $N_{\alpha=0.51}(b = 7.63, d = 0.9684)$. To highlight the robustness of our proposed L_α-CUSUM statistics, we compare our proposed schemes $N_{\alpha=0.21}(b = 11.69, d = 1.6831)$ and $N_{\alpha=0.51}(b = 7.63, d = 0.9684)$ with other four schemes: two baseline CUSUM schemes and Chan's two methods.

Figure 1.7 reports the curve of $\log \mathbf{E}_{\theta_0}^{(\infty)}(T)$ as the contamination ratio ϵ varies from 0.02 to 0.2 with stepsize 0.02. It is clear from the figure that all curves decrease with the increasing of contaminations, meaning that all schemes will raise false alarm more frequently when there are more outliers. However, the curves for the CUSUM or likelihood-ratio based methods decreased very quickly, whereas our proposed L_α-CUSUM statistics-based method with $\alpha_{oracle} = 0.21$ and $\alpha_{opt} = 0.51$ decrease rather slowly. This suggests that our proposed scheme is more robust in the sense of keeping $\log \mathbf{E}_{\theta_0}^{(\infty)}(T)$ more stable with a small departure from the assumed model. Moreover, note the curve for $\alpha_{opt} = 0.51$ decreases slower than the curve for $\alpha_{oracle} = 0.21$, which implies the performance of α_{opt} is better than α_{oracle} in term of keeping the false alarm constraint stable to the contaminations.

Table 1.3: A comparison of the detection delays of 6 schemes with $\gamma = 5000$, $m = 10$.

Gross error model with $\epsilon = 0.1$.					
True post-change θ value	$\theta = 1$	$\theta = 1.5$	$\theta = 2$	$\theta = 2.5$	$\theta = 3$
Our proposed robust scheme					
$N_{\alpha=0.21}(b = 16.40, d = 1.6831)$	10.1 ± 0.06	6.5 ± 0.03	5.2 ± 0.02	4.6 ± 0.02	4.5 ± 0.01
$N_{\alpha=0.51}(b = 9.26, d = 0.9684)$	10.9 ± 0.06	7.4 ± 0.03	6.4 ± 0.02	6.5 ± 0.02	7.4 ± 0.02
CUSUM-based scheme					
$N_{\alpha=0}(b = 84.74, d = 2.3026)$	17.0 ± 0.08	10.0 ± 0.05	7.2 ± 0.03	5.7 ± 0.02	4.8 ± 0.02
$N_{Chan,1}(b = 22.55, p_0 = 0.1)$	16.8 ± 0.10	9.9 ± 0.04	7.1 ± 0.03	5.6 ± 0.02	4.7 ± 0.02
$N_{Chan,2}(b = 48.7, p_0 = 0.1)$	32.8 ± 0.18	14.9 ± 0.08	8.5 ± 0.05	5.6 ± 0.03	3.9 ± 0.02
$N_{XS}(b = 130, p_0 = 0.1)$	32.3 ± 0.61	14.7 ± 0.26	8.4 ± 0.15	5.6 ± 0.08	3.9 ± 0.07

In the last experiment, we focus on the sensitivity of our proposed scheme $N_\alpha(b, d)$ with the misspecified post-change parameter θ. Specifically, we fix the number of affected features $m = 10$ and set the true post-change parameter θ to be $1, 1.5, 2, 2.5$, and 3. Then, we simulate the detection delay of our proposed schemes $N_{\alpha=0.21}(b = 11.69, d = 1.6831)$, $N_{\alpha=0.51}(b = 7.63, d = 0.9684)$, the CUSUM-based scheme $N_{\alpha=0}(b = 84.74, d = 2.3026)$, $N_{Chan,1}(b = 22.55, p_0 = 0.1)$, $N_{Chan,2}(b = 48.7, p_0 = 0.1)$ and $N_{XS}(b = 130, p_0 = 0.1)$. The results are summarized in Table 1.3. First, we can see although $\alpha_{oracle} = 0.21$ is designed to be optimal when the true post-change parameter $\theta = \theta_1 = 1$ with $\epsilon = 0.1$ and $g = N(0, 3^2)$, it still has the smallest detection delay among those three schemes with the true change parameter is larger than 1. Second, although the overall performance of our proposed scheme with the choice of α to be $\alpha_{opt} = 0.51$ is not as good as the the choice of α to be $\alpha_{oracle} = 0.21$, it still has a smaller detection delay than the CUSUM-based method when the true post-change post-change parameter is smaller than 2. Moreover, it does not use any knowledge of outliers ϵ and g. Those results demonstrate that generally our proposed scheme $N_\alpha(b, d)$ are not sensitive to the small misspecified post-change parameter θ.

1.6 Case study

In this section, we conduct a case study based on a real dataset of tonnage signal collected from a progressive forming manufacturing process. The dataset includes 307 normal samples and 2 different groups of fault samples . Each group contains 69 samples which are collected under the faults due to missing part occurring in the forming station (hereafter called Fault #1) and the pre-forming station (hereafter called Fault #2). Additionally, there are $p = 2^{11} = 2048$ measurement points in each tonnage signal. We want to build efficient monitoring scheme to detect the faults due to missing part occurring in the forming station while avoid making false alarm on the random fault #2 samples.

In literature, wavelet-based approaches have been widely used for analyzing and monitoring nonlinear profile data (Fan, 1996; Zhou, Sun, and Shi, 2006; Lee, Hur, Kim, and Wilson, 2012). In this chapter, Haar transform is chosen as an illustration of our proposed scheme because Haar coefficients have an explicit interpretation of the changes in the profile observations, see Zhou, Sun, and Shi (2006) as an example about applying Haar transform and the physical interpretation of the Haar coefficients. Specifically, discrete Haar transform is applied on each tonnage signal data and we just keep the first $p = 512$ Haar coefficients.

We use $c_{k,n}$ denotes the k^{th} Haar coefficient of the n^{th} tonnage signal data. Then we consider the normalized standardized Haar coefficients by

$$X_{k,n} = \frac{c_{k,n} - \hat{\mu}_k}{\hat{\sigma}_k}, \tag{1.29}$$

where $\hat{\mu}_k$ and $\hat{\sigma}_k^2$ are the sample mean and variance of all in-control normal tonnage signal data on the k^{th} Haar coefficient. Figure 1.8 shows the projection of all normal and faulty samples on two selected standardized Haar coefficients. Clearly, we may not detect the fault 1 samples if we just using the first Haar coefficient. This illustrates the necessary to monitor a large number of coefficients to effectively detect some small but persistent

changes.

After standardizing those Haar coefficients, we assume those $X_{k,n}$'s are i.i.d with standard normal distribution $N(0,1)$ for the in-control tonnage samples and have some mean shifts for those faulty tonnage samples. To apply our proposed scheme, we set $\theta_1 = 1$, i.e., the minimal magnitude of shift is 1 and the number of affected coefficients $m = 50$. We will use our proposed scheme with the choice of $\alpha = 0.51$, which maximizes the false alarm breakdown point, and the choice of $\alpha = 0.21$, which minimizes efficiency improvement for $\epsilon = 0.1$ and $g = N(0, 3^2)$. We compare the performance of those two choices of α with the baseline CUSUM-based scheme $N_{\alpha=0}(b, d)$, Xie and Siegmund method $N_{XS}(b, p_0 = 0.1)$ and Chan's two methods $N_{Chan,1}(b, p_0 = 0.1)$ and $N_{Chan,2}(b, p_0 = 0.1)$. All of those schemes are conducted by using the normalized Haar coefficients data $X_{k,n}$ in (1.29).

To evaluate the detection efficiency of those methods, we first find the appropriate values of the global threshold b such that the average run length of each scheme is 300 when the samples are collected by sampling from the 307 in-control tonnage samples with probability 90% and from the 69 Fault #1 tonnage samples with probability 10%. Then, using the obtained global threshold value b, we simulate the detection delay when the samples are sequentially collected by sampling from the 69 Fault #1 tonnage samples with probability 90% and from the Fault 2 tonnage samples with probability 10%. All Monte Carlo simulations are based on 100 repetitions. The results of detection delay and standard error are summarized in Table 1.4.

From Table 1.4, we can see our proposed schemes yield very small detection delay for detecting the smaller persistent change caused by Fault #1 compared with other methods. Thus, they are robust to the larger but transient change caused by Fault #2.

1.7 Proofs

In this section, we provide the detailed proofs for Theorem 1.3.1, Theorem 1.3.2, Theorem 1.4.1 and Corollary 1.3.1.

Table 1.4: A comparison of the detection delays of 6 methods with in-control average run length equal to 300 based on 100 repetitions in Monte Carlo simulations. The standard errors of the detection delays are reported in the bracket.

Method	Detection delay (Standard deviation)
$N_{\alpha=0.21}(b=133, d=1.5056)$	$5.96(0.08)$
$N_{\alpha=0.51}(b=80, d=0.7235)$	$6.45(0.09)$
$N_{\alpha=0}(b=4400, d=3.9357)$	$43.44(0.46)$
$N_{Chan,1}(b=2120, p_0=0.1)$	$43.2(0.42)$
$N_{Chan,2}(b=1950, p_0=0.1)$	$26.43(0.48)$
$N_{XS}(b=4050, p_0=0.1)$	$23.13(0.67)$

1.7.1 Proof of Theorem 1.3.1

For any $x \geq 0$, by Chebyshev's inequality,

$$\begin{aligned}
\mathbf{E}_\epsilon^{(\infty)}[N_\alpha(b,d)] &\geq x\mathbf{P}_\epsilon^{(\infty)}(N_\alpha(b,d) \geq x) \\
&= x\left[1 - \mathbf{P}_\epsilon^{(\infty)}(N_\alpha(b,d) < x)\right] \\
&= x\left[1 - \mathbf{P}_\epsilon^{(\infty)}(\sum_{k=1}^{K} \max\{0, W_{\alpha,k,n} - d\} \geq b) \text{ for some } 1 \leq n \leq x\right] \\
&\geq x\left[1 - x\mathbf{P}_\epsilon^{(\infty)}(\sum_{k=1}^{K} \max\{0, W^*_{\alpha,k} - d\} \geq b)\right], \qquad (1.30)
\end{aligned}$$

where $W^*_{\alpha,k} = \limsup_{n\to\infty} W_{\alpha,k,n}$. We will show that $W^*_{\alpha,k}$ exists later, and when it does exist, it is clear that $W^*_{\alpha,k}$ are i.i.d. across different k under the pre-change measure $\mathbf{P}_\epsilon^{(\infty)}$. Now if we define the log-moment generating function of the $W^*_{\alpha,k}$'s

$$\psi_\alpha(\theta) = \log \mathbf{E}_\epsilon^{(\infty)} \exp\{\theta \max(0, W^*_{\alpha,k} - d)\} \qquad (1.31)$$

for some $\theta \geq 0$, then another round application of Chebyshev's inequality yields

$$\begin{aligned}
\exp(K\psi_\alpha(\theta)) &= \mathbf{E}_\epsilon^{(\infty)} \exp\{\theta \sum_{k=1}^{K} \max(0, W^*_{\alpha,k} - d)\} \\
&\geq e^{\theta b}\mathbf{P}_\epsilon^{(\infty)}(\sum_{k=1}^{K} \max\{0, W^*_{\alpha,k} - d\} \geq b) \qquad (1.32)
\end{aligned}$$

for $\theta > 0$. Combining (1.30) and (1.32) yields that

$$\mathbf{E}_{\epsilon}^{(\infty)}[N_{\alpha}(b,d)] \;\; \geq \;\; x\left[1 - x\exp(-\theta b + K\psi_{\alpha}(\theta))\right] \tag{1.33}$$

for all $x \geq 0$. Since $x(1-xu)$ is maximized at $x = 1/(2u)$ with the maximum value $1/(4u)$. We conclude from (1.33) that

$$\mathbf{E}_{\epsilon}^{(\infty)}[N_{\alpha}(b,d)] \;\; \geq \;\; \frac{1}{4}\exp\left(\theta b - K\psi_{\alpha}(\theta)\right). \tag{1.34}$$

for any $\theta > 0$ as long as $\psi_{\alpha}(\theta)$ in (1.31) is well-defined.

The remaining proof is to utilize the assumption of $\lambda(\epsilon,\alpha) > 0$ in (1.12) in Assumption (1.3.2) to show that the upper limiting $W_{\alpha,k}^{*}$ of the proposed L_{α}-CUSUM statistics is well-defined and derive a careful analysis of $\psi_{\alpha}(\theta)$ in (1.31). When $\alpha = 0$, the L_{α}-CUSUM statistics become the classical CUSUM statistics, and the corresponding analysis is well-known, see Liu, Zhang, and Mei (2019). Here our main insight is that our proposed L_{α}-CUSUM statistics $W_{\alpha,k,n}$ for detecting a change from $h_0(x)$ to $h_1(x)$ in (1.1) can be thought of as the classical CUSUM statistic for detecting a local change from $h_0(x)$ to another new density function $h_2(x)$. Hence, under the pre-change hypothesis of $h_0(\cdot)$, the false alarm properties of our proposed L_{α}-CUSUM statistics can be derived through those of the classical CUSUM statistics.

By the assumption of $\lambda(\epsilon,\alpha) > 0$ in (1.12) in Assumption 1.3.2, if we define a new function

$$h_2(x) := \exp\left\{\lambda(\epsilon,\alpha)(\frac{(f_1(x))^{\alpha} - (f_0(x))^{\alpha}}{\alpha})\right\} h_0(x), \tag{1.35}$$

then $h_2(x)$ is a well-defined probability density function. Then in the problem of detection a local change from $h_0(x)$ to $h_2(x)$, the local CUSUM statistics for the kth local data stream

is defined recursively by

$$\begin{aligned} W'_{k,n} &= \max\{0, W'_{k,n-1} + \log \frac{h_2(X_{k,n})}{h_0(X_{k,n})}\} \\ &= \max\{0, W'_{n-1} + \lambda(\epsilon,\alpha)\frac{[f_1(X_{k,n})]^\alpha - [f_0(X_{k,n})]^\alpha}{\alpha}\}. \end{aligned}$$

Compared with our proposed L_α-CUSUM statistics $W_{\alpha,k,n}$, it is clear that $W'_{k,n} = \lambda(\epsilon,\alpha)W_{\alpha,k,n}$, and thus our proposed L_α-CUSUM statistics $W_{\alpha,k,n}$'s are equivalent to the standard CUSUM statistics $W'_{k,n}$ up to a positive constant $\lambda(\epsilon,\alpha)$. By the classical results on the CUSUM, see Appendix 2 on Page 245 of Siegmund (1985), as $n \to \infty$, $W'_{k,n}$ converges to a limit and thus $W_{\alpha,k,n}$ also converges to a limit, denoted by $W^*_{\alpha,k}$. Moreover, the tail probability of $W^*_{\alpha,k}$ satisfies

$$G(x) = \mathbf{P}_\epsilon^{(\infty)}(W^*_{\alpha,k} \geq x) = \mathbf{P}_\epsilon^{(\infty)}(\limsup_{n\to\infty} W'_{k,n} \geq \lambda(\epsilon,\alpha)x) \leq e^{-\lambda(\epsilon,\alpha)x}. \tag{1.36}$$

Now we shall use (1.36) to derive information bound of $\psi_\alpha(\theta)$ in (1.31). In order to simplify our arguments, we abuse the notation and simply denote $\lambda(\epsilon,\alpha)$ by λ in the remaining proof of the theorem. By the definition of $\psi_{\alpha,k}(\theta)$ in (1.31) and the tail probability $G(x)$ in (1.36), for $\theta > 0$,

$$\begin{aligned} \psi_\alpha(\theta) &= \log[\mathbf{P}^{(\infty)}(W^*_{\alpha,k} \leq d) - \int_d^\infty e^{\theta(x-d)} dG(x)] \qquad (1.37) \\ &= \log[1 + \theta \int_d^\infty e^{\theta(x-d)} G(x) dx] \\ &\leq \log[1 + \theta \int_d^\infty e^{\theta(x-d)} e^{-\lambda x} dx] \\ &= \log\left(1 + \frac{\theta}{\lambda - \theta} e^{-d\lambda}\right) \leq \frac{\theta}{\lambda - \theta} e^{-d\lambda}, \end{aligned}$$

where the second equation is based on the integration by parts. Clearly, relation (1.37) holds for any $0 < \theta < \lambda = \lambda(\epsilon,\alpha)$.

By (1.34) and (1.37), we have

$$\mathbf{E}_{\epsilon}^{\infty} N_{\alpha}(b,d) \;\geq\; \frac{1}{4}\exp\left(\theta b - \frac{K\theta}{\lambda-\theta}e^{-d\lambda}\right) \tag{1.38}$$

for all $0 < \theta < \lambda = \lambda(\epsilon,\alpha)$. When $\lambda b > K\exp\{-\lambda d\}$, relation (1.13) follows at once from (1.38) by letting $\theta = \sqrt{\lambda/b}\left(\sqrt{\lambda b} - \sqrt{K\exp\{-d\lambda\}}\right) \in (0,\lambda)$. This completes the proof of Theorem 1.3.1.

1.7.2 Proof of Theorem 1.3.2

To prove the detection delay bound (1.15) in Theorem 1.3.2, without loss of generality, assume the first m data streams are affected. Consider a new stopping time

$$T'(b,d) \;=\; \inf\{n \geq 1 : \sum_{k=1}^{m}(W_{\alpha,k,n} - d) \geq b\} = \inf\{n \geq 1 : \sum_{k=1}^{m} W_{\alpha,k,n} \geq b + md\}.$$

Clearly $N_{\alpha}(b,d) \leq T'(b,d)$, and thus

$$D_{\epsilon}(N_{\alpha}(b,d)) \;\leq\; D_{\epsilon}(T'(b,d)).$$

Next, by the recursive definition of $W_{\alpha,k,n}$ in (1.4), using the same approach in Theorem 2 of Lorden (1971) that connects the recursive CUSUM-type scheme to the random walks, we have

$$D_{\epsilon}(T'(b,d))) \;\leq\; \mathbf{E}_1 T''(b,d),$$

where $\mathbf{E}_1$ denotes the expectation when the change happen at time $\nu = 1$, and $T''(b,d)$ is the first passage time when the random walk with i.i.d. increment of mean $mI_{\theta}(\epsilon,\alpha)$

exceeds the bound $b + md$, and is defined as

$$T''(b,d) \;=\; \inf\{n \geq 1 : \sum_{i=1}^{n}\sum_{k=1}^{m} \frac{[f_1(X_{k,i})]^\alpha - [f_0(X_{k,i})]^\alpha}{\alpha} \geq b + md\}.$$

By standard renewal theory, as $(\frac{b}{m} + d) \to \infty$, we have

$$\mathbf{E}_1 T''(b,d) \leq \frac{1+o(1)}{mI_\theta(\epsilon,\alpha)}\,(b+md)\,.$$

Relation (1.15) then follows at once from the above relations, which completes the proof of Theorem 1.3.2.

1.7.3 Proof of Corollary 1.3.1

The choice of $b = b_\gamma$ in (1.16) follows directly from Theorem 1.3.1. To prove (1.17), we abuse the notation and use λ to denote $\lambda(\epsilon,\alpha)$ for simplification. By Theorem 1.3.2, the optimal d is the non-negative value that minimize the function

$$\ell(d) := \frac{b_\gamma}{m} + d = \frac{1}{\lambda m}(\sqrt{\log(4\gamma)} + \sqrt{Ke^{-\lambda d}})^2 + d. \tag{1.39}$$

This is an elementary optimization problem, and the optimal d can be found by taking derivative of $\ell(d)$ with respect to d, since $\ell(d)$ is a convex function of d. To see this,

$$\begin{aligned}
\ell'(d) &= -\frac{1}{m}(\sqrt{Ke^{-\lambda d}} + \frac{\sqrt{\log(4\gamma)}}{2})^2 + 1 + \frac{\log(4\gamma)}{4m} \\
\ell''(d) &= \frac{\lambda}{m}(\sqrt{Ke^{-\lambda d}} + \frac{\sqrt{\log(4\gamma)}}{2})\sqrt{Ke^{-\lambda d}} > 0.
\end{aligned}$$

Thus $\ell(d)$ is a convex function on $[0,+\infty)$, and the optimal d_{opt} value can be found by setting $\ell'(d) = 0$:

$$\sqrt{Ke^{-\lambda d}} = \sqrt{m + \frac{\log(4\gamma)}{4}} - \frac{1}{2}\sqrt{\log(4\gamma)}.$$

This gives an unique optimal value

$$
\begin{aligned}
d_{opt} &= \frac{1}{\lambda}\log\frac{K}{(\sqrt{m+\frac{1}{4}\log(4\gamma)}-\frac{1}{2}\sqrt{\log(4\gamma)})^2} \\
&= \frac{1}{\lambda}\left\{\log\frac{\left[\sqrt{m+\frac{1}{4}\log(4\gamma)}+\frac{1}{2}\sqrt{\log(4\gamma)}\right]^2}{m}+\log\frac{K}{m}\right\},
\end{aligned}
\tag{1.40}
$$

which is equivalent to those in (1.17) under the assumption that $m = m(K) << \min(\log\gamma, K)$. Plugging $d = d_{opt}$ in (1.40) back to (1.16) yields (1.18), and thus the corollary is proved.

□

1.7.4 Proof of Theorem 1.4.1

Before providing the detailed proof of Theorem 1.4.1, let us prove the following probability result that is interesting on its own.

Lemma 1.7.1. *Suppose that Y is a continuous random variable that takes both positive and negative values, and assume that its moment generating function $\varphi(\lambda) = \mathbf{E}[e^{\lambda Y}]$ is well defined over $-\infty < \lambda < \infty$. Then there exists a constant $\lambda^* > 0$ satisfying $\mathbf{E}[e^{\lambda^* Y}] = 1$ if and only if $\mathbf{E}(Y) < 0$.*

Proof: Let us first present several facts of the moment generating function $\varphi(\lambda) = \mathbf{E}[e^{\lambda Y}]$. First, $\varphi(\lambda)$ is a strict convex function of λ since $\varphi''(\lambda) = \mathbf{E}[Y^2 e^{\lambda Y}] > 0$, as Y is not identical 0. Second, under our assumption, $\varphi(\lambda) \to +\infty$ as $\lambda \to \pm\infty$. To see this, note that there exists a constant $y_0 > 0$, such that $\mathbf{P}(Y \geq y_0) > 0$. By Chebyshev's inequality, as $\lambda > 0$, $\varphi(\lambda) = \mathbf{E}[e^{\lambda Y}] \geq e^{\lambda y_0}\mathbf{P}(Y \geq y_0)$, which goes to ∞ as $\lambda \to \infty$. Similarly, we can show that $\lim_{\lambda\to-\infty}\varphi(\lambda) = +\infty$.

To show the "if" direction, assume $\mathbf{E}(Y) < 0$. Since $\varphi(0) = 1$ and $\varphi'(0) = \mathbf{E}(Y)$, there must exist a positive $\lambda_0 > 0$ such that $\varphi(\lambda_0) < 1$. However, $\varphi(\lambda) \to \infty$ as $\lambda \to \infty$. Hence, there exists a $\lambda^* \in (\lambda_0, \infty)$ such that $\varphi(\lambda^*) = 1$.

For the "only if" direction, since $\varphi(0) = \varphi(\lambda^*) = 1$, there exists a positive value $\lambda_1 \in (0, \lambda^*)$ such that $\varphi'(\lambda_1) = 0$. Since $\varphi(\lambda)$ is convex, $\varphi'(\lambda)$ must be decreasing. Thus $\mathbf{E}(Y) = \varphi'(0) < \varphi'(\lambda_1) = 0$. This completes the proof of the lemma. □

Now we are ready to prove Theorem 1.4.1. Let us begin with a high-level sketch of the proof. To find the breakdown point of our proposed scheme $T_\alpha(b, d)$, we need to investigate the asymptotic properties of $\mathbf{E}_h^{(\infty)}[N_\alpha(b, d)]$ for any $h = (1 - \epsilon)f_0 + \epsilon g$ as $b \to \infty$, where $\mathbf{E}_h^{(\infty)}$ denotes the expectation of run length when there is no change and all data come from the density function h here and the remaining of the proof. Since we assume $f_0(x) - f_1(x)$ take both positive and negative values, $Y = \frac{[f_1(X)]^\alpha - [f_0(X)]^\alpha}{\alpha}$ is a continuous random variable that takes both positive and negative values. By Lemma 1.7.1, it turns out the asymptotic properties depend on whether the following expectation is positive or negative:

$$\begin{aligned} \mu_{\epsilon,h} &= \mathbf{E}_h \frac{[f_1(X)]^\alpha - [f_0(X)]^\alpha}{\alpha} \\ &= (1 - \epsilon)\mathbf{E}_{f_0} \frac{[f_1(X)]^\alpha - [f_0(X)]^\alpha}{\alpha} + \epsilon \mathbf{E}_g \frac{[f_1(X)]^\alpha - [f_0(X)]^\alpha}{\alpha} \end{aligned} \tag{1.41}$$

As we will show below, $\log \mathbf{E}_h^{(\infty)}[N_\alpha(b, d)]$ is of order b if $\mu_{\epsilon,h} < 0$ but becomes of order $\log(b)$ if $\mu_{\epsilon,h} > 0$. Next, in order for $T_\alpha(b, d)$ to satisfy the false alarm constraint γ under the idealized model with $\epsilon = 0$, we must have $b \sim \log \gamma$ as it can be shown that $\mu_{\epsilon=0,h} < 0$ for any $\alpha \geq 0$ when f_0 and f_1 are from the same location family. Hence, the false alarm breakdown point can be found by finding the smallest ϵ value such that $\mu_{\epsilon,h} > 0$.

Next, let us show that $\mu_{\epsilon,h} < 0$ is a sufficient condition that $\log \mathbf{E}_h^{(\infty)}[N_\alpha(b, d)]$ is of order b. By Lemma 1.7.1, if $\mu_{\epsilon,h} < 0$, then there exists a positive real value $\lambda > 0$ such that

$$\mathbf{E}_h \exp\left\{\lambda\left(\frac{[f_1(X)]^\alpha - [f_0(X)]^\alpha}{\alpha}\right)\right\} = 1.$$

This is exactly Assumption 1.3.2 with $h_0 = h$, and thus the conclusions of Theorem 1.3.1

holds when h_0 is replaced by h. In particular, for fixed d and K, as b goes to ∞, we have

$$\log \mathbf{E}_h^{\infty} N_\alpha(b,d) \geq (1+o(1))\lambda b. \tag{1.42}$$

Meanwhile, if $\mu_{\epsilon,h} > 0$, we will show that $\log \mathbf{E}_h^{(\infty)}[N_\alpha(b,d)]$ is of order $\log(b)$. To see this, $\mathbf{E}_h^{(\infty)} N_\alpha(b,d)$ is the expected sample size of $N_\alpha(b,d)$ when the data are i.i.d. from h, which can also be regarded as the detection delay with the post-change distribution $h_1 = h$ when the change occurs at time $\nu = 1$. Indeed, $\mu_{\epsilon,h} > 0$ is actually Assumption 1.3.1 with $h_1 = h$, and thus the arguments on the detection delay analysis in Theorem 1.3.2 applies. Hence,

$$\log \mathbf{E}_h^{(\infty)} N_\alpha(b,d) \leq (1+o(1)) \log b. \tag{1.43}$$

Therefore, combining the above results with the definition of breakdown point in Definition 1.4.1, the breakdown point of our proposed scheme $N_\alpha(b,d)$ is

$$\epsilon^*(N_\alpha) = \inf\{\epsilon \geq 0 : \sup_{h \in \hbar_{0,\epsilon}} \mu_{\epsilon,h} > 0\}, \tag{1.44}$$

where $\mu_{\epsilon,h}$ is defined in (1.41).

The remaining proof is based on a careful analysis of $\mu_{\epsilon,h}$ in (1.41) for any arbitrary outlier density function g. For any $h(x) = (1-\epsilon)f_0(x) + \epsilon g(x) \in \hbar_{0,\epsilon}$, by (1.41), we have

$$\mu_{\epsilon,h} \;=\; -\frac{1-\epsilon}{1+\alpha} d_\alpha(f_0, f_1) + \epsilon \int \left(\frac{[f_1(x)]^\alpha - [f_0(x)]^\alpha}{\alpha}\right) g(x) dx, \tag{1.45}$$

where $d_\alpha(f_0, f_1)$ is defined in (1.24) and is the density power divergence between f_0 and f_1 proposed by Basu, Harris, Hjort, and Jones (1998). Here we use the fact that $\int [f_1(x)]^{1+\alpha} dx = \int [f_0(x)]^{1+\alpha} dx$ when $f_0(x)$ and $f_1(x)$ come from the same location family.

By the definition of $M(\alpha)$ in (1.23), it is clear from (1.45) that

$$\sup_{h\in\hbar_{0,\epsilon}} \mu_{\epsilon,h} = -\frac{1-\epsilon}{1+\alpha} d_\alpha(f_0, f_1) + \epsilon M(\alpha). \tag{1.46}$$

Therefore, by (1.44), if both $d_\alpha(f_0, f_1)$ and $M(\alpha)$ are finite, the false alarm breakdown point of N_α should be

$$\epsilon^*(N_\alpha) = \frac{d_\alpha(f_0, f_1)}{d_\alpha(f_0, f_1) + (1+\alpha)M(\alpha)}. \tag{1.47}$$

If $d_\alpha(f_0, f_1)$ is finite but $M(\alpha) = +\infty$, by (1.44) and (1.46), $\epsilon^*(N_\alpha) = 0$. If $d_\alpha(f_0, f_1) = +\infty$ but $M(\alpha)$ is finite, $\epsilon^*(N_\alpha) = 1$. If both $d_\alpha(f_0, f_1)$ and $M(\alpha)$ are $+\infty$ and $\frac{d_\alpha(f_0,f_1)}{M(\alpha)} = \rho$, by (1.44) and (1.46), we have $\epsilon^*(N_\alpha) = \frac{\rho}{\rho+(1+\alpha)}$ no matter ρ is finite or not. Therefore, for all cases, the false alarm breakdown point of N_α have the same expression in (1.47), which completes the proof of Theorem 1.4.1.

CHAPTER 2
COMMUNICATION-EFFICIENT QUICKEST DETECTION IN SENSOR NETWORKS

2.1 Introduction

Sensor networks have broad applications including health and environmental monitoring, biomedical signal processing, wireless communication, intrusion detection in computer networks, and surveillance for national security. There are many important dynamic decision problems in sensor networks, as information is accumulated (or updated) over time in the network systems. One of them is the quickest detection of a "trigger" event when sensor networks are deployed to monitor the changing environments over time and space, see Veeravalli (2001).

In this chapter, we consider a general scenario of quickest detection problems when some unknown, but not necessarily all, sensors might be affected by the "trigger event." A naive approach is to monitor each local sensor individually and to raise a global alarm as soon as any local sensor raises a local alarm. Unfortunately, this specific parallel local monitoring approach does not take advantage of global information and may lead to large detection delays if several sensors can provide information about the occurring event. Indeed, one allegation often made to the parallel local monitoring approach is that one loses much information at the global level by combining local detection procedures, not raw observations themselves, to make a global decision.

The main purpose of this chapter is to demonstrate that the problem is not on the parallel local monitoring approach itself, but on how to combine the local detection statistics suitably when the number of affected data streams is moderate. Our proposed methodolo-

[1]The materials in this chapter were published in *Sequential Analysis*, 2018.

gies are motivated by the communication efficiency in censoring sensor network, which was introduced by Rago, Willett, and Bar-Shalom (1996) and later by Appadwedula, Veeravalli, and Jones (2005) and by Tay, Tsitsiklis, and Win (2007). Figure 2.1 illustrates the general setting of a widely used configuration of censoring sensor networks, in which the data streams $X_{k,n}$'s are observed at the remote sensors (typically low-cost battery-powered devices), but the final decision is made at a central location, called the fusion center. The key feature of such a network is that while sensing (i.e., taking observations at the local sensors) is generally cheap and affordable, communication between remote sensors and fusion center is expensive in terms of both energy and limited bandwidth. Thus, to prolong the reliability and lifetime of the network system, practitioners often allow the local sensors to send summary messages $U_{k,n}$'s to the fusion center only when necessary. The question then becomes when and how to send summary messages so that the fusion center can still monitor the network system effectively.

This consideration motivates us to propose communication-efficient schemes that raise a global alarm based on the sum of those local detection statistics (e.g., local CUSUM statistics) that are "large" under either hard-, soft- or order- thresholding. We will then investigate the statistical properties of our proposed communication-efficient schemes under two asymptotic regimes: one is the classical asymptotic regime for fixed dimension K, and the other is the modern asymptotic regime when the dimension K goes to ∞. Our theoretical results illustrate the deep connections between communication efficiency and statistical efficiency.

It is worth pointing out that a well-known view in the standard off-line statistical inference literature is the necessity of shrinkage or thresholding for high-dimensional data in order to improve statistical power or efficiency (see Neyman (1937), Donoho and Johnstone (1994), Fan and Lin (1998), and Candes (2006)). In the sequential change-point detection or quickest detection literature, shrinkage or thresholding has been applied in two different directions for sparse post-change scenarios: one direction is the application

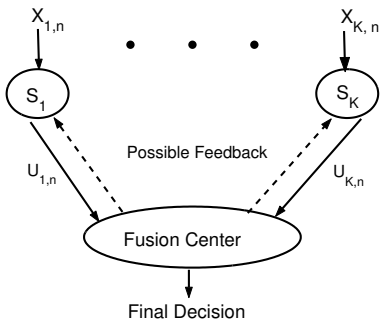

Figure 2.1: A widely used configuration of censoring sensor networks.

on the shrinkage estimation of sparse post-change parameters of local data streams, see Xie and Siegmund (2013), Wang and Mei (2015), and Chan (2017), and the other is an indirect approach of filtering out non-changing local data streams through the local summary statistics, which was first proposed in a conference paper by the author in Mei (2011) and were shown to be effective in real-world applications of profile or image monitoring (Liu, Mei, and Shi, 2015; Zhang, Mei, and Shi, 2018). This chapter investigates the asymptotic statistical properties of the indirect approach, and hopefully it will provide a deeper insight and popularize its use in practice to balance the tradeoff between communication efficiency and statistical efficiency.

The remainder of this chapter is organized as follows. In Section 2.2 we present a rigorous mathematical formulation of sequential change-point detection problems in the context of globally monitoring multiple data streams and also discuss existing methodologies. In Section 2.3, we develop our proposed methodologies from the communication-efficient viewpoint and provide guidelines how to choose tuning parameters. Asymptotic statistical properties of our proposed communication-efficient schemes are presented in Section 2.4 and numerical Monte Carlo simulation results are provided in Section 2.5. The detailed technical proofs are postponed in the Section 2.6.

2.2 Problem formulation and backgrounds

Suppose that in a network system as in Figure 2.1, there are K sensors, and each local sensor S_k observes a local data stream over time, say, $\{X_{k,n}\}_{n=1}^{\infty}$ for $k = 1, \ldots, K$. Initially, the system is "in control" and the distribution of the $X_{k,n}$'s is f_k at the k-th sensor. At some *unknown* time ν, a "trigger" event occurs to the network system, and the density function of the sensor observations $X_{k,n}$'s changes from one density f_k to another density g_k at time $\nu_k = \nu + \delta_k$. Here the term $\delta_k \in [0, \infty]$ denotes the (unknown) delay of the occurring event's impact at the k-th sensor, and $\delta_k = \infty$ implies that the k-th sensor is not affected. The problem is to find an efficient global monitoring scheme, so that the system can detect the occurring event as quickly as possible.

To be more rigorous, we assume that the f_k's and g_k's are completely specified densities with respect to a suitable measure μ_k, see, for example, Tartakovsky and Veeravalli (2004). For each $1 \leq k \leq K$, we assume that the Kullback-Leibler (KL) information number

$$I(g_k, f_k) = \int \log \frac{g_k(x)}{f_k(x)} g_k(x) d\mu_k(x) \tag{2.1}$$

is finite and positive, and

$$\int \Big(\log \frac{g_k(x)}{f_k(x)} \Big)^2 g_k(x) d\mu_k(x) < \infty. \tag{2.2}$$

Denoted by $\mathbf{P}^{(\nu)}_{\delta_1,\delta_2,\ldots,\delta_K}$ and $\mathbf{E}^{(\nu)}_{\delta_1,\delta_2,\ldots,\delta_K}$ the probability measure and expectation of the sensor observations when the event occurs at time ν, and denoted by $\mathbf{P}^{(\infty)}$ and $\mathbf{E}^{(\infty)}$ the same when there are no changes. Note that $\mathbf{P}^{(\nu)}_{\infty,\infty,\ldots,\infty}$ is the same as $\mathbf{P}^{(\infty)}$. A global monitoring scheme can be defined as a stopping time T with respect to the sequence of K-dimensional random vectors $\{(X_{1,n}, \cdots, X_{K,n})\}_{n\geq 1}$, and the interpretation of T is that, when $T = n$, we stop at time n and declare that a change has occurred somewhere at or before time n. As in the classical quickest change detection problems in Lorden (1971), our problem can

then be formulated as to find a stopping time T such that the "worse-case" detection delay

$$\overline{\mathbf{E}}_{\delta_1,\delta_2,\ldots,\delta_K}(T) = \sup_{\nu\geq 1} \operatorname{ess\,sup} \mathbf{E}^{(\nu)}_{\delta_1,\delta_2,\ldots,\delta_K}\Big((T-\nu+1)^+\Big|\mathcal{F}_{\nu-1}\Big) \tag{2.3}$$

is as small as possible for those reasonable combinations of nonnegative δ_k's subject to the global false alarm constraint

$$\mathbf{E}^{(\infty)}(T) \geq \gamma, \tag{2.4}$$

where $\gamma > 0$ is a pre-specified constant.

When $K = 1$ or when monitoring a single local data stream, say, the k-th data stream, such a problem has been well studied in the sequential change-point detection literature, see Page (1954), Shiryaev (1963), Roberts (1966), Lorden (1971), Pollak (1985), Moustakides (1986), Pollak (1987), Basseville and Nikiforov (1993), Lai (2001), and Kulldorff (2001). For a review, see the books such as Basseville and Nikiforov (1993), Poor and Hadjiliadis (2009), and Tartakovsky, Nikiforov, and Basseville (2014). One efficient local detection procedure is Page's CUSUM procedure: it raises a local alarm at the first time n when the local CUSUM statistic $W_{k,n}$ exceeds some pre-specified threshold, where $W_{k,n}$ can be computed conveniently online via a recursive formula

$$\begin{aligned} W_{k,n} &= \max\Big\{0,\ \max_{1\leq\nu\leq n}\sum_{i=\nu}^{n}\log\frac{g_k(X_{k,i})}{f_k(X_{k,i})}\Big\} \\ &= \max\Big(W_{k,n-1}+\log\frac{g_k(X_{k,n})}{f_k(X_{k,n})},\ 0\Big). \end{aligned} \tag{2.5}$$

Below we will develop global monitoring schemes based on the local CUSUM statistics $W_{k,n}$ in (2.5), although the ideas can be easily extended to other local detection statistics (in the logarithm scale of the likelihood) such as Shiryeav-Roberts statistics or scan statistics (Glaz, Naus, Wallenstein, Wallenstein, and Naus, 2001).

Now let us go back to our global monitoring problem when K is moderately large, and

it is known that the generalized likelihood ratio based methods do not have recursive forms and are computationally expensive, see Mei (2010) and Fuh and Mei (2015). In order to develop efficient scalable global monitoring schemes, it is natural to combine the local detection procedures together to make a global decision, and there are two intuitive approaches. The first one is the "MAX" scheme that raises an alarm at the global level if the maximum of the local CUSUM statistics is too large, i.e., if one of the local CUSUM procedures raises a local alarm, see Tartakovsky, Rozovskii, Blažek, and Kim (2006). Mathematically, the "MAX" scheme raises a global alarm at time

$$T_{\max}(c) = \inf\{n \geq 1 : \max_{1\leq k\leq K} W_{k,n} \geq c\}, \tag{2.6}$$

($= \infty$ if such n does not exist) where $c > 0$ is a pre-specified constant chosen to satisfy the false alarm constraint (2.4). The second approach is the "SUM" scheme, proposed in Mei (2010), in which one raises an alarm if the sum of local CUSUM statistics is too large. Specifically, at time n, each data stream calculates its local CUSUM statistic $W_{k,n}$'s as in (2.5), and then one will raise an alarm at the global level at time

$$T_{\text{sum}}(d) = \inf\{n \geq 1 : \sum_{k=1}^{K} W_{k,n} \geq d\}, \tag{2.7}$$

where the constant $d > 0$ is some suitably chosen constant. Intuitively, the "MAX" scheme $T_{\max}(c)$ in (2.6) works better when one or very few data streams are affected, whereas the "SUM" scheme $T_{\text{sum}}(d)$ in (2.7) works better when many data streams are affected, and numerical simulations in Mei (2010) indeed verified this intuition.

2.3 Communication-efficient methodology

In this section, we propose our global monitoring schemes from the communication efficiency viewpoint in the censoring sensor networks in Figure 2.1. To have a better illustration, we divide this section to two subsections. In the first subsection, we will present our

proposed schemes and provides the motivation of our proposed schemes in the censoring sensor networks. In the second subsection, we will discuss the relation between the tuning parameters in our proposed schemes and the communication costs in the censoring sensor networks and provide guidelines about how to choose the tuning parameters.

2.3.1 Our proposed schemes

From the communication efficiency viewpoint, in the censoring sensor networks in Figure 2.1, the local sensors need to summarize the information and only send "significant" information to the fusion center to prolong the reliability and lifetime of the network. This inspires us to propose to transmit only those local CUSUM statistics $W_{k,n}$'s that are larger than their respective local thresholds.

Specifically, at time n, each local sensor calculates its local CUSUM statistic $W_{k,n}$ recursively as in (2.5), and then sends the following sensor message $U_{k,n}$ to the fusion center:

$$U_{k,n} = \begin{cases} W_{k,n}, & \text{if } W_{k,n} \geq b_k \\ \text{NULL}, & \text{if } W_{k,n} < b_k \end{cases}, \tag{2.8}$$

where $b_k \geq 0$ is the local censoring (hard threshold) parameter at the k-th sensor. Here the message "NULL" is a special sensor symbol to indicate the local CUSUM statistic is not large. In practice, "NULL" could be represented by the situation when the sensor does not send any messages to the fusion center, e.g., the sensor is silent.

After receiving the local sensor messages $U_{k,n}$'s in (2.8), the fusion center then combines them together suitably to make a global decision. There are several reasonable approaches to do so, and the first two schemes are based on the summation of all sensor messages $U_{k,n}$'s, depending on how to interpret the "NULL" values. The first approach is to treat the "NULL" values as lower limit 0, and to raise a global alarm at the fusion center

at time

$$\begin{aligned} N_{hard}(a) &= \inf\Big\{n \geq 1 : \sum_{k=1}^{K} U_{k,n} \geq a\Big\} \\ &= \inf\Big\{n \geq 1 : \sum_{k=1}^{K} W_{k,n}\mathbf{1}\{W_{k,n} \geq b_k\} \geq a\Big\}. \end{aligned} \tag{2.9}$$

Below this scheme will be referred as the hard-thresholding scheme, since it involve the hard-thresholding transformation $h(w) = w\mathbf{1}\{w \geq b\}$ of the local CUSUM statistics $W_{k,n}$.

The second approach is to treat the "NULL" values as the upper limit b_k's, in which the fusion center will compute the global monitoring statistic

$$G_n = \sum_{k=1}^{K} U_{k,n} = \sum_{k=1}^{K} \max\{W_{k,n}, b_k\} = \sum_{k=1}^{K} \max\{W_{k,n} - b_k, 0\} + \sum_{k=1}^{K} b_k.$$

This is closely related to the soft-thresholding transformation $h(w) = \max(w - b, 0)$ of the local CUSUM statistic $W_{k,n}$, and we can define the soft-thresholding scheme that raises an alarm at time

$$N_{soft}(a) = \inf\Big\{n \geq 1 : \sum_{k=1}^{K} \max\{W_{k,n} - b_k, 0\} \geq a\Big\}. \tag{2.10}$$

Here we keep the threshold of $N_{soft}(a)$ as a instead of $a - \sum_{k=1}^{K} b_k$, so that both $N_{hard}(a)$ in (2.9) and $N_{soft}(a)$ in (2.10) can be written in a common SUM-shrinkage family of schemes

$$N_G(a) = \inf\{n \geq 1 : \sum_{k=1}^{K} h_k(W_{k,n}) \geq a\}, \tag{2.11}$$

also see Liu, Zhang, and Mei (2019).

The third approach occurs when the fusion center has a prior knowledge that (at most) r out of K data streams will be affected by the occurring event. Such a prior knowledge may be defined by the network fault-tolerant design to avoid risking failure. In this case, it is reasonable for the fusion center to order all sensor messages $U_{k,n}$'s as $U_{(1),n} \geq \ldots \geq$

$U_{(K),n}$, and raise an alarm if the sum of the r largest $U_{k,n}$'s is too large. This yields a global monitoring scheme that is based on the order-thresholding transformation of $U_{k,n}$'s:

$$N_{comb,r}(a) = \inf\Big\{n \geq 1 : \sum_{k=1}^{r} U_{(k),n} \geq a\Big\}, \tag{2.12}$$

where one might treats the "NULL" values as lower limit 0, upper limit b_k or any other reasonable values. In this chapter, $U_{k,n}$ in the combined scheme $N_{comb,r}(a)$ is chosen as the hard-shrinkage of the local CUSUM statistics, i.e., $W_{k,n}\mathbf{1}\{W_{k,n} \geq b_k\}$.

From the statistical viewpoint, a special case of $N_{comb,r}(a)$ in (2.12) is when the order-thresholding transformation is applied directly to the local detection statistics $W_{k,n}$'s in (2.5) themselves. Specifically, we order the K local CUSUM statistics $W_{1,n}, \ldots, W_{K,n}$ from largest to smallest: $W_{(1),n} \geq W_{(2),n} \geq \ldots \geq W_{(K),n}$. Then the order-thresholding scheme can be defined by the stopping time

$$N_{order,r}(a) = \inf\Big\{n \geq 1 : \sum_{k=1}^{r} W_{(k),n} \geq a\Big\}. \tag{2.13}$$

Clearly, $N_{order,r}(a)$ is a special case of $N_{comb,r}(a)$ if the local censoring parameter $b_k \equiv 0$, since the local CUSUM statistics $W_{k,n}$'s are non-negative.

Note that each family of schemes, $N_{hard}(a)$ in (2.9), $N_{soft}(a)$ in (2.10), $N_{order,r}(a)$ in (2.13), and $N_{comb,r}(a)$ in (2.12), can be thought of as a large family that includes both "MAX" and "SUM" schemes. For instance, the "SUM" scheme $T_{\mathrm{sum}}(d)$ in (2.7) correspond to the hard thresholding scheme $N_{hard}(a)$ with $b_k \equiv a$ and $a = d$, or the order-thresholding scheme $N_{order,r}(a)$ in (2.13) with $r = 1$. Similarly, if all threshold parameter $b_k = 0$, then the hard thresholding scheme $N_{hard}(a)$ in (2.9), the soft-thresholding schemes $N_{soft}(a)$, and $N_{comb,r}(a)$ in (2.12) with $r = K$ will become the "SUM" scheme $T_{\mathrm{sum}}(d)$ in (2.7).

It is useful to mention that our proposed schemes, $N_{hard}(a)$ in (2.9), $N_{soft}(a)$ in (2.10), $N_{order,r}(a)$ in (2.13), and $N_{comb,r}(a)$ in (2.12), take advantage of the same high-level in-

sights: little information seems to be lost at the fusion center if we do not observe those local data streams with small values of $W_{k,n}$'s since they make limited contributions to detect the true changes. These ideas and similar techniques have been applied in other contexts. Banerjee and Veeravalli (2015) essentially use the hard-thresholding transformation in (2.9) tackle the quickest detection problem when one purposely miss the observations to reduce costs. Wang, Mei, and Paynabar (2018) borrowed the soft-threshold schemes in (2.10) for profile monitoring when a change only affects some but not all principle components in the principal component analysis. Liu, Mei, and Shi (2015) applied the order-thresholding transformation in (2.13) for efficient adaptive sampling policy when one only has ability to observe r out of K data streams at each time step. This may occur in manufacturing process control when there are K possible stages in the process but there are only r expensive sensors available to monitor the process. In such a problem, the order-thresholding scheme allows us to adaptively observe those r data streams with the largest $W_{k,n}$'s values at each time step. Zhang, Mei, and Shi (2018) also used the order-thresholding transformation in (2.13) for monitoring nonlinear profiles when small shifts may occurred on some unknown regions of the profile data. In addition, along the idea of order statistics, Banerjee and Fellouris (2016) proposed the stopping time $\hat{N}_r(a) = \inf\{n : W_{(r),n} \geq a\}$. This is asymptotically equivalent to our proposed order-thresholding scheme $N_{order,r}(a)$ in (2.13) when the prior knowledge of exactly r affected data streams is true. However, our proposed order-thresholding scheme $N_{order,r}(a)$ in (2.13) is more robust when the prior knowledge is inaccurate and the true affected number of data streams $r_{true} < r$.

2.3.2 Choice of thresholding parameters

So far we simply follow our intuition without discussing how to choose the local threshold parameters b_k's. Intuitively we should choose identical local threshold parameters b_k's when the local sensors are homogeneous, but choose sensor-specified local threshold parameters b_k's when the sensors are nonhomogeneous. The homogeneous case was dis-

cussed in our previous research in Liu, Zhang, and Mei (2019), and here we focus on the possible nonhomogeneous case.

Under the assumption of the finiteness of local KL information numbers $I(g_k, f_k)$ in (2.1), we propose to choose the local threshold parameter b_k's as

$$b_k = \rho_k b \tag{2.14}$$

for $k = 1, \ldots, K$, where

$$\rho_k = \frac{I(g_k, f_k)}{\sum_{k=1}^{K} I(g_k, f_k)} \tag{2.15}$$

and $b \geq 0$ is the common global-level thresholding parameter that will be discussed in a little bit. The rigorous statistical justification of (2.14)-(2.15) will be postponed to the next section, and it is useful to think at the high-level that ρ_k can be thought of as the weight of the k-th data stream in the overall final decision, and those local sensors with larger KL information numbers or larger signal-to-noise ratios will play more important roles in the final decision. Meanwhile, note that when the sensors are the homogeneous, we have $\rho_k \equiv 1/K$ and thus local threshold parameters $b_k \equiv b/K$ are the same. Hence, our proposed choices of thresholding parameters in (2.14)-(2.15) match our intuition in the homogeneous case.

The choice of global-level thresholding parameter b is nontrivial, and may need to consider some non-statistical constraints. As an illustration, in certain applications of censoring sensor networks, the censoring parameter b may be chosen to satisfy the constraints on the average fraction of transmitting sensors when no events occur. For our proposed scheme $N_{hard}(a, b)$, when no event occurs, the average fraction of transmitting sensors at

any time step n is

$$\begin{aligned}\frac{1}{K}\sum_{k=1}^{K}\mathbf{P}^{(\infty)}(U_{k,n}\neq \text{NULL}) &= \frac{1}{K}\sum_{k=1}^{K}\mathbf{P}^{(\infty)}(W_{k,n}\geq \rho_k b)\\ &\leq \frac{1}{K}\sum_{k=1}^{K}\exp(-\rho_k b),\end{aligned}$$

where the last inequality follows from the well-known properties of the local CUSUM statistics, see, Appendix 2 on Page 245 of Siegmund (1985). In particular, if all K sensors are homogeneous in the sense that the $I(g_k, f_k)$'s are the same for all k, then $\rho_k = 1/K$, and the average fraction of transmitting sensors at any time step is $\exp(-b/K)$ when no event occurs. Hence for our proposed scheme $N_{hard}(a, b)$, a choice of

$$b = K\log\eta^{-1},$$

or equivalently, the local hard threshold $b_k = \rho_k b = b/K = \log\eta^{-1}$, will guarantee that on average, at most $100\eta\%$ of K homogeneous sensors will transmit messages at any given time when no event occurs. It is interesting to note that the local threshold $b_k = \log\eta^{-1}$ at each local sensor is a constant that does not depend on K.

The choice of b becomes more complicated for the combined thresholding schemes $N_{comb,r}(a, b)$ if the thresholding parameter r has been given beforehand. We do not have an explicit answer, and a general rule of thumb is that the censoring parameter b in (2.12) shall not be too large, as one generally should keep at least r non-zero $U_{k,n}$'s when r data streams are affected by the event.

The choice of thresholding parameter r is straightforward and depends on whether one has any prior knowledge about the maximum number of affected data streams. If such a knowledge exists and it is believed that at most r_0 data streams will be affected by the occuring event, then one should use this r_0 as the value of thresholding parameter r. Otherwise one may want to be conservative to choose $r = K$, e.g., consider the "SUM" scheme

or the hard-thresholding scheme $N_{hard}(a, b)$ in (2.9).

2.4 Statistical efficiency

In this section, we investigate the statistical efficiency of our proposed communication-efficient schemes, $N_{hard}(a)$ in (2.9), $N_{soft}(a)$ in (2.10), $N_{order,r}(a)$ in (2.13), and $N_{comb,r}(a)$ in (2.12). Here we assume that the local thresholds ρ_k are given in (2.14)-(2.15), and rewrite our proposed schemes as $N_{hard}(a, b)$, $N_{soft}(a, b)$, $N_{order,r}(a, b)$, $N_{comb,r}(a, b)$ so as to emphasize the role of the common threshold b in (2.14). Our statistical efficiency analysis allows us to provide a rationale justification of the choice of ρ_k in (2.15), or b_k in (2.14)-(2.15), although we should emphasize that these choices are a sufficient but not necessarily necessary condition in order for our proposed schemes in (2.9)-(2.13) to enjoy good properties.

For easy understanding our theoretical results, we divide this section into three subsections. In the first subsection, we provide the asymptotic upper bound of detection delay of our proposed schemes under the settings when the number of affected data streams are fixed. In the second subsection, we derive the upper bound of detection delay of our proposed scheme when the false alarm constraint (2.4) γ goes to ∞ under the classical asymptotic regime when the number of data streams K is fixed. The delay analysis on the high-dimension regime when K goes to ∞ will be presented in the last subsection.

2.4.1 Detection delay analysis

In this subsection, we consider a general setting when the change is not necessarily instantaneous. We assume that when the occurring event occurs at time ν, the k-th data stream is affected at time $\nu_k = \nu + \delta_k$, where the term $\delta_k \in [0, \infty]$ denotes the delay of the occurring event's impact on the k-th data stream. In particular, $\delta_k = \infty$ implies that the k-th data stream is not affected. In other words, the density function of the sensor observations $X_{k,n}$'s of the k-th data stream changes from f_k to g_k at time $\nu_k = \nu + \delta_k$. Most research in

the literature assumes that the delay effect δ_k only takes two possible values, 0 or ∞. Here we relax such an assumption a little bit, and assume that the delay effects δ_k's satisfy the following post-change hypothesis set Δ :

$$\Delta = \big\{(\delta_1, \ldots, \delta_K) : \text{the } \delta_k\text{'s either} = \infty \text{ or satisfy } 0 \leq \delta_k << \log\gamma \text{ and } \min_{1\leq k\leq K} \delta_k = 0\big\}. \tag{2.16}$$

where γ is the false alarm constraint in (2.4), and $x(t) << y(t)$ implies that $x(t)/y(t) \to 0$ as $t \to \infty$. Note that the assumption of $\min_{1\leq k\leq K} \delta_k = 0$ is trivial, since otherwise the system is actually affected by the occurring event at the "new" change-point $\nu' = \nu + \min_{1\leq k\leq K} \delta_k$. The assumption of $\delta_k << \log\gamma$ is a technical assumption to ensure that one is able to utilize all affected data streams to raise a global alarm subject to the false alarm constraint γ in (2.4). In other words, we only consider the scenario when the differences on the finite delay effects δ_k's are not too large as compared to the typical order ($\log\gamma$) of detection delays. A sufficient condition to satisfy this assumption is when all finite δ_k's are uniformly bounded by some constants that do not depend on the false alarm constraint γ in (2.4).

In the detection delay analysis, the following constant plays a crucial role:

$$J(\delta_1, \ldots, \delta_K) = \sum_{k=1}^{K} I(g_k, f_k) I\{\delta_k < \infty\}, \tag{2.17}$$

and $I(g_k, f_k)$ is the KL information number defined in (2.1), and $I\{A\}$ is the indicator function of set A. Essentially, the constant $J(\delta_1, \ldots, \delta_K)$ in (2.17) states that only those affected data streams can make contributions in quickest detection.

The following theorem establishes the detection delay properties of our proposed schemes, $N_{hard}(a, b)$ in (2.9), $N_{soft}(a, b)$ in (2.10), $N_{order,r}(a, b)$ in (2.13), and $N_{comb,r}(a, b)$ in (2.12), as the global threshold a goes to ∞. The proof of this theorem is presented in detail in the appendix.

Theorem 2.4.1. *Suppose* $a \to \infty$.

(i) *For any combination* $(\delta_1, \ldots, \delta_K) \in \Delta$ *defined in (2.16), as* $b \to \infty$

$$\begin{aligned} \overline{\mathbf{E}}_{\delta_1,\ldots,\delta_K}(N_{hard}(a,b)) \;\; \leq \;\; & \max\left\{ \frac{a}{J(\delta_1,\ldots,\delta_K)}, \frac{b}{\sum_{k=1}^K I(g_k,f_k)} \right\} \\ & + O(\sqrt{b}) + O\Big(\max_{\delta_k:\delta_k<\infty} (\delta_k) \Big), \end{aligned} \tag{2.18}$$

where $J(\delta_1, \ldots, \delta_K)$ *is defined in (2.17).*

(ii) *For all* $b \geq 0$, *the soft-thresholding scheme* $N_{soft}(a,b)$ *in (2.10) satisfies*

$$\begin{aligned} \overline{\mathbf{E}}_{\delta_1,\ldots,\delta_K}(N_{soft}(a,b)) \;\; \leq \;\; & \frac{a}{J(\delta_1,\ldots,\delta_K)} + \frac{b}{\sum_{k=1}^K I(g_k,f_k)} \\ & + O(\sqrt{b}) + O\Big(\max_{\delta_k:\delta_k<\infty} (\delta_k) \Big), \end{aligned} \tag{2.19}$$

(iii) *For any integer* $1 \leq r \leq K$, *the order-*r *thresholding scheme* $N_{order,r}(a)$ *in (2.13) and the combined thresholding scheme* $N_{comb,r}(a,b)$ *in (2.12) satisfy (2.18) whenever* $\sum_{k=1}^K I\{\delta_k < \infty\} \leq r$, *i.e., when the occurring event affects at most* r *sensors.*

2.4.2 Classical asymptotic regime with fixed dimension K

In this subsection, we present the asymptotic optimality properties of our proposed schemes, $N_{hard}(a,b)$, $N_{soft}(a,b)$, $N_{order,r}(a)$, and $N_{comb,r}(a,b)$, under the classical asymptotic regime in which the number of data streams K is fix and the false alarm constraint γ goes to ∞.

The following lemma derives the information bound on the detection delays of any globally monitoring schemes when Δ is defined in (2.16), as the false alarm constraint γ in (2.4) goes to ∞.

Lemma 2.4.1. *Assume a scheme* $T(\gamma)$ *satisfies the false alarm constraint (2.4). Then for*

any given post-change hypothesis $(\delta_1, \ldots, \delta_K) \in \Delta$, *as* γ *goes to* ∞,

$$\overline{\mathbf{E}}_{\delta_1,\ldots,\delta_K}(T(\gamma)) \geq (1 + o(1)) \frac{\log \gamma}{J(\delta_1, \ldots, \delta_K)}, \tag{2.20}$$

where $J(\delta_1, \ldots, \delta_K)$ *is defined in (2.17).*

When the local censoring parameters b_k's are defined in (2.14)-(2.15) with the common parameter b, the asymptotic optimality properties of our proposed schemes under the classical asymptotic regime can be summarized as follow.

Theorem 2.4.2. *For a given* K *and for any* $b \geq 0$, *with the choice of*

$$a = a_\gamma = \log \gamma + (K - 1 + o(1)) \log \log \gamma, \tag{2.21}$$

the hard-thresholding scheme $N_{hard}(a_\gamma, b)$ *satisfies the false alarm constraint (2.4). Moreover, if* $a - b$ *goes to* ∞ *as* γ *goes to* ∞, *then for all* $b \geq 0$,

$$\overline{\mathbf{E}}_{\delta_1,\ldots,\delta_K}(N_{hard}(a, b)) \leq \frac{\log \gamma + (K - 1 + o(1)) \log \log \gamma}{J(\delta_1, \ldots, \delta_K)} + O(\sqrt{b}) + O(1) \tag{2.22}$$

for all possible post-change hypothesis $(\delta_1, \ldots, \delta_K) \in \Delta$ *in (2.16). Therefore, for any given* $b = o((\log \log \gamma)^2)$, *the hard-thresholding schemes* $N_{hard}(a, b)$ *in (2.9) asymptotically minimize* $\overline{\mathbf{E}}_{\delta_1,\ldots,\delta_K}(N_{hard}(a, b))$ *(up to the second-order) for each and every post-change hypothesis* $(\delta_1, \ldots, \delta_K) \in \Delta$ *subject to the false alarm constraint (2.4), as* γ *in (2.4) goes to* ∞. *The conclusion also holds if* $N_{hard}(a, b)$ *is replaced by the soft-thresholding scheme* $N_{soft}(a, b)$ *in (2.10), the order-thresholding scheme* $N_{order,r}$ *in (2.13) or the combined thresholding scheme* $N_{comb,r}(a, b)$ *in (2.12) when the occurring event affects at most* r *data streams, i.e., when* $(\delta_1, \ldots, \delta_K) \in \Delta$ *satisfies* $\sum_{k=1}^{K} I\{\delta_k < \infty\} \leq r$.

Theorem 2.4.2 validated our choices of the local censoring parameters b_k's in (2.14) and the weights ρ_k's in (2.15) in the general nonhomogeneous scenario, as the corresponding

schemes are asymptotically optimal when the KL information numbers $I(g_k, f_k)$ in (2.1) might be different for different k. Moreover, by Theorem 2.4.2, when $b = o((\log\log\gamma)^2)$, the upper bound of the detection delay in the right hand side of (2.22) is asymptotically first-order equivalent to those with $b = 0$. This indicates that we can choose the local threshold $b = o((\log\log\gamma)^2)$ to achieve both communication efficiency and statistical efficiency simultaneously.

2.4.3 Modern asymptotic regime when the dimension $K \to \infty$

In this subsection, we present the asymptotic properties of our proposed schemes, $N_{hard}(a, b)$, $N_{soft}(a, b)$, $N_{order,r}(a)$, and $N_{comb,r}(a, b)$, under the modern asymptotic regime in which both the dimension K and the false alarm constraint γ in (2.4) go to ∞ in a suitable rate. In order to be tractable, we consider the homogenous case when $(f_k, g_k) = (f, g)$ for all k, and the local censoring parameters b_k's defined in (2.14)-(2.15) will become $b_k = b/K$ with the common parameter b. In this subsection, denote by $I = I(g, f)$ the KL information number defined in (2.1).

Here we consider the sparse post-change scenario when the number of affected data streams m is fixed, and focus on the impact of the dimension K on the performance of our proposed schemes. Two different scenarios will be investigated: $K = o(\log\gamma)$ and $K >> \log\gamma$. When K and $\log\gamma$ have the same order, research becomes more challenging and is out of the scope of this chapter. Note that Chan (2017) considers the not-so-sparse and not-so-dense post-change scenario when the number of affected data streams m goes to ∞ by assuming that $\log(m), \log(K)$, and $\log\log\gamma$ have the same order. Here our asymptotic setting is different, and we consider the case of fixed m when K and $\log\gamma$ go to ∞.

First, when both the dimension K and the false alarm constraint γ in (2.4) go to ∞, the choice of a in (2.21) for fixed K might no longer work, and thus it is crucial to find the threshold a to satisfy the false alarm constraint γ in (2.4) in the modern asymptotic setting when $K \to \infty$. The following theorem characterizes a general non-asymptotic result on

the conservative choice of the threshold a.

Theorem 2.4.3. *For any given b and K, a choice of*

$$a = (\sqrt{\log(4\gamma) + K - Ke^{-b/K}} + \sqrt{K})^2 \tag{2.23}$$

will guarantee the hard-shrinkage scheme $N_{hard}(a,b)$, the soft-thresholding scheme $N_{soft}(a,b)$, the order-thresholding scheme $N_{order,r}(a,b)$ or the combined thresholding scheme $N_{comb,r}(a,b)$ satisfy the false alarm constraint (2.4).

It is clear from Theorem 2.4.3 that the asymptotic property of the conservative threshold a in (2.23) depends on the relation between K and $\log\gamma$. The following corollary summarizes the asymptotic detection delays of our proposed schemes, and it shows that the classical asymptotic detection delay bounds for fixed K still hold when $K = o(\log\gamma)$, but we will have new asymptotic delay bounds when $K >> \log\gamma$.

Corollary 2.4.1. *Assume the number m of affected data streams is fixed, and assume K and $\log\gamma$ go to ∞,*

(i) *if $K = o(\log\gamma)$, for any $b \geq 0$, with the choice of*

$$a = a_\gamma = \log(4\gamma) + o(\log\gamma) \tag{2.24}$$

the hard-thresholding scheme $N_{hard}(a,b)$ in (2.9) satisfies the false alarm constraint in (2.4) and has the detection delay

$$\overline{\mathbf{E}}_{\delta_1,\ldots,\delta_K}(N_{hard}(a,b)) \leq (1+o(1))\frac{\log\gamma}{mI} + O(1), \tag{2.25}$$

for all possible post-change hypothesis $(\delta_1,\ldots,\delta_K) \in \Delta$ in (2.16).

(ii) *If* $K >> \log\gamma$ *and* $b \geq 0$, *with the choice of*

$$a = (1 + o(1))K \tag{2.26}$$

the hard-thresholding scheme $N_{hard}(a, b)$ *in (2.9) satisfies the false alarm constraint in (2.4). Moreover, if the local censoring parameters* b_k*'s are not too large, i.e.,* $b_k = o(K)$, *or equivalently, the global censoring parameter* $b = o(K^2)$, *we have*

$$\overline{\mathbf{E}}_{\delta_1,\ldots,\delta_K}(N_{hard}(a, b)) \leq (1 + o(1))\frac{K}{mI} + O(1), \tag{2.27}$$

for all possible post-change hypothesis $(\delta_1, \ldots, \delta_K) \in \Delta$ *in (2.16).*

(iii) *The conclusions of (i) and (ii) also hold if* $N_{hard}(a, b)$ *is replaced by the soft-thresholding scheme* $N_{soft}(a, b)$ *in (2.10), the order-thresholding scheme* $N_{order,r}$ *in (2.13) or the combined thresholding scheme* $N_{comb,r}(a, b)$ *in (2.12) when the occurring event affects at most* r *data streams, i.e., when* $(\delta_1, \ldots, \delta_K) \in \Delta$ *satisfies* $\sum_{k=1}^{K} I\{\delta_k < \infty\} \leq r$.

2.5 Numerical simulations

In this subsection we report our numerical simulation results to illustrate the usefulness of the proposed schemes in (2.9)-(2.13). Suppose that there are $K = 100$ independent and identical sensors in a system, and the observations at each sensor are iid with mean 0 and variance 1 before the change and with mean 1 and variance 1 after the change if affected. In our simulation study, we simply assume that the change is instantaneous if a sensor is affected, but we do not know which subset of sensors will be affected.

For the purpose of comparison, we conduct numerical simulations for six families of global monitoring schemes:

- the "MAX" scheme $T_{\max}(a)$ in (2.6),

- the "SUM" scheme $T_{\rm sum}(a)$ in (2.7),
- the order thresholding scheme $N_{order,r}(a)$ in (2.13) with $r = 10$,
- the hard thresholding scheme $N_{hard}(a)$ in (2.9),
- the soft thresholding scheme $N_{soft}(a)$ in (2.10),
- the combined thresholding schemes $N_{comb,r}(a)$ in (2.12) with $r = 10$.

The first three schemes require all local sensors to send all local CUSUM statistics $W_{k,n}$'s values to the fusion center at each and every time step, and corresponds to the case when the local censoring parameter $b_k \equiv 0$ for all $k = 1, \cdots, K$. For order-thresholding in the families of $N_{order,r}(a)$ and $N_{comb,r}(a)$, we choose $r = 10$ to better understand the scenario when 10 out of 100 sensors are affected by the occurring event. For each of the last three schemes in the list, i.e., our three proposed schemes (2.9)-(2.12), we further consider three different values of the local censoring parameters b_k's:

(i) $b_k \equiv 1/2 \approx -\log(0.607)$ for all k,

(ii) $b_k \equiv -\log(0.1) = 2.3026$ for all k,

(iii) $b_k \equiv -\log(0.01) = 4.6052$ for all k.

The choices of these values will guarantee that when no event occurs, on average at most $\eta = 60.7\%, 10\%$, and 1% of $K = 100$ homogeneous sensors will transmit messages at any given time, respectively. Therefore, there are a total of $3 + 3 * 3 = 12$ specific schemes in our numerical simulation study.

For each of these 12 specific schemes $T(a)$, we first find the appropriate values of the global threshold a to satisfy the false alarm constraint $\mathbf{E}^{(\infty)}(T(a)) \approx \gamma = 5000$ (within the range of sampling error). Next, using the obtained global threshold value a, we simulate the detection delay when the change-point occurs at time $\nu = 1$ under several different post-change scenarios, i.e., different number of affected sensors. All Monte Carlo simulations are based on $m = 2500$ repetitions.

Table 2.1: A comparison of the detection delays of six families of schemes with $\gamma = 5000$. The smallest and largest standard errors of these 12 schemes are also reported under each post-change hypothesis based on 2500 repetitions in Monte Carlo simulations.

	# sensors affected								
	1	3	5	8	10	20	30	50	100
Smallest standard error	0.18	0.07	0.05	0.03	0.03	0.02	0.01	0.01	0.00
Largest standard error	0.35	0.12	0.07	0.06	0.05	0.04	0.03	0.03	0.03
Schemes with $b_k \equiv 0$									
$T_{\max}(a = 11.27)$	23.3	16.3	14.4	13.0	12.4	10.9	10.2	9.5	8.7
$T_{\text{sum}}(a = 88.66)$	52.1	21.8	14.7	10.3	8.7	5.2	3.9	2.9	2.0
$N_{order,r=10}(a = 44.11)$	34.1	15.5	11.2	8.5	7.5	5.5	4.8	4.1	3.4
Schemes $N_{hard}(a)$ in (2.9) with different positive b_k's									
$N_{hard}(a = 85.60, b_k = 0.50)$	52.9	21.9	14.9	10.3	8.7	5.2	4.0	2.9	2.0
$N_{hard}(a = 52.21, b_k = 2.3026)$	50.6	20.7	13.8	9.6	8.2	5.2	4.2	3.2	2.4
$N_{hard}(a = 26.31, b_k = 4.6052)$	39.8	16.0	11.5	8.8	7.9	5.9	5.2	4.4	3.8
Schemes $N_{soft}(a)$ in (2.10) with different positive b_k's									
$N_{soft}(a = 63.92, b_k = 0.50)$	48.2	20.2	13.7	9.7	8.2	5.1	4.0	3.0	2.0
$N_{soft}(a = 21.56, b_k = 2.3026)$	33.9	15.4	11.2	8.5	7.5	5.3	4.5	3.7	3.0
$N_{soft}(a = 8.29, b_k = 4.6052)$	25.2	13.8	11.1	9.2	8.4	6.7	5.9	5.2	4.4
Schemes $N_{comb,r}(a)$ in (2.12) with $r = 10$ and different positive b_k's									
$N_{comb,r}(a = 44.11, b_k = 0.50)$	34.1	15.5	11.2	8.5	7.5	5.5	4.8	4.1	3.4
$N_{comb,r}(a = 43.88, b_k = 2.3026)$	38.5	16.8	11.7	8.6	7.5	5.5	4.7	4.0	3.3
$N_{comb,r}(a = 26.31, b_k = 4.6052)$	39.8	16.0	11.5	8.8	7.9	5.9	5.2	4.4	3.8

Table 2.1 summarizes our simulated detection delays of these 12 schemes under 8 different post-change hypothesis, depending on the number of affected sensors. From Table 2.1, among these 12 specific schemes, when a small number ($1 \sim 3$) of 100 homogeneous sensors are affected by the event, the "MAX" scheme $T_{\max}(a)$ is the best (in the sense of smallest detection delay), the "SUM" scheme $T_{\text{sum}}(a)$ is the worst, and all other schemes are in-between. Similarly, when a large number (20 or more) of 100 homogeneous sensors are affected, the order is reserved: $T_{\text{sum}}(a)$ is the best, $T_{\max}(a)$ is the worst, and all other schemes are in-between. However, when $5 \sim 10$ sensors are affected, the schemes with order-thresholding $r = 10$ yield the smallest detection delays, since they are designed to detect the scenario when 10 sensors are affected by the event. In addition, it is clear from Table 2.1 that for each given scheme, the fewer affected sensors we have, the larger detection delay it will have. All these results are consistent with our intuition.

It is worth emphasizing that for the families of the hard- and soft- thresholding schemes, $N_{hard}(a)$ in (2.9) and $N_{soft}(a)$ in (2.10), a larger censoring value of b_k actually leads to a smaller detection delay when only a few sensors are affected. This suggests that a larger censoring value b_k may actually be necessary for efficient detection when the affected sensors are sparse.

A surprising and possibly counter-intuitive result in Table 2.1 is the effect of not so large values of censoring parameters b_k's in finite sample simulations. For instance, the performances of the "SUM" scheme $T_{\rm sum}(a)$ and the hard thresholding scheme $N_{hard}(a, b_k = 0.50)$ are similar in view of sampling errors. Likewise, the top-r thresholding scheme $N_{order,r=10}(a)$ and the combined thresholding scheme $N_{comb,r=10}(a, b_k = 0.50)$ also have identical performances. The interpretation in the censoring sensor networks context is as follows: using our proposed communication policy in (2.8), we only need $\exp(-b_k) = \exp(-0.5) = 60.7\%$ of 100 sensors to transmit information to the fusion center at any given time when no event occurs, but we can still be as effective as the full transmission scenario when all sensors transmit information at all time steps. In other words, much communication costs can be saved by our proposed schemes $N_{hard}(a)$ or $N_{comb,r}(a)$ with not so large values of b_k's.

It is also interesting to see the effect of the order-thresholding parameter r in finite sample simulations when the hard-thresholding parameters b_k's are large. From Table 2.1, when the false alarm constraint γ in (2.4) is only moderately large, e.g., $\gamma = 5000$, the performances of $N_{hard}(a, b_k)$ and $N_{comb,r=10}(a, b_k)$ are identical when $b_k = 4.6052$ — they not only have the same global threshold a, but also have the same detection delays. Intuitively, the stopping time $N_{comb,r}(a, b_k)$ is decreasing as a function of r, and thus we have $N_{hard}(a, b_k) = N_{comb,r=K}(a, b_k) \leq N_{comb,r=10}(a, b_k)$ when $b_k = 4.6052$. So one may wonder why our numerical simulations lead to identical results? One explanation is that with such a choice of $b_k = 4.6052$, when no event occurs, on average there is at most 1 non-zero sensor message received in the fusion center at any given time, and thus there is

little difference whether one uses the sum of the largest $r = 10$ sensor messages or uses the sum of all $K = 100$ sensor messages. Hence similar performances are observed in finite-sample simulations.

2.6 Proofs

Proof of Theorem 2.4.1. Let us first focus part (i) on the properties of the hard-thresholding scheme $N_{hard}(a, b)$ in (2.9) with $b \geq 0$ being the common constant for b_k's in (2.14)-(2.15).

To prove relation (2.18), it is clear that the worst-case detection delay of $N_{hard}(a, b)$ occurs at the change-point $\nu = 1$, and thus it suffices to show that $\mathbf{E}^{(\nu=1)}_{\delta_1,\ldots,\delta_K}(N_{hard}(a, b))$ satisfies (2.18). Without loss of generality, we assume that only the first m data steams are affected and no other data streams are affected. To simplify our notation below, denote $\delta_{\max} = \max_{1\leq i\leq m} \delta_i$. It suffices to show that

$$\mathbf{E}^{(\nu=1)}_{\delta_1,\ldots,\delta_K}(N_{hard}(a, b)) \leq \max\left\{\frac{a}{\sum_{k=1}^m I(g_k, f_k)}, \frac{b}{\sum_{k=1}^K I(g_k, f_k)}\right\} + O(\sqrt{b}) + O(1) + \delta_{\max}, \tag{2.28}$$

for any $b \geq 0$.

The essential idea in the proof of (2.28) is to compare $N_{hard}(a, b)$ with new stopping times that are only based on those affected m data streams. Define a stopping time that is in the form of the one-sided sequential probability ratio test (SPRT):

$$\tau(a, b) = \text{first } n \text{ such that } \sum_{i=1}^n \sum_{k=1}^m \log \frac{g_k(X_{k,i})}{f_k(X_{k,i})} \geq a \text{ and } \sum_{i=1}^n \log \frac{g_k(X_{k,i})}{f_k(X_{k,i})} \geq \rho_k b \text{ for all } 1 \leq k \leq m, \tag{2.29}$$

where the weights ρ_k's are defined in (2.15), and let $\hat{\tau}_\delta(a, b)$ be the new stopping time that applies $\tau(a, b)$ to the new observations after time $\delta_{\max}$.

Now whenever $\hat{\tau}_\delta(a, b)$ stops at time $n_0+\delta_{\max}$, we know that $\tau(a, b)$ stops after applying

it to n_0 observations $(X_{k,\delta_{\max}+1}, \cdots, X_{k,\delta_{\max}+n_0})$ for each k. By the definition of the local CUSUM statistics in (2.5), we have

$$W_{k,n_0+\delta_{\max}} \geq \sum_{i=\delta_{\max}+1}^{\delta_{\max}+n_0} \log \frac{g_k(X_{k,i})}{f_k(X_{k,i})} \geq \rho_k b$$

for all $1 \leq k \leq m$. Hence,

$$\sum_{k=1}^{K} W_{k,n_0+\delta_{\max}} \mathbf{1}\{W_{k,n_0+\delta_{\max}} \geq \rho_k b\} \geq \sum_{k=1}^{m} \sum_{i=\delta_{\max}+1}^{\delta_{\max}+n_0} \log \frac{g_k(X_{k,i})}{f_k(X_{k,i})} \geq a,$$

where the last relation is from the definition of $\tau(a,b)$. This implies that the scheme $N_{hard}(a,b)$ must stop at time $n_0 + \delta_{\max}$, and possibly earlier. Thus

$$\mathbf{E}^{(\nu=1)}_{\delta_1,\ldots,\delta_K}(N_{hard}(a,b)) \quad \leq \quad \mathbf{E}^{(\nu=1)}_{\delta_1,\ldots,\delta_K}(\hat{\tau}_\delta(a,b)) = \delta_{\max} + \mathbf{E}^{(\nu=1)}_{\delta_1^*,\ldots,\delta_K^*}(\tau(a,b)),$$

where δ_k^* is the binary version of δ_k's defined in (2.37). To simplify the notation, denote by $\mathbf{E}^{(1)}$ the expectation when the change occurs at time $\nu = 1$ and the event affects the first m data streams immediately but does not affect the other remaining $K - m$ data streams. So it suffices to show that the stopping time $\tau(a,b)$ in (2.29) satisfies

$$\mathbf{E}^{(1)}(\tau(a,b)) \leq \max\left\{ \frac{a}{\sum_{k=1}^{m} I(g_k, f_k)}, \frac{b}{\sum_{k=1}^{K} I(g_k, f_k)} \right\} + O(\sqrt{b}) + O(1). \quad (2.30)$$

To prove (2.30), for $1 \le k \le m$, let

$$\begin{aligned}
M_k &= \inf\Big\{n \ge 1 : \sum_{i=1}^{n} \log \frac{g_k(X_{k,i})}{f_k(X_{k,i})} \ge \rho_k b\Big\}, \\
\tau_k(M_k) &= \sup\Big\{n \ge 1 : \sum_{i=M_k+1}^{M_k+n} \log \frac{g_k(X_{k,i})}{f_k(X_{k,i})} \le 0\Big\} \\
\hat{M} &= \max_{1\le k\le m}\Big(M_k + \tau_k(M_k) + 1\Big) \\
t(\hat{M}) &= \inf\Big\{n \ge 1 : \sum_{i=\hat{M}+1}^{\hat{M}+n} \Big(\sum_{k=1}^{m} \log \frac{g_k(X_{k,i})}{f_k(X_{k,i})}\Big) \ge \max\{a - (\sum_{k=1}^{m} \rho_k)b, 0\}\Big\}.
\end{aligned}$$

Combining these definitions with those of $\tau(a,b)$ in (2.29) yields that

$$\begin{aligned}
\tau(a,b) &\le \hat{M} + t(\hat{M}) = \max_{1\le k\le m}\Big(M_k + \tau_k(M_k) + 1\Big) + t(\hat{M}) \\
&\le \sum_{k=1}^{m} \tau_k(M_k) + 1 + t(\hat{M}) + \max_{1\le k\le m} M_k.
\end{aligned}$$

Hence, relation (2.30) holds if we can establish the following three relations:

$$\mathbf{E}^{(1)}\Big(\tau_k(M_k)\Big) = O(1) \qquad \text{for all } 1 \le k \le m; \tag{2.31}$$

$$\mathbf{E}^{(1)}\Big(t(\hat{M})\Big) \le \max\left\{\frac{a}{\sum_{k=1}^{m} I(g_k, f_k)} - \frac{b}{\sum_{k=1}^{K} I(g_k, f_k)}, 0\right\} + O(1); \tag{2.32}$$

$$\mathbf{E}^{(1)}\Big(\max_{1\le k\le m} M_k\Big) \le \frac{b}{\sum_{k=1}^{K} I(g_k, f_k)} + O(\sqrt{b}) + O(1). \tag{2.33}$$

Relation (2.31) is well-known in renewal theory, e.g., Theorem D in Kiefer and Sacks, 1963, since $\log\big(g_k(X)/f_k(X)\big)$ has positive mean and finite variance under $\mathbf{E}^{(1)}$ by our assumptions in (2.1) and (2.2).

For relation (2.32), by the definition of $t(\hat{M})$, when $a \le (\sum_{k=1}^{m} \rho_k)b$, the threshold becomes 0 and thus $t(\hat{M}) = 0$. When $a \ge (\sum_{k=1}^{m} \rho_k)b$, the stopping time $t(\hat{M})$ is defined when a random walk exceeds the bound $a-(\sum_{k=1}^{m} \rho_k)b$, the application of standard renewal

theory yields that

$$\begin{aligned}\mathbf{E}^{(1)}(t(\hat{M})) &= \frac{a-(\sum_{k=1}^{m}\rho_k)b}{\sum_{k=1}^{m} I(g_k,f_k)}+O(1)\\ &= \frac{a}{\sum_{k=1}^{m} I(g_k,f_k)}-\frac{b}{\sum_{k=1}^{K} I(g_k,f_k)}+O(1),\end{aligned}$$

see, for example, Siegmund (1985). Here the second equation follows from the definition of ρ_k in (2.15) that

$$\frac{\sum_{k=1}^{m}\rho_k}{\sum_{k=1}^{m} I(g_k,f_k)}=\frac{1}{\sum_{k=1}^{K} I(g_k,f_k)}.$$

Thus relation (2.32) holds.

The proof of relation (2.33) is a little more complicated, but it can be done along the same line as that in Mei (2005). The key fact is that the choice of $b_k = \rho_k b$'s in (2.14)-(2.15) makes sure that the stopping times M_k's have roughly the same mean under $\mathbf{P}^{(1)}$. Specifically, by renewal theory and the assumptions of (f_k, g_k) in (2.1) and (2.2), under $\mathbf{P}^{(1)}$,

$$\mathbf{E}^{(1)}(M_k)=\frac{\rho_k b}{I(g_k,f_k)}+O(1)=\frac{b}{\sum_{k=1}^{K} I(g_k,f_k)}+O(1)$$

and $\text{Var}^{(1)}(M_k) = O(b)$, as $b \to \infty$, see Siegmund Siegmund (1985). Thus as $b \to \infty$,

$$\begin{aligned}\Big(\mathbf{E}^{(1)}\big|M_k-\frac{b}{\sum_{k=1}^{K} I(g_k,f_k)}\big|\Big)^2 &\le \mathbf{E}^{(1)}\Big(M_k-\frac{b}{\sum_{k=1}^{K} I(g_k,f_k)}\Big)^2\\ &= \text{Var}^{(1)}(M_k)+\Big(\mathbf{E}^{(1)}M_k-\frac{b}{\sum_{k=1}^{K} I(g_k,f_k)}\Big)^2\\ &\le C_{1k}b,\end{aligned}$$

where $C_{1k} > 0$ is a constant. Taking square root both sides, and noticing that $M_k = M_k(b)$ is an increasing function of $b \ge 0$, it is not difficult to show that for each $k = 1, \cdots, K$, there exists a constant $C_{2k} > 0$ so that

$$\mathbf{E}^{(1)}\big|M_k-\frac{b}{\sum_{k=1}^{K} I(g_k,f_k)}\big| \le \max(C_{2k}, \sqrt{C_{1k}}\sqrt{b}),$$

for all $b > 0$.

Therefore,

$$\begin{aligned}
\mathbf{E}^{(1)}\Big(\max_{1\le k\le m} M_k\Big) &= \frac{b}{\sum_{k=1}^K I(g_k, f_k)} + \mathbf{E}^{(1)} \max_{1\le k\le m}\Big(M_k - \frac{b}{\sum_{k=1}^K I(g_k, f_k)}\Big) \\
&\le \frac{b}{\sum_{k=1}^K I(g_k, f_k)} + \sum_{k=1}^m \mathbf{E}^{(1)}\Big|M_k - \frac{b}{\sum_{k=1}^K I(g_k, f_k)}\Big| \\
&\le \frac{b}{\sum_{k=1}^K I(g_k, f_k)} + \sum_{k=1}^m \max(C_{2k}, \sqrt{C_{1k}}\sqrt{b}) \\
&\le \frac{b}{\sum_{k=1}^K I(g_k, f_k)} + C(\sqrt{b}+1),
\end{aligned}$$

where the constant $C = \sum_{k=1}^K \max(C_{2k}, \sqrt{C_{1k}})$ does not depend on b. This proves relation (2.33). Therefore, relations (2.31)-(2.33) hold, and thus relation (2.18) holds for the hard-thresholding scheme $N_{hard}(a, b)$ in (2.9).

The proof for the soft-thresholding scheme $N_{soft}(a, b)$ in (2.10) is similar, except defining the stoping time $\tau(a, b)$ by

$$\tau(a,b) = \text{first } n \text{ such that } \sum_{i=1}^n \sum_{k=1}^m \log\frac{g_k(X_{k,i})}{f_k(X_{k,i})} \ge a + b\sum_{k=1}^m \rho_k \text{ and} \tag{2.34}$$

$$\sum_{i=1}^n \log\frac{g_k(X_{k,i})}{f_k(X_{k,i})} \ge \rho_k b \text{ for all } 1 \le k \le m, \tag{2.35}$$

instead of (2.29) and prove

$$\mathbf{E}^{(1)}(\tau(a,b)) \le \frac{a}{\sum_{k=1}^m I(g_k, f_k)} + \frac{b}{\sum_{k=1}^K I(g_k, f_k)} + O(\sqrt{b}) + O(1). \tag{2.36}$$

by replacing the threshold $\max\{a-(\sum_{k=1}^m \rho_k)b, 0\}$ in the stopping time $t(\hat{M})$ by the threshold a. The remaining arguments are identical and thus omitted.

Now let us provide a sketch of the proof for part (iii) of Theorem 2.4.1 on the order-thresholding scheme $N_{order,r}(a)$ in (2.13) and the combined thresholding scheme $N_{comb,r}(a, b)$ in (2.12). Since $N_{order,r}(a)$ is a special case of $N_{comb,r}(a, b)$ with $b = 0$, it suffices to prove

the theorem for $N_{comb,r}(a,b)$ in (2.12) with $b \geq 0$. Clearly relation (2.38) also holds for $N_{comb,r}(a,b)$ for any $b \geq 0$, because the "SUM" scheme $T_{\rm sum}(a)$ again provides the lower bound for $N_{comb,r}(a,b)$.

It remains to show that relation (2.18) holds for $N_{comb,r}(a,b)$ with $b \geq 0$ in the scenario when the occurring event affects at most r data streams, i.e., when $\sum_{k=1}^{K} I\{\delta_k < \infty\} \leq r$. Without loss of generality, assume that the affected data streams are just the first m data streams with $m \leq r$. Recall that $U_{k,n} = W_{k,n} I\{W_{k,n} \geq \rho_k b\}$, and we order the $U_{k,n}$'s as $U_{(1),n} \geq \ldots \geq U_{(K),n}$, and $N_{comb,r}(a,b)$ stops if $\sum_{k=1}^{r} U_{(k),n} \geq a$. Note that if $m \leq r$,

$$\sum_{k=1}^{r} U_{(k),n} \;\; \geq \;\; \sum_{k=1}^{r} U_{k,n} \geq \sum_{k=1}^{m} U_{k,n},$$

since $U_{k,n} \geq 0$. Thus, if at some time n_0 we have $W_{k,n_0} \geq \rho_k b$ and $\sum_{k=1}^{m} W_{k,n_0} \geq a$ for $1 \leq k \leq m$ (i.e., for the first m data streams), then $N_{comb,r}(a,b)$ will also stop at time n_0 and possibly earlier. Hence, whenever $m \leq r$, the stopping time $\tau(a,b)$ in (2.29) also provides an upper bound on the detection delay of $N_{comb,r}(a,b)$. Thus the proposed combined thresholding scheme $N_{comb,r}(a,b)$ in (2.12) satisfies relation (2.18) whenever the occurring event affects at most r data streams. This completes the proof of the theorem. □

Proof of Lemma 2.4.1. Intuitively, only those affected sensors provide information to detect the occurring events, and the quickest possible way to detect the occurring event is when the event affects the sensors instantaneously. More rigorously, if we define

$$\delta_k^* = \begin{cases} 0, & \text{if } \delta_k \text{ is finite} \\ \infty, & \text{if } \delta_k = \infty \end{cases}, \tag{2.37}$$

then for any given scheme $T(\gamma)$,

$$\overline{\mathbf{E}}_{\delta_1,\ldots,\delta_K}(T(\gamma)) \geq \inf_{\tau} \overline{\mathbf{E}}_{\delta_1^*,\ldots,\delta_K^*}(\tau),$$

where the infimum is taken over all possible schemes τ satisfying the false alarm constraint γ in (2.4). An alternative and possible better viewpoint is based on a time-shifting argument in which one imagines that at time n one observes the observations $X_{k,n+\delta_k}$ (instead of $X_{k,n}$) when δ_k is finite, and then applies $T(\gamma)$ to the new aligned observations.

Without loss of generality, assume that the first m data streams are affected abruptly and simultaneously by the event at unknown time ν, and other data streams are unaffected. That is, m out of K data streams are affected by the event, and $\delta_i^* = 0$ for $1 \leq i \leq m$, and $= \infty$ for $m+1 \leq i \leq K$. By (2.17), we have

$$J(\delta_1, \ldots, \delta_K) = J(\delta_1^*, \ldots, \delta_K^*) = \sum_{i=1}^{m} I(g_i, f_i).$$

In this case, we face the sequential change detection problem when the distribution of $(X_{1,n}, \cdots, X_{K,n})$ changes from $(f_1, \cdots, f_m, f_{m+1}, \cdots, f_K)$ to $(g_1, \cdots, g_m, f_{m+1}, \cdots, f_K)$. It is well-known (Lorden, 1971) that

$$\inf_{\tau} \overline{\mathbf{E}}_{\delta_1^*, \ldots, \delta_K^*}(\tau) \geq (1 + o(1)) \frac{\log \gamma}{\sum_{i=1}^{m} I(g_i, f_i)}.$$

subject to the false alarm constraint γ in (2.4) as $\gamma \to \infty$. Combining the above results yields relation (2.20), completing the proof of Lemma 3.2.1. □

Proof of Theorem 2.4.2: First, we will prove for any $a, b \geq 0$,

$$\mathbf{E}^{(\infty)}(N_{hard}(a,b)) \geq (1 + o(1)) \frac{e^a}{1 + a + \frac{a^2}{2!} + \cdots + \frac{a^{K-1}}{(K-1)!}}. \tag{2.38}$$

To prove (2.38), note that $N_{hard}(a,b)$ in (2.9) is increasing as a function of $b \geq 0$, and when $b = 0$, $N_{hard}(a, b=0)$ reduces to the "SUM" scheme $T_{\text{sum}}(a)$ in (2.7). Hence, for any $b \geq 0$, $N_{hard}(a,b) \geq T_{\text{sum}}(a)$ and of course, $\mathbf{E}^{(\infty)}(N_{hard}(a,b)) \geq \mathbf{E}^{(\infty)}(T_{\text{sum}}(a))$. By Theorem 1 of Mei (2010), the "SUM" scheme $T_{\text{sum}}(a)$ satisfies relation (2.38), and so are the hard-thresholding schemes $N_{hard}(a,b)$ for all $b \geq 0$.

Theorem 2.4.2 follows at once from Theorem 2.4.1 and (2.38). In particular, the choice of a_γ in (2.21) follows from (2.38) and the fact that $1 + a + \frac{a^2}{2!} + \cdots + \frac{a^{K-1}}{(K-1)!} \sim \frac{a^{K-1}}{(K-1)!}$ if K is fixed and a goes to ∞. □

Proof of Theorem 2.4.3: Clearly, we can see for any fixed combination of (a, b), $\mathbf{E}^{(\infty)}N_{hard}(a, b)$ is smaller than $\mathbf{E}^{(\infty)}N_{soft}(a, b)$ or $\mathbf{E}^{(\infty)}N_{comb,r}(a, b)$. Therefore, it is sufficient to prove the choice of a in (2.23) could guarantee the hard-thresholding scheme $N_{hard}(a, b)$ satisfies false alarm constraint (2.4).

First, define $W_k^* = \lim_{n\to\infty} W_{k,n}$ as the limit of the CUSUM statistics, which has the following non-asymptotic result: for *any* $x > 0$, the tail probability

$$G(x) = \mathbf{P}^{(\infty)}(W_k^* > x) \leq e^{-x}, \tag{2.39}$$

see Appendix 2 on Page 245 of Siegmund (1985). It is clear that W_k^* are i.i.d. across different k. Now we define the log-moment generating function of the W_k^*'s

$$\psi(\theta) = \log \mathbf{E}^{(\infty)} \exp\{\theta W_k^* \mathbf{1}\{W_k^* \geq b/K\}\} \tag{2.40}$$

For any $x \geq 0$, by Chebyshev's inequality,

$$\begin{aligned}
\mathbf{E}^{(\infty)}[N_{hard}(a, b)] &\geq x\mathbf{P}^{(\infty)}(N_{hard}(a, b) \geq x) \\
&= x\left[1 - \mathbf{P}^{(\infty)}(N_{hard}(a, b) < x)\right] \\
&= x\left[1 - \mathbf{P}^{(\infty)}(\sum_{k=1}^{K} W_{k,n}\mathbf{1}\{W_{k,n} \geq b_k\} \geq a) \text{ for some } 1 \leq n \leq x\right] \\
&\geq x\left[1 - x\mathbf{P}^{(\infty)}(\sum_{k=1}^{K} W_k^*\mathbf{1}\{W_k^* \geq b_k\} \geq a)\right], \\
&\geq x\left[1 - xe^{-\theta a}\mathbf{E}^{(\infty)} \exp(\theta \sum_{k=1}^{K} W_k^*\mathbf{1}\{W_k^* \geq b/K\})\right] \\
&= x\left[1 - x\exp\left(-\theta a + K\psi(\theta)\right)\right]. \qquad (2.41)
\end{aligned}$$

Note that for any $u > 0$, the function $x(1 - xu)$ is maximized at $x = 1/(2u)$ with the

maximum value $1/(4u)$. Therefore, we can get for any $0 < \theta < 1$,

$$\mathbf{E}^{(\infty)}[N_{hard}(a,b)] \;\geq\; \frac{1}{4}\exp\left(\theta a - K\psi(\theta)\right). \tag{2.42}$$

By the definition of $\psi(\theta)$ in (2.40) and the tail probability W_k^* in (2.39), for all $0 < \theta < 1$,

$$\begin{aligned}
\psi(\theta) &= \log[\mathbf{P}^{(\infty)}(W_k^* \leq b/K) - \int_{b/K}^{\infty} e^{\theta x} dG(x)] \\
&= \log[1 + (e^{\theta b/K} - 1)G(b) + \theta \int_{b/K}^{\infty} e^{\theta x} G(x) dx)] \\
&\leq \log[1 + (e^{\theta b/K} - 1)e^{-b/K} + \theta \int_{b/K}^{\infty} e^{\theta x} G(x) dx] \\
&\leq \log[1 + (e^{\theta b/K} - 1)e^{-b/K} + \theta \int_{b/K}^{\infty} e^{\theta x} e^{-x} dx] \\
&= \log\left(1 + \frac{1}{1-\theta} e^{-b(1-\theta)/K} - e^{-b/K}\right) \\
&\leq \frac{1}{1-\theta} e^{-b(1-\theta)/K} - e^{-b/K} \\
&\leq \frac{1}{1-\theta} - e^{-b/K}
\end{aligned} \tag{2.43}$$

where the second equation is based on the integration by parts. By (2.42) and (2.43), we have

$$\mathbf{E}^{(\infty)} N_{hard}(a,b) \;\geq\; \frac{1}{4}\exp\left(\theta a - \frac{K}{1-\theta} + Ke^{-b/K}\right) \tag{2.44}$$

for all $0 < \theta < 1$. If $K < a$, by letting $\theta = 1 - \sqrt{K/a}$ yield

$$\mathbf{E}^{(\infty)} N_{hard}(a,b) \;\geq\; \frac{1}{4}\exp\left((\sqrt{a} - \sqrt{K})^2 + Ke^{-b/K} - K\right) \tag{2.45}$$

Therefore a choice of

$$a = (\sqrt{\log(4\gamma) + K - Ke^{-b/K}} + \sqrt{K})^2 \tag{2.46}$$

will guarantee the hard-shrinkage scheme $N_{hard}(a,b)$ satisfies the false alarm constraint (2.4).

Note using the continuity of the soft-thresholding transformation function, a tighter bound for $N_{soft}(a,b)$ was derived for the soft-thresholding scheme in Liu, Zhang, and Mei (2019), although they are asymptotically equivalent to those in Theorem 2.4.3 and Corollary 2.4.1 $N_{hard}(a,b)$ as the dimension K goes to ∞. □

Proof of Corollary 2.4.1:

If $K = o(\log\gamma)$, the corresponding $a = a_\gamma = \log(4\gamma)+o(\log\gamma)$ will guarantee the false alarm constraint. Moreover, if m is fixed and $b = o(\log\gamma)$, the upper bound of detection delay in Theorem 2.4.1 could be applied and yields

$$\overline{\mathbf{E}}_{\delta_1,\ldots,\delta_K}(N_{hard}(a,b)) \leq (1+o(1))\left(\frac{\log\gamma}{mI}\right)+O(1), \tag{2.47}$$

which implies the first order detection efficiency will be kept as long as $b = o(\log\gamma)$.

If $K >> \log\gamma$, the corresponding $a = (1+o(1))K$ will guarantee the false alarm constraint. Moreover, since m is fixed and $b = o(K^2)$, the upper bound of detection delay in Theorem 2.4.1 could be applied and yields

$$\overline{\mathbf{E}}_{\delta_1,\ldots,\delta_K}(N_{hard}(a,b)) \leq (1+o(1))\left(\frac{K}{mI}\right)+O(1), \tag{2.48}$$

which completes the proof of corollary. □

CHAPTER 3

ROBUSTNESS AND TRACTABILITY FOR NON-CONVEX M-ESTIMATORS

3.1 Introduction

M-estimation plays an important role in linear regression due to its robustness and flexibility. From the statistical viewpoint, it has been shown that many M-estimators enjoy desirable robustness properties in the presence of outliers, as well as asymptotic normality when the data are normally distributed without outliers. Some general theoretical properties and review of robust M-estimators can be found in Bai, Rao, and Wu (1992), Huber and Ronchetti (2009), Hampel, Ronchetti, Rousseeuw, and Stahel (2011), and El Karoui, Bean, Bickel, Lim, and Yu (2013). In the high-dimensional setting, where the dimensionality is greater than the number of samples, penalized M-estimators have been widely used to tackle the challenges of outliers and have been used for sparse recovery and variable selection, see Lambert-Lacroix and Zwald (2011), Li, Peng, and Zhu (2011), Wang, Jiang, Huang, and Zhang (2013), and Loh (2017). However, from the computational tractability perspective, it is often not easy to compute the M-estimators, since optimization problems over non-convex loss functions are usually involved. Moreover, the tractability issue becomes more challenging when the data are contaminated by some arbitrary outliers, which is essentially the situation where robust M-estimator is designed to tackle.

This chapter aims to investigate two important properties of M-estimators, *robustness* and *tractability*, simultaneously under *the gross error model*. Specifically, we assume the data generation model is $y_i = \langle \theta_0, x_i \rangle + \epsilon_i$, where $y_i \in \mathbb{R}, x_i \in \mathbb{R}^p$, , for $i = 1, \cdots, n$, and the noise term ϵ_i's are from Huber's gross error model (Huber, 1964): $\epsilon_i \sim (1-\delta) f_0 + \delta g$, for $i = 1, \cdots, n$. Here, f_0 denotes the probability density function (pdf) of the noise of the normal samples, which has the desirable properties, such as zero mean and finite

variance; g denotes the pdf of the outliers (contaminations), which may also depend on the explanatory variable x_i, for $i = 1, \cdots, n$. One thing to notice is that we do not require the mean of g to be 0. The parameter $\delta \in [0, 1]$, denotes the percentage of the contaminations, which is also known as the contamination ratio in robust statistics literature. The gross error model indicates that for the i^{th} sample, the residual term ϵ_i is generated from the pdf f_0 with probability $1 - \delta$, and from the pdf g with probability δ. It is important to point out that the residual ϵ_i is independent of x_i and other x_j's when it is from the pdf f_0, but can be dependent with the variable x_i when it is from the pdf g.

In the first part of this chapter, we start with the low-dimensional case when the dimension p is fixed. We consider the robust M-estimation with a constraint on the norm of θ. Mathematically, we study the following optimization problem:

$$\begin{aligned} \underset{\theta}{\text{Minimize:}} \quad & \hat{R}_n(\theta) := \frac{1}{n} \sum_{i=1}^{n} \rho(y_i - \langle \theta, x_i \rangle), \\ \text{subject to:} \quad & \|\theta\|_2 \leq r. \end{aligned} \tag{3.1}$$

Here, $\rho : \mathbb{R} \to \mathbb{R}$ is the loss function, and is often *non-convex*. We consider the problem with the ℓ_2 constraint due to three reasons: first, it is well know the constraint optimization problem in (3.1) is equivalent to the unconstraint optimization problem with a ℓ_2 regularizer. Therefore, it is related to the Ridge regression, which can alleviate multicollinearity amongst regression predictors. Second, by considering the problem of (3.1) in a compact ball with radius r, it guarantees the existence of the global optimal, which is necessary for establishing the tractability properties of the M-estimator. Finally, by working on the constrained optimization problem, we can avoid technical complications and establish the uniform convergence theorems of the empirical risk and population risk. Besides, the constrained M-estimators are widely used and studied in the literature, see Mei, Bai, and Montanari (2018) and Loh (2017) for more details. To be consistent with the assumptions used in the literature, in the current work, we assume r is a constant and the true parameter θ_0 is

inside of the ball.

In the second part, we extend our research to the high-dimensional case, where $p \gg n$ and the true parameter θ_0 is sparse. In order to achieve the sparsity in the resulting estimator, we consider the penalized M-estimator with ℓ_1 regularizer:

$$\begin{aligned} &\underset{\theta}{\text{Minimize:}} \quad \hat{L}_n(\theta) := \frac{1}{n}\sum_{i=1}^{n} \rho(y_i - \langle \theta, x_i \rangle) + \lambda_n ||\theta||_1, \\ &\text{subject to:} \quad \|\theta\|_2 \leq r. \end{aligned} \tag{3.2}$$

Note the corresponding penalized M-estimator with a L_2 constraint is related to the Elastic net, which overcomes the limitations of the LASSO type regularization (Zou and Hastie, 2005).

In both parts, we will show that (in the finite sample setting,) the M-estimator obtained from (3.1) or (3.2) is robust in the sense that all stationary points of empirical risk function $\hat{R}_n(\theta)$ or $\hat{L}_n(\theta)$ are bounded in the neighborhood of the true parameter θ_0 when the proportion of outliers is small. In addition, we will show that with a high probability, there is a unique stationary point of the empirical risk function, which is the global minimizer of (3.1) or (3.2) for some general (possibly nonconvex) loss functions ρ. This implies that the M-estimator can be computed efficiently. To illustrate our general theoretical results, we study some specific M-estimators with Huber's loss (Huber, 1964) and Welsch's exponential squared loss (Dennis Jr and Welsch, 1978), and explicitly discuss how the tuning parameter and percentage of outliers affect the robustness and tractability of the corresponding M-estimators.

Our research makes several fundamental contributions on the field of robust statistics and non-convex optimization. First, we demonstrate the uniform convergence results for the gradient and Hessian of the empirical risk to the population risk under the gross error model. Second, we provide nonasymptotic upper bound of the estimation error for the general M-estimators, which nearly achieve the minimax error bound in Chen, Gao, and

Ren (2016). Third, we investigate the computational tractability of the general nonconvex M-estimators under the gross error model and show when the contamination ratio δ is small, there is only one unique stationary point of the empirical risk function. Therefore, efficient algorithms such as gradient descent or proximal gradient decent can be guaranteed to converge to a unique global minimum irrespective of the initialization. Our general results also imply the following interesting and to some extent surprising statement: the percentage of outliers has an impact on the *tractability* of non-convex M-estimators. In a nutshell, the estimation and the corresponding optimization problem become more difficult both in terms of solution quality and computational efficiency when more outliers appear. While the former is well expected, we find the latter – that more outliers make M-estimators more difficult to numerically compute – an interesting and somewhat surprising discovery. Our simulation results and case study also verify this phenomenon.

Related work

Since Huber's pioneer work on robust M-estimators (Huber, 1964), many M-estimators with different choices of loss functions have been proposed, e.g., Huber's loss (Huber, 1964), Andrews sine loss (Andrews, Bickel, Hampel, Huber, Rogers, and W.Tukey, 1972), Tukey's Bisquare loss (Beaton and Tukey, 1974), Welsch's exponential squared loss (Dennis Jr and Welsch, 1978), to name a few. From the statistical perspective, much research has been done to investigate the robustness of M-estimators such as large breakdown point (Donoho and Huber, 1983; Mizera and Müller, 1999; Alfons, Croux, and Gelper, 2013), finite influent function (Hampel, Ronchetti, Rousseeuw, and Stahel, 2011) and asymptotic normality (Maronna and Yohai, 1981; Lehmann and Casella, 2006; El Karoui, Bean, Bickel, Lim, and Yu, 2013). Recently, in the high-dimensional context, regularized M-estimators have received a lot of attentions. Lambert-Lacroix and Zwald (2011) proposed a robust variable selection method by combing Huber's loss and adaptive lasso penalty. Li, Peng, and Zhu (2011) show the nonconcave penalized M-estimation method can perform

parameter estimation and variable selection simultaneously. Welsch's exponential squared loss combined with adaptive lasso penalty is used by Wang, Jiang, Huang, and Zhang (2013) to construct a robust estimator for sparse estimation and variable selection. Chang, Roberts, and Welsh (2018) proposed a robust estimator by combining the Tukey's biweight loss with adaptive lasso penalty. However, those statistical works did not discuss the computational tractability of the M-estimators even though many of these loss functions are non-convex.

During the last several years, non-convex optimization has attracted fast growing interests due to its ubiquitous applications in machine learning and in particular deep learning, such as dictionary learning (Mairal, Bach, Ponce, and Sapiro, 2009), phase retrieval (Candes, Li, and Soltanolkotabi, 2015), orthogonal tensor decomposition (Anandkumar, Ge, Hsu, Kakade, and Telgarsky, 2014) and training deep neural networks (Bengio, 2009). It is well known that there is no efficient algorithm that can guarantee to find the global optimal solution for general non-convex optimization.

Fortunately, in the context of estimating non-convex M-estimators for high-dimensional linear regression (*without outliers*), under some mild statistical assumptions, Loh (2017) establishes the uniqueness of the stationary point of the non-convex M-estimator when using some non-convex bounded regularizers instead of ℓ_1 regularizer. By investigating the uniform convergence of gradient and Hessian of the empirical risk, Mei, Bai, and Montanari (2018) prove that with a high probability, there exists one unique stationary point of the regularized empirical risk function with ℓ_1 regularizer. Thus regardless of the initial points, many computational efficient algorithm such as gradient descent or proximal gradient descent algorithm could be applied and are guaranteed to converge to the global optimizer, which implies the high tractability of the M-estimator. However, their analysis is restricted to the standard linear regression setting without outliers. In particular, they assume the distribution of the noise terms in the linear regression model should have some desirable properties such as zero mean, sub-gaussian and independent of feature vector x,

which might not hold when the data are contaminated with outliers. To the best of our knowledge, no research has been done on analyzing the computational tractability properties of the non-convex M-estimators when data are contaminated by arbitrary outliers, although the very reason why M-estimators are proposed is to handle outliers in linear regression in the robust statistics literature. Our research is the first to fill the significant gap on the tractability of non-convex M-estimators. We prove that under mild assumptions, many M-estimators can tolerate a small amount of arbitrary outliers in the sense of keeping the tractability, even if the loss functions are non-convex.

Notations. Given $\mu, \nu \in \mathbb{R}^p$, their standard inner product is defined by $\langle \mu, \nu \rangle = \sum_{i=1}^{p} \mu_i \nu_i$. The ℓ_p norm of a vector x is denoted by $||x||_p$. The p by p identity matrix is denoted by $I_{p\times p}$. Given a matrix $M \in \mathbb{R}^{m\times m}$, let $\lambda_{\max}(M), \lambda_{\min}(M)$ denote the largest and the smallest eigenvalue of M, respectively. The operator norm of M is denoted by $||M||_{op}$, which is equal to $\max(\lambda_{\max}(M), -\lambda_{\min}(M))$ when $M \in \mathbb{R}^{m\times m}$. Let $B_q^p(a, r) = \{x \in \mathbb{R}^p : ||x - a||_q \leq r\}$, be the ℓ_q ball in the $\mathbb{R}^p$ space with center a and radius r. Given a random variable X with probability density function f, we denote the corresponding expectation by $\mathbf{E}_f$. We will often omit the density function subscript f when it is clear from the context, the expectation is taken for all variables.

Organization. The rest of this chapter is organized as follows. In Section 3.2, we present the theorems about the robustness and tractability of general M-estimators under the low-dimensional setup when dimension p is fixed and less than n. Then in Section 3.3, we consider the penalized M-estimator with ℓ_1 regularizer in the high-dimensional regression when $p \gg n$. The ℓ_2 error bounds of the estimation and the scenario when the M-estimator has nice tractability are provided. In Section 3.4, we discuss two special families of robust estimator constructed by Huber's and Welsch's exponential loss as examples to illustrate our general theorems of robustness and tractability of M-estimators. Simulation results are presented in Section 3.5 and a case study is shown in Section 3.6 to illustrate the robustness and tractability properties. We relegate all proofs to the Section 3.7 due to space limits.

3.2 M-estimators in the low-dimensional regime

In this section, we investigate two key properties of M-estimators, namely *robustness* and *tractability*, in the setting of linear regression with arbitrary outliers in the low-dimensional regime where the dimension p is fixed and smaller than the number of samples n. In terms of robustness, we show that under some mild conditions, any stationary point of the objective function in (3.1) will be well bounded in a neighborhood of the true parameter θ_0. Moreover, the neighborhood shrinks when the proportion of outliers decreases. In terms of tractability, we show that when the proportion of outliers is small and the sample size is large, with a high probability, there is a *unique stationary point* of the empirical risk function, which is the global optimum (and hence the corresponding M-estimator). Consequently, many first order methods are guaranteed to converge to the global optimum, irrespective of initialization.

Before presenting our main theorems, we make the following mild assumptions on the loss function ρ, the explanatory or feature vectors x_i, and the idealized noise distribution f_0. We define the score function $\psi(z) := \rho'(z)$.

Assumption 3.2.1. **(a)** *The score function $\psi(z)$ is twice differentiable and odd in z with $\psi(z) \geq 0$ for all $z \geq 0$. Moreover, we assume $\max\{||\psi(z)||_\infty, ||\psi'(z)||_\infty, ||\psi''(z)||_\infty\} \leq L_\psi$.*

(b) *The feature vector x_i are i.i.d with zero mean and τ^2-sub-Gaussain, that is $\mathbf{E}[e^{\langle \lambda, x_i \rangle}] \leq \exp(\frac{1}{2}\tau^2||\lambda||_2^2)$, for all $\lambda \in \mathbb{R}^p$.*

(c) *The feature vector x_i spans all possible directions in $\mathbb{R}^p$, that is $\mathbf{E}[x_i x_i^T] \succeq \gamma\tau^2 I_{p\times p}$, for some $0 < \gamma < 1$.*

(d) *The idealized noise distribution $f_0(\epsilon)$ is symmetric. Define $h(z) := \int_{-\infty}^{\infty} f_0(\epsilon)\psi(z+\epsilon)d\epsilon$ and $h(z)$ satisfies $h(z) > 0$, for all $z > 0$ and $h'(0) > 0$.*

Assumption (a) requires the smoothness of the loss function in the objective function,

which is crucial to study the tractability of the estimation problem; Assumption (b) assumes the sub-Gaussian design of the observed feature matrix; Assumption (c) assumes that the covariance matrix of the feature vector is positive semidefinite. We remark that the condition on $h(z)$ is mild. It is not difficult to show that it is satisfied if the idealized noise distribution $f_0(\epsilon)$ is strictly positive for all ϵ and decreasing for $\epsilon > 0$, e.g., if $f_0 =$ pdf of $N(0, \sigma^2)$.

Before presenting our main results in this section, we first define the population risk as follows:

$$R(\theta) = \mathbf{E}\hat{R}_n(\theta) = \mathbf{E}[\rho(Y - \langle \theta, X \rangle)]. \tag{3.3}$$

The high level idea is to analyze the population risk first, and then we build a link between the population risk and the empirical risk, which solves the original estimation problem. Theorem 3.2.1 below summarizes the results for the population risk function $R(\theta)$ in (3.3).

Theorem 3.2.1. *Assume that Assumption 3.2.1 holds and the true parameter θ_0 satisfies $||\theta_0||_2 \leq r/3$.*

(a) *There exists a constant $\eta_0 = \frac{\delta}{1-\delta} C_1$ such that any stationary point θ^* of $R(\theta)$ satisfies $||\theta^* - \theta_0||_2 \leq \eta_0$, where δ is the contamination ratio, and C_1 is a positive constant that only depends on $\gamma, r, \tau, \psi(z)$ and the pdf f_0, but does not depend on the outlier pdf g.*

(b) *When δ is small, there exist a constant $\eta_1 = C_2 - C_3\delta > 0$, where C_2, C_3 are two positive constants that only depend on $\gamma, r, \tau, \psi(z)$ and the pdf f_0 but not depend on the outlier pdf g, such that*

$$\lambda_{\min}(\nabla^2 R(\theta)) > 0 \tag{3.4}$$

for every θ with $||\theta_0 - \theta||_2 < \eta_1$.

(c) *There is a unique stationary point of $R(\theta)$ in the ball $B_2^p(0, r)$ as long as $\eta_0 < \eta_1$ for a given contamination ratio δ.*

It is useful to add some remarks for better understanding Theorem 3.2.1. First, recall that the noise term ϵ_i follows the gross error model: $\epsilon_i \sim (1-\delta)f_0+\delta g$, where the outlier pdf g may also depend on x_i. While the true parameter θ_0 may no longer be the stationary point of the population risk function $R(\theta)$, Theorem 3.2.1 implies that the stationary points of $R(\theta)$ will always bounded in a neighborhood of the true parameter θ_0 when the percentage of contamination δ is small. This indicates the robustness of M-estimators in the population case.

Second, Theorem 3.2.1 asserts that when there are no outliers, i.e., $\delta = 0$, the stationary point is indeed the true parameter θ_0. In addition, since the constant η_0 in (a) is an increasing function of δ whereas the constant η_1 in (b) is a decreasing function of δ, stationary points of $R(\theta)$ may disperse from the true parameter θ_0 and the strongly convex region around θ_0 will be decreasing, as the contamination ratio δ is increasing. This indicates the difficulty of optimization for large contamination ratio cases.

Third, part (c) is a direct result from part (a) and (b). Note that $\eta_0(\delta = 0) = 0 < \eta_1(\delta = 0) = C_2$, thus there exists a positive δ^*, such that $\eta_0 < \eta_1$ for any $\delta < \delta^*$. A simple lower bound on δ^* is $C_3/(C_1 + C_2 + C_3)$, since $C_1\delta < (1-\delta)(C_2 - C_3\delta)$ whenever $0 \le \delta \le C_3/(C_1 + C_2 + C_3)$.

Our next step is to link the empirical risk function (and the corresponding M-estimator) with the population version. To this end, we need the following lemma, which shows the global uniform convergence theorem of the sample gradient and Hessian.

Lemma 3.2.1. *Under Assumption 3.2.1, for any $\pi > 0$, there exists a constant C_π depending on $\pi, \gamma, r, \tau, \psi(z), h(z)$ but independent of p, n, δ and g, such that for any $\delta \ge 0$, the following hold:*

(a) *The sample gradient converges uniformly to the population gradient in Euclidean norm, i.e., if* $n \geq C_\pi p \log n$, *we have*

$$\mathbf{P}\left(\sup_{\theta \in B_2^p(0,r)} ||\nabla \hat{R}_n(\theta) - \nabla R(\theta)||_2 \leq \tau \sqrt{\frac{C_\pi p \log n}{n}}\right) \geq 1 - \pi. \tag{3.5}$$

(b) *The sample Hessian converges uniformly to the population Hessian in operator norm, i.e., if* $n \geq C_\pi p \log n$, *we have*

$$\mathbf{P}\left(\sup_{\theta \in B_2^p(0,r)} ||\nabla^2 \hat{R}_n(\theta) - \nabla^2 R(\theta)||_{op} \leq \tau^2 \sqrt{\frac{C_\pi p \log n}{n}}\right) \geq 1 - \pi. \tag{3.6}$$

We are now ready to present our main result about M-estimators by investigating the empirical risk function $\hat{R}_n(\theta)$.

Theorem 3.2.2. *Assume Assumption 3.2.1 holds and* $||\theta_0||_2 \leq r/3$. *Let us use the same notation* η_0 *and* η_1 *as in Theorem 3.2.1. Then for any* $\pi > 0$, *there exist constant* C_π *depends on* $\pi, \gamma, r, \tau, \psi, f_0$ *but independent of* n, p, δ *and* g , *such that as* $n \geq C_\pi p \log n$, *the following statements hold with probability at least* $1 - \pi$:

(a) *for all* $||\theta - \theta_0||_2 > 2\eta_0$,

$$\langle \theta - \theta_0, \nabla \widehat{R}_n(\theta) \rangle > 0. \tag{3.7}$$

(b) *for all* $||\theta - \theta_0||_2 \leq \eta_1$,

$$\lambda_{\min}(\nabla^2 \widehat{R}_n(\theta)) > 0. \tag{3.8}$$

Thus, as long as $2\eta_0 < \eta_1$, $\widehat{R}_n(\theta)$ *has a unique stationary point, which lies in the ball* $B^p(0, r)$. *This is the unique global optimal solution of (3.1), and denote this unique stationary point by* $\widehat{\theta}_n$.

(c) *There exists a positive constant κ that depends on $\pi, \gamma, r, \psi, \delta, f_0$ but independent of n, p and g, such that*

$$||\widehat{\theta}_n - \theta_0||_2 \leq \eta_0 + \frac{4\tau}{\kappa}\sqrt{\frac{C_\pi p \log n}{n}}. \tag{3.9}$$

A few remarks are in order. First, since η_0 is independent of n, p and g, Theorem 3.2.2(a) asserts that the M-estimator which minimizes $\widehat{R}_n(\theta)$ is always bounded in the ball $B_2^p(\theta_0, 2\eta_0)$, regardless of g (and hence the outliers observed). This indicates the robustness of the M-estimator, i.e., the estimates are not severely skewed by a small amount of "bad" outliers. Next, when the contamination ratio δ is small such that $2\eta_0 < \eta_1$, there is a unique stationary point of $\widehat{R}_n(\theta)$. Therefore, although the original optimization problem (3.1) is non-convex and the sample contains some arbitrary outliers, the optimal solution of $\widehat{R}_n(\theta)$ can be computed efficiently via most off-the-shelf first-order algorithms such as gradient descent or stochastic gradient descent. This indicates the tractability of the M-estimator. Interestingly, as in the population risk case, the tractability is closely related to the amount of outliers – the problem is easier to optimize when the data contains fewer outliers. Finally, when the number of samples $n \gg p \log n$, the estimation error bound η_0 is in the order of $O(\delta + \sqrt{\frac{p \log n}{n}})$, which nearly achieves the minimax lower bound of $O(\delta + \sqrt{\frac{p}{n}})$ in Chen, Gao, and Ren (2016).

3.3 Penalized M-estimator in the high-dimensional regime

In this section, we investigate the tractability and the robustness of the penalized M-estimator in the high-dimension region where the dimension of parameter p is much greater than the number of samples n. Specifically, we consider the same data generation model $y_i = \langle \theta_0, x_i \rangle + \epsilon_i$, where $y_i \in \mathbb{R}, x_i \in \mathbb{R}^p$, and the noise term ϵ_i are from Huber's gross error model (Huber, 1964): $\epsilon_i \sim (1-\delta)f_0 + \delta g$. Moreover, we assume $p \gg n$ and the true parameter θ_0 is sparse.

We consider the ℓ_1-regularized M-estimation under a ℓ_2-constraint on θ:

$$\begin{aligned} &\underset{\theta}{\text{Minimize:}} \quad \hat{L}_n(\theta) := \frac{1}{n}\sum_{i=1}^{n} \rho(y_i - \langle \theta, x_i \rangle) + \lambda_n ||\theta||_1, \\ &\text{subject to:} \quad \|\theta\|_2 \leq r. \end{aligned} \tag{3.10}$$

Before presenting our main theorem, we need additional assumptions on the feature vector x.

Assumption 3.3.1. *The feature vector x has a probability density function in $\mathbb{R}^p$. In addition, there exists constant $M > 1$ that is independent of dimension p such that $||x||_\infty \leq M\tau$ almost sure.*

The following lemma shows the uniform convergence of gradient and Hessian under the Huber's contamination model in the high-dimensional setting where $p >> n$.

Lemma 3.3.1. *Under assumption 3.2.1 and 3.3.1, there exist constants C_1, C_2 that depend on $r, \tau, \pi, \delta, L_\psi$, such that the following hold:*

a *The sample directional gradient converges uniformly to the population directional gradient, along the direction $(\theta - \theta_0)$.*

$$\begin{aligned} &\mathbf{P}\left(\sup_{\theta \in B_2^p(r)\setminus\{0\}} \frac{|\langle \nabla R_n(\theta) - \nabla R(\theta), \theta - \theta_0 \rangle|}{||\theta - \theta_0||_1} \leq (T_0 + L_0\tau)\sqrt{\frac{C_1 \log(np)}{n}} \right) \\ &\geq 1 - \pi. \end{aligned} \tag{3.11}$$

b *As $n \geq C_2 s_0 \log(np)$, we have*

$$\begin{aligned} &\mathbf{P}\left(\sup_{\theta \in B_2^p(r) \cap B_2^p(s_0), \nu \in B_2^p(1) \cap B_0^p(s_0)} |\langle \nu, \left(\nabla^2 R_n(\theta) - \nabla^2 R(\theta) \right) \nu \rangle| \leq \tau^2 \sqrt{\frac{C_2 s_0 \log(np)}{n}} \right) \\ &\geq 1 - \pi. \end{aligned}$$

Now we are ready for our main theorem.

Theorem 3.3.1. *Assume that Assumption 3.2.1 and Assumption 3.3.1 hold and the true parameter θ_0 satisfies $||\theta_0||_2 \leq r/3$ and $||\theta_0||_0 \leq s_0$. Then there exist constants C, C_0, C_1, C_2 that are dependent on $(\rho, L_\psi, \tau^2, r, \gamma, \pi)$ but independent on (δ, s_0, n, p, M) such that as $n \geq C s_0 \log p$ and $\lambda_n = C_0 M \sqrt{\frac{\log p}{n}} + \frac{C_1}{\sqrt{s_0}}\delta$, the following hold with probability as least $1 - \pi$:*

(a) *All stationary points of problem (3.10) are in $B_2^p(\theta_0, \eta_0 + \frac{\sqrt{s_0}}{1-\delta}\lambda_n C_2)$*

(b) *As long as n is large enough such that $n \geq C s_0 \log^2 p$ and the contamination ratio δ is small such that $(\eta_0 + \frac{1}{1-\delta}\sqrt{s_0}\lambda_n C_2) \leq \eta_1$, the problem (3.10) has a unique local stationary point which is also the global minimizer.*

The proof of Theorem 3.3.1 is based on several lemmas, which are postponed to the appendix. We believe that some of our lemmas are of interest in their own right. Theorem 3.3.1 implies the estimation error of the penalized M-estimator is bounded as the order of $O(\delta + \sqrt{\frac{s_0 \log p}{n}})$, which achieves the minimax estimation rate (Chen, Gao, and Ren, 2016). Moreover, it implies that the penalized M-estimator has good tractability when the percentage of outliers δ is small.

3.4 Example

In this section, we use some examples to illustrate our general theoretical results about the robustness and tractability of M-estimators. In the first subsection, we consider the low-dimensional regime and study a family of M-estimators with a specific loss function known as Huber's loss (Huber, 1964). In the second subsection, we consider the high-dimensional regime and study the penalized M-estimator with Welsch's exponential squared loss (Dennis Jr and Welsch, 1978; Rey, 2012; Wang, Jiang, Huang, and Zhang, 2013). In both subsections, we will derive the explicit expression of the two critical radius η_0, η_1 and discuss the robustness and tractability of the corresponding M-estimators.

3.4.1 M-estimator via Huber's loss

In this subsection, we illustrate the general results presented in Section 3.2 by studying the Huber's loss function (Huber, 1964)

$$\rho_\alpha(t) = \begin{cases} \frac{1}{2}t^2, & \text{if } |t| \le \alpha \\ \alpha(|t| - \alpha/2), & \text{if } |t| > \alpha. \end{cases} \tag{3.12}$$

where $\alpha > 0$ is a tuning parameter. The corresponding M-estimator is obtained by solving the optimization problem

$$\begin{aligned} \min_{\theta} \quad & \hat{R}_n(\theta) := \frac{1}{n}\sum_{i=1}^{n} \rho_\alpha(y_i - \langle \theta, x_i \rangle), \\ \text{subject to} \quad & ||\theta||_2 \le r. \end{aligned} \tag{3.13}$$

First, note the loss function $\rho_\alpha(t)$ in (3.12) is convex. Thus, the corresponding M-estimator should be tractable even though there are some outliers. Second, when α goes to 0, $\rho_\alpha(t)$ will converges to $t^2/2$. Thus, the least square estimator is a special case of the M-estimator obtained from (3.13), which is not robust to outliers. Third, for fixed $\alpha > 0$, $\rho'_\alpha(t), \rho''_\alpha(t)$ are all bounded. Intuitively, this implies that the impact of outlier observations of y_i will be controlled and thus the corresponding statistical procedure will be robust.

We now study the robustness and tractability of the M-estimator of (3.13) based on our framework in Theorem 3.2.2. In order to emphasize on the effects of the tuning parameter α and the contamination ratio δ on the robustness property and tractability property, we consider a simplified assumption on the feature vector x_i and the pdf of idealized residual f_0.

Assumption 3.4.1. **(a)** *The feature vector x_i are i.i.d multivariate Gaussian distribution $N(0, \tau^2 I_{p\times p})$.*

(b) *The idealized noise pdf $f_0(\epsilon)$ has Gaussian distribution $N(0, \sigma^2)$.*

(c) *The true parameter* $||\theta_0||_2 \leq r/3$.

Corollary 3.4.1. *Under Assumption 3.4.1, for any* $\delta, \alpha \geq 0$, *there exist two constants* $\eta_0(\delta, \alpha), \eta_1(\delta, \alpha)$:

$$\eta_0(\delta, \alpha) = \frac{\delta}{1-\delta} \frac{4\sqrt{2\pi}\sigma^3}{(\alpha^2 + 3\sigma^2)\tau} e^{\frac{\alpha^2 + 22\tau^2 r^2}{2\sigma^2}} \tag{3.14}$$

$$\eta_1(\delta, \alpha) = +\infty, \tag{3.15}$$

such that when the number of data points n *is large, with high probability, any stationary points of the empirical risk function* $\hat{R}_n(\theta)$ *in (3.13) belongs in the ball* $B_2^p(\theta_0, 2\eta_0(\delta, \alpha))$. *Moreover, the empirical risk function* $\hat{R}_n(\theta)$ *in (3.13) is strongly convex in the ball* $B_2^p(\theta_0, \eta_1(\delta, \alpha))$. *Thus, there exists a unique stationary point of* $\hat{R}_n(\theta)$, *which is the corresponding M-estimator.*

Note $\eta_1(\delta, \alpha) = \infty$, which means the corresponding Huber's estimator will be tractable, no matter there are outliers or not. This is consistent with the fact that the Huber's loss function is convex. Moreover, it is interesting to see the special case of Corollary 3.4.1 with $\alpha = +\infty$, which reduces to the least square estimator. As we can see, with $\delta > 0$, we have $\eta_0(\delta, \alpha = +\infty) = +\infty$, which implies the solution of the optimization problem in (3.13) can be arbitrarily in the ball $B_2^p(0, r = 10)$, even when the proportion of outliers is small. Thus it is not robust to the outliers. This recovers the well-known fact: the least square estimator is easy to compute, but is very sensitive to outliers.

Additionally, for another special case with $\delta = 0$ and $\alpha > 0$, we have $\eta_0(\delta = 0, \alpha) = 0$, which means the true parameter θ_0 is the unique stationary point of the risk function. This implies the Huber's estimator is consistent when there are no outliers.

3.4.2 Penalized M-estimator via Welsch's exponential squared loss

In this subsection, we illustrate the general results presented in Section 3.3 by considering a family of M-estimators with a specific nonconvex loss function known as Welsch's ex-

ponential squared loss (Dennis Jr and Welsch, 1978; Rey, 2012; Wang, Jiang, Huang, and Zhang, 2013),

$$\rho_\alpha(t) = \frac{1 - \exp(-\alpha t^2/2)}{\alpha}, \tag{3.16}$$

where $\alpha \geq 0$ is a tuning parameter. The corresponding penalized M-estimator is obtained by solving the optimization problem

$$\begin{aligned} \min_{\theta} \quad & \hat{L}_n(\theta) := \frac{1}{n}\sum_{i=1}^{n} \rho_\alpha(y_i - \langle \theta, x_i \rangle) + \lambda_n ||\theta||_1, \\ \text{subject to} \quad & ||\theta||_2 \leq r. \end{aligned} \tag{3.17}$$

The non-convex loss function $\rho_\alpha(t)$ in (3.16) has been used in other contexts such as robust estimation and robust hypothesis testing, see Ferrari and Yang (2010) and Qin and Priebe (2017), as it has many nice properties. First, it is a smooth function of both α and t, and the gradient and Hessian are well-defined. Second, when α goes to 0, $\rho_\alpha(t)$ will converges to $t^2/2$. Thus, the LASSO estimator is a special case of the M-estimator obtained from (3.17). Third, for fixed $\alpha > 0$, $\rho_\alpha(t), \rho'_\alpha(t), \rho''_\alpha(t)$ are all bounded. Intuitively, this implies that the impact of outlier observations of y_i will be controlled and thus the corresponding statistical procedure will be robust.

We now study the robustness and tractability of the penalized M-estimator of (3.17) based on our framework in Theorem 3.3.1. When α goes to 0, the M-estimator reduces to the LASSO estimator, which can be computed easily. However, it is also known to be very sensitive to the outliers. On the other hand, when α increases, the estimator becomes more robust, but may lose tractability due to the highly non-convexity of the function $\rho_\alpha(t)$ as well as the presence of the outliers.

In order to emphasize on the relation between the tuning parameter α and the contamination ratio δ, we consider a simplified assumption on the feature vector x_i and the pdf of idealized residual f_0.

Assumption 3.4.2. **(a)** *The feature vector x_i are i.i.d uniform distribution $[-\tau, \tau]^p$.*

(b) *The idealized noise pdf $f_0(\epsilon)$ has Gaussian distribution $N(0, \sigma^2)$.*

(c) *The true parameter $||\theta_0||_2 \leq 10/3$.*

With Assumption 3.4.2 and Theorem 3.3.1, we can get the following corollary, which characterizes the robustness and tractability of the penalized M-estimator with Welsch's exponential squared loss in (3.17):

Corollary 3.4.2. *Assume that Assumption 3.4.2 holds and the true parameter θ_0 satisfies $||\theta_0||_2 \leq r/3$, for any $\pi \in (0, 1)$, there exist a constant C_π such that if choose $\lambda_n = 2C_\pi\tau\sqrt{\frac{\log p}{n}} + \frac{\alpha\tau}{2}\frac{\delta}{\sqrt{s_0}}$, as $n >> s_0 \log p$, the following hold with probability as least $1 - \pi$:*

(a) *All stationary points of problem (3.17) are in $B_2^p(\theta_0, (1 + 2\tau)\eta_0)$*

(b) *The empirical risk function $\hat{L}_n(\theta)$ are strong convex in the ball $B_2^p(\theta_0, \eta_1)$*

(c) *As long as n is large enough and the contamination ratio δ is small such that $(1 + 2\tau)\,\eta_0 \leq \eta_1$, the problem (3.17) has a unique local stationary point which is also the global minimizer.*

Here

$$\eta_0(\delta, \alpha) = \frac{\delta}{1-\delta}\sqrt{\frac{e}{\alpha}}\frac{4(1+\alpha\sigma^2)^{3/2}}{\tau}e^{\frac{32\alpha r^2\tau^2}{3(1+\alpha\sigma^2)}} \tag{3.18}$$

$$\eta_1(\delta, \alpha) = \frac{1}{3\sqrt{3\alpha}(1+\alpha\sigma^2)^{3/2}\tau}\left[\tau^2 - \delta(\tau^2 + (1+\alpha\sigma^2)^{3/2})\right], \tag{3.19}$$

It is interesting to see the special case of Corollary 3.4.2 with $\alpha = 0$, which reduces to the LASSO estimator. On the one hand, with $\alpha = 0$, we have $\eta_1(\delta, \alpha = 0) = +\infty$ for any $\delta > 0$. This means that the corresponding risk function is strongly convex in the entire region of $B_2^p(0, r = 10)$, and hence it is always tractable. On the other hand, since $\eta_0(\delta, \alpha = 0) = +\infty$, the solution of the optimization problem in (3.17) can be arbitrarily

in the ball $B_2^p(0, r = 10)$, even when the proportion of outliers is small. Thus it is not robust to the outliers. This recovers the well-known fact: the LASSO estimator is easy to compute, but is very sensitive to outliers.

Additionally, for another special case with $\delta = 0$ and $\alpha > 0$, we have $\eta_0(\delta = 0, \alpha) = 0$, which means the true parameter θ_0 is the unique stationary point of the risk function. This implies the Welsch's estimator has nice tractability when there is no outliers. However, when the percentage of outlier δ is increasing, $\eta_1(\delta, \alpha)$ will decrease, which implies more outliers will reduce the tractability of the M-estimator.

3.5 Simulation results

In this section, we report the simulation results by using Welsch's exponential loss (Dennis Jr and Welsch, 1978) when the data are contaminated, using synthetic data setting. We first generate covariates $x_i \sim N(0, I_{p\times p})$ and responses $y_i = \langle \theta_0, x_i \rangle + \epsilon_i$, where $||\theta_0||_2 = 1$. We consider the case when the residual term ϵ_i have gross error model with contamination ratio δ, i.e., $\epsilon_i \sim (1-\delta)N(0,1) + \delta N(\mu_i, 3^2)$ where $\mu_i = ||x_i||_2^2 + 1$. The outlier distribution is chosen to highlight the effects of outliers when they are dependent on x_i and has non-zero mean.

In the first part, we consider the low-dimensional case when the dimension $p = 10$. Specifically, we generate $n = 200$ pairs of data $(y_i, x_i)_{i=1,..,n}$ with dimension $p = 10$ and with different choices of contamination ratios δ. We use projected gradient descent to solve the optimization problem in (3.13) with $r = 10$. In order to make the iteration points be inside the ball, we will project the points back into $B_2^p(0, r = 10)$ if they fall out of the ball. The step size is fixed as 1. In order to test the tractability of the M-estimator, we run gradient descent algorithm with 20 random initial values in the ball $B_2^p(0, r = 10)$ to see whether the gradient descent algorithm can converge to the same stationary point or not. Denote $\hat{\theta}(k)$ as the k^{th} iteration points, Figure 3.1 shows the convergence of the gradient descent algorithm for the exponential loss with the choice of $\alpha = 0.1$ under the gross error model

with different δ. From Figure 3.1 we observe when the proportion of outliers is small (i.e., $\delta \leq 0.1$,) gradient descent could converge to the same stationary point fast. However, when the contamination ratio δ becomes larger, gradient descent may not converge to the same point for different initial points, indicating the loss of tractability *for the same objective function* with increasing proportion of outliers. Those observations are consistent to our Theorem 3.2.2, which asserts the M-estimator is tractable when the contamination ratio δ is small.

To illustrate the robustness of the M-estimator, we generate 100 realizations of (Y, X) and run gradient descent algorithm with different initial values. The average estimation errors between the M-estimator and the true parameter θ_0 are presented in Figure 3.2. As we can see, when $\delta = 0$, all estimators have small estimation errors, which are well expected as those M-estimators are consistent without outliers (Huber, 1964; Huber and Ronchetti, 2009). However, for the M-estimator with $\alpha = 0$, i.e., the least square estimator, the estimation error will increase dramatically as the proportion of outliers increases. This confirms that the least square estimator is not robust to the outliers.

Meanwhile, when $\alpha = 0.1$, the overall estimation error does not increase much even with 40% outliers, which clearly demonstrate the robustness of the M-estimator. Note that when α is further increased from 0.1 to 0.3, although the estimator error is still very small for $\delta \leq 0.2$, it will increase dramatically when δ is greater than 0.2. We believe that two reasons contribute to this phenomenon: robustness starts to decrease when α becomes too large; and more importantly, the algorithm fails to find the global optimum due to multiple stationary points when α is large. Thus for each α, there exists a critical bound of δ, such that the estimator will be robust and tractable efficiently when the proportion of outliers is smaller than that bound.

In the second part, we present our results in the high-dimensional region when $p = 400$. Data (y_i, x_i) are generated from the same gross error model in the previous simulation study, with the true parameter θ_0 a sparse vector with 10 nonzero entries. All nonzero

entries are set to be $1/\sqrt{10}$. We use proximal gradient descent algorithm to solve problem (3.10). Similarly, we will project the points back into $B_2^p(0, r = 10)$ if they fall out of the ball. Figure 3.3 shows the convergence of the proximal gradient descent algorithm for the nonconvex exponential loss with the choice of $\alpha = 0.1$ and L_1 regularizer with the parameter $\lambda = 0.1$ under the gross error model with different δ. From Figure 3.3 we observe when the percentage of outliers is small, the algorithm will converge to the same stationary point fast, which implies there is only one unique stationary point. When δ is larger, the converge rate become slower, which implies there may exist another stationary points. Those simulation results reflect our theoretical result for the tractability of the penalized M-estimator in high-dimensional regression.

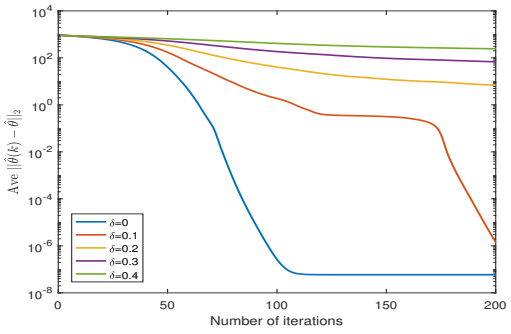

Figure 3.1: The convergence of gradient descent algorithm for different δ. Y-axis is with log scale.

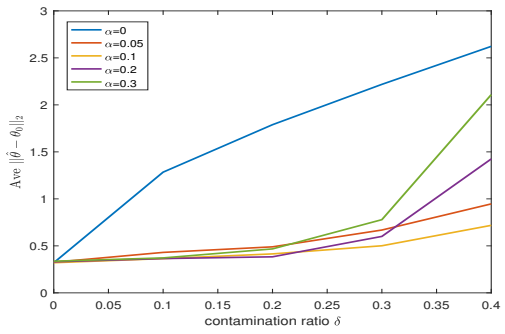

Figure 3.2: The estimation error for different α and δ

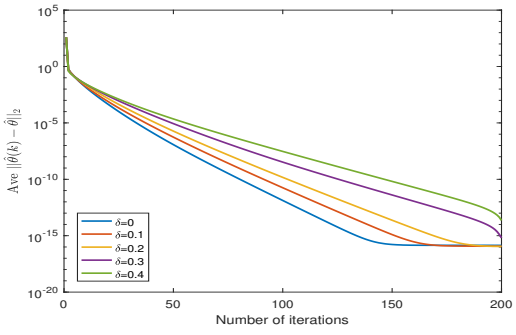

Figure 3.3: The convergence of gradient descent algorithm for different δ. Y-axis is with log scale.

3.6 Case study

In this section, we present a case study of the robust regression problem for the Airfoil Self-Noise dataset (Brooks, Pope, and Marcolini, 2014). The dataset was processed by NASA and is commonly used for regression study to learn the relation between the airfoil self-noise and five explanatory variables. Specifically, the dataset contain the following 5 explanatory variables: Frequency (in Hertzs), Angle of attack (in degrees), Chord length,(in meters), Free-stream velocity (in meters per second), and Suction side displacement thickness (in meters). There are 1503 observations in the dataset. The response variable is Scaled sound pressure level (in decibels). In this section, the five explanatory variables are scaled to have zero mean and unit variance. Then, we corrupt the response by adding noise ϵ from the same gross error model as the previous section: $\epsilon_i \sim (1-\delta)N(0,1)+\delta N(\mu_i, 3^2)$ with $\mu_i = ||x_i||_2^2 + 1$.

We consider the M-estimator using Welsch's exponential loss (Dennis Jr and Welsch, 1978) on the dataset to validate the tractability and the robustness of the corresponding M-estimator. First, we run 100 Monte Carlo simulations. At each time, we split the dataset which consists of 1503 pairs of data into a training dataset of size 1000 and a testing dataset of size 503. Then for the training dataset, we use gradient descent method with 20 different initial values to update the iteration points.

Figure 3.4 shows the average distance between each iteration point and the optimal point with the choice of $\alpha = 0.7$ and step size 0.5. Clearly, when δ is smaller than 0.3, gradient descent will converge to the same local minimizer, which implies the uniqueness of the stationary point. This result demonstrates the nice tractability of the M-estimator under the gross error model when the proportion of outliers is small. Then, using the optimal point as the M-estimator, we calculate the prediction error, which is the mean square error on the testing data. Figure 3.5 shows the average prediction error on the testing data. As we can see, the prediction error with the choice of $\alpha = 0$ will increase dramatically when

the percentage of outliers increases. In contrast, the prediction errors of M-estimators with $\alpha = 0.4$ is stable even with a large percentage of outliers. This illustrates the robustness of M-estimators for some positive α.

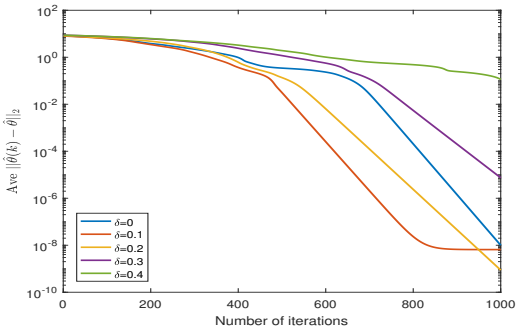

Figure 3.4: The convergence of gradient descent algorithm for different δ. Y-axis is with log scale.

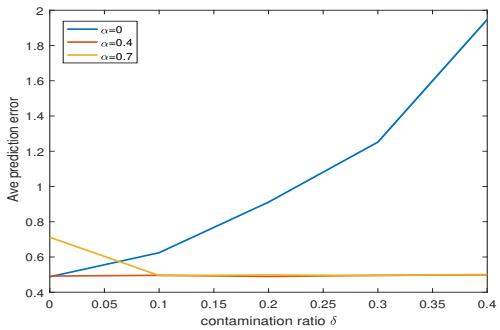

Figure 3.5: The prediction error for different α and δ

3.7 Proof

Proof of Lemma 3.2.1: In order to prove the uniform convergency theorem, it is suffice to verify assumption 1, 2 and 3 in Mei, Bai, and Montanari (2018). Specifically, first, we will verify that the directional gradient of the population risk is sub-Gaussian (Assumption 1 in Mei, Bai, and Montanari (2018)). Note the directional gradient of the population risk is given by $\langle \nabla \rho(Y - \langle X, \theta \rangle), \nu \rangle = \psi(Y - \langle X, \theta \rangle)\langle X, \nu \rangle$. Since $|\psi(Y - \langle X, \theta \rangle)| \leq L_\psi$, and $\langle X, \nu \rangle$ is mean zero and τ^2-sub-Gaussian by our assumption 1, due to Lemma 1 in Mei, Bai, and Montanari (2018), there exists a universal constant C_1, such that $\langle \nabla \rho(Y - \langle X, \theta \rangle), \nu \rangle$ is $C_1 L_\psi \tau^2-$sub-Gaussian. Second, we will verify that the directional Hessian of the loss is sub-exponential (Assumption 2 in Mei, Bai, and Montanari (2018)). The directional Hessian of the loss gives $\langle \nabla^2 \rho(Y - \langle X, \theta \rangle)\nu, \nu \rangle = \psi'(Y - \langle X, \theta \rangle)\langle X, \nu \rangle^2$. Since $|\psi'(Y - \langle X, \theta \rangle)| \leq L_\psi$, by Lemma 1 in Mei, Bai, and Montanari (2018), $\langle \nabla^2 \rho(Y - \langle X, \theta \rangle)\nu, \nu \rangle$ is $C_2\tau^2$-sub-exponential. Third, let $H = ||\nabla^2 R(\theta_0)||_{op}$ and $J^* = \mathbf{E}\left[\sup_{\theta_1 \neq \theta_2} \frac{||(\psi'(\theta_1) - \psi'(\theta_2))xx^T||_{op}}{||\theta_1 - \theta_2||_2}\right]$. Then, we can show $H \leq L_\psi \tau^2$ and $J^* \leq L_\psi (p\tau^2)^{3/2}$. Therefore, there exists a constant c_h such that $H \leq \tau^2 p^{c_h}$ and $J^* \leq \tau^3 p^{c_h}$, which verifies the assumption 3 in Mei, Bai, and Montanari (2018). Therefore, the uniform convergency of gradient and Hessian in theorem 1 in Mei, Bai, and Montanari (2018) holds for our gross error model. □

Proof of Theorem 3.2.1: Part (a): It is suffice to show that $\langle \theta - \theta_0, \nabla R(\theta) \rangle > 0$ for all $||\theta - \theta_0||_2 > \eta_0$. Note by Assumption 3.2.1(d), we have $h(z) = \int_{-\infty}^{+\infty} \psi(z + \epsilon) f_0(\epsilon) d\epsilon > 0$ as $z > 0$ and $h'(0) > 0$. Define $H(s) := \inf_{0 \leq z \leq s} \frac{h(z)}{z}$, it is easy to see that $H(s) > 0$ for all

$s > 0$. Then, we have

$$\begin{aligned}
\langle \theta - \theta_0, \nabla R(\theta) \rangle &= \mathbf{E}\left[\mathbf{E}[\psi(z+\epsilon)z | z = \langle \theta_0 - \theta, X\rangle]\right] \\
&= (1-\delta)\mathbf{E}[h(\langle \theta - \theta_0, X\rangle)\langle \theta - \theta_0, X\rangle] + \delta \mathbf{E}\left[\mathbf{E}_g(\psi(z+\epsilon)z | z = \langle \theta_0 - \theta, X\rangle)\right] \\
&\geq (1-\delta)H(s)\mathbf{E}[\langle \theta - \theta_0, X\rangle^2 I_{(|\langle \theta - \theta_0, X\rangle| \leq s)}] - \delta L_\psi \mathbf{E}|\langle \theta_0 - \theta, X\rangle| \\
&= (1-\delta)H(s)\mathbf{E}[\langle \theta - \theta_0, X\rangle^2 - \langle \theta - \theta_0, X\rangle^2 I_{(|\langle \theta - \theta_0, X\rangle| > s)}] - \delta L_\psi \mathbf{E}|\langle \theta - \theta_0, X\rangle| \\
&\geq (1-\delta)H(s)\left[\mathbf{E}[\langle \theta - \theta_0, X\rangle^2] - \left(\mathbf{E}[\langle \theta - \theta_0, X\rangle^4] \cdot \mathbf{P}(|\langle \theta - \theta_0, X\rangle| > s)\right)^{1/2}\right] \\
&\quad -\delta L_\psi (\mathbf{E}|\langle \theta - \theta_0, X\rangle|^2)^{1/2} \\
&\overset{(i)}{\geq} (1-\delta)H(s)||\theta - \theta_0||_2^2 \tau^2 \left(\gamma - \sqrt{c_2 \mathbf{P}(|\langle \theta - \theta_0, X\rangle| > s)}\right) - \delta L_\psi ||\theta - \theta_0||_2 \tau \\
&\overset{(ii)}{\geq} (1-\delta)H(s)||\theta - \theta_0||_2^2 \tau^2 \left(\gamma - \sqrt{\frac{c_2 \mathbf{E}(|\langle \theta - \theta_0, X\rangle|^4)}{s^4}}\right) - \delta L_\psi ||\theta - \theta_0||_2 \tau \\
&\geq (1-\delta)H(s)||\theta - \theta_0||_2^2 \tau^2 \left(\gamma - \sqrt{\frac{c_2 \cdot c_2 \tau^4 ||\theta - \theta_0||_2^4}{s^4}}\right) - \delta L_\psi ||\theta - \theta_0||_2 \tau \\
&\geq (1-\delta)H(s)||\theta - \theta_0||_2^2 \tau^2 \left(\gamma - \frac{c_2 \tau^2 ||\theta - \theta_0||_2^2}{s^2}\right) - \delta L_\psi ||\theta - \theta_0||_2 \tau \\
&\geq (1-\delta)H(s)||\theta - \theta_0||_2^2 \tau^2 \left(\gamma - \frac{16 c_2 \tau^2 r^2}{9 s^2}\right) - \delta L_\psi ||\theta - \theta_0||_2 \tau.
\end{aligned}$$

Here (i) holds from the fact that if X has mean zero and is τ^2-sub-Gaussian, then for all $u \in \mathbb{R}^p$,

$$\begin{aligned}
\mathbf{E}|\langle u, X\rangle|^2 &\leq ||u||_2^2 \tau^2, \\
\mathbf{E}|\langle u, X\rangle|^4 &\leq c_2 ||u||_2^4 \tau^4,
\end{aligned}$$

where c_2 is a constant (Boucheron, Lugosi, and Massart, 2013). (ii) holds from Chebyshev's inequality. Thus, a choice of $\tilde{s} = \frac{8\tau r}{3}\sqrt{\frac{c_2}{\gamma}}$ will ensure that

$$\langle \theta - \theta_0, \nabla R(\theta) \rangle \geq (1-\delta)\frac{3}{4}H(\frac{8\tau r}{3}\sqrt{\frac{c_2}{\gamma}})||\theta - \theta_0||_2^2 \tau^2 \gamma - \delta L_\psi ||\theta - \theta_0||_2 \tau, \qquad (3.20)$$

which is greater than 0 when

$$||\theta - \theta_0||_2 > \frac{\delta L_\psi}{(1-\delta)\frac{3}{4}H(\frac{8\tau r}{3}\sqrt{\frac{c_2}{\gamma}})\tau\gamma} := \eta_0. \tag{3.21}$$

Therefore, there are no stationary point outside of the ball $B_2^p(\theta_0, \eta_0)$.

Part(b): We first look at the minimum eigenvalue of the Hessian $\nabla^2 R(\theta)$ at $\theta = \theta_0$. For any $u \in \mathbb{R}^p, ||u||_2 = 1$,

$$\begin{aligned}
\langle u, \nabla^2 R(\theta_0) u\rangle &= (1-\delta)\mathbf{E}_{f_0}[\psi'(\epsilon)\langle X, u\rangle^2] + \delta\mathbf{E}_g[\psi'(\epsilon)\langle X, u\rangle^2] \\
&= (1-\delta)\mathbf{E}_{f_0}[\psi'(\epsilon)]\mathbf{E}[\langle X, u\rangle^2] + \delta\mathbf{E}_g[\psi'(\epsilon)\langle X, u\rangle^2] \\
&\geq (1-\delta)h'(0)\gamma\tau^2 - \delta L_\psi \tau^2.
\end{aligned}$$

Therefore, we have the minimum eigenvalue of $\nabla^2 R(\theta_0)$ is greater than 0 as long as $\delta < \frac{h'(0)\gamma}{h'(0)\gamma + L_\psi}$.

Then we look at the operator norm of $\nabla^2 R(\theta) - \nabla^2 R(\theta_0)$. For any $u \in \mathbb{R}^p, ||u||_2 = 1$,

$$\begin{aligned}
|\langle u, (\nabla^2 R(\theta) - \nabla^2 R(\theta_0))u\rangle| &= |\mathbf{E}[(\psi'(\langle X, \theta_0 - \theta\rangle + \epsilon) - \psi'(\epsilon))\langle X, u\rangle^2]| \\
&= |\mathbf{E}[\psi''(\xi)\langle X, \theta_0 - \theta\rangle\langle X, u\rangle^2]| \\
&\leq \mathbf{E}|\psi''(\xi)|\mathbf{E}|\langle X, \theta_0 - \theta\rangle\langle X, u\rangle^2| \\
&\leq L_\psi\{\mathbf{E}[\langle X, \theta_0 - \theta\rangle^2]\mathbf{E}[\langle X, u\rangle^4]\}^{1/2} \\
&\leq L_\psi(||\theta_0 - \theta||_2^2\tau^2 c_2\tau^4)^{1/2} \\
&= L_\psi\sqrt{c_2}||\theta_0 - \theta||_2\tau^3.
\end{aligned}$$

Hence, taking

$$||\theta - \theta_0||_2 \leq \eta_1 := \frac{(1-\delta)h'(0)\gamma - \delta L_\psi}{2\sqrt{c_2}\tau L_\psi} \tag{3.22}$$

guarantees that $(\nabla^2 R(\theta) - \nabla^2 R(\theta_0))_{op} \leq \frac{(1-\delta)h'(0)\gamma\tau^2 - \delta L_\psi \tau^2}{2}$. Therefore, for all $\theta \in B_2^p(\theta_0, \eta_1)$, we have

$$\lambda_{\min}(\nabla^2 R(\theta)) \geq \kappa := \frac{(1-\delta)h'(0)\gamma - \delta L_\psi}{2}\tau^2, \tag{3.23}$$

which yields there is at most one minimizer of $R(\theta)$ in the ball $B_2^p(\theta_0, \eta_1)$, as long as $\delta < \frac{h'(0)\gamma}{h'(0)\gamma + L_\psi}$.

Part (c): Note $R(\theta)$ is a continuous function on $B_2^p(r)$. Thus there exists a global minimizer, denoted by θ^*. Since we have shown that there is no stationary points outside the ball $B_2^p(\theta_0, \eta_0)$, θ^* should be in the ball $B_2^p(\theta_0, \eta_0)$. Therefore, as long as $\eta_1 > \eta_0$, i.e.,

$$\frac{(1-\delta)h'(0)\gamma - \delta L_\psi}{2\sqrt{c_2}\tau L_\psi} > \frac{\delta L_\psi}{(1-\delta)\frac{3}{4}H(\frac{8\tau r}{3}\sqrt{\frac{c_2}{\gamma}})\tau\gamma}, \tag{3.24}$$

there exists and only exists a unique stationary point of $R(\theta)$, which is also the global optimum θ^*. □

Proof of Theorem 3.2.2 Based on Lemma 3.2.1, there exists a constant C such that when $n \geq Cp\log p$,

$$\mathbf{P}\left(\sup_{\theta \in B^p(0,r)} ||\nabla \hat{R}_n(\theta) - \nabla R(\theta)||_2 \leq \tau\delta L_\psi\right) \geq 1 - \pi \tag{3.25}$$

$$\mathbf{P}\left(\sup_{\theta \in B^p(0,r)} ||\nabla^2 \hat{R}_n(\theta) - \nabla^2 R(\theta)||_{op} \leq \kappa/2\right) \geq 1 - \pi. \tag{3.26}$$

Part (a): Note

$$\begin{aligned} \langle \theta - \theta_0, \nabla \widehat{R}_n(\theta)\rangle &\geq \langle \theta - \theta_0, \nabla R(\theta)\rangle - ||\nabla \hat{R}_n(\theta) - \nabla R(\theta)||_2 ||\theta - \theta_0||_2 \qquad (3.27) \\ &\geq (1-\delta)\frac{3}{4}H(\frac{8\tau r}{3}\sqrt{\frac{c_2}{\gamma}})||\theta - \theta_0||_2^2\tau^2\gamma - 2\tau\delta L_\psi ||\theta - \theta_0||_2 \end{aligned}$$

which is greater than 0 when

$$||\theta - \theta_0||_2 > \frac{2\delta L_\psi}{(1-\delta)\frac{3}{4}L(\frac{8\tau r}{3}\sqrt{\frac{c_2}{\gamma}})\tau\gamma} = 2\eta_0. \tag{3.28}$$

Therefore, there are no stationary points outside of the ball $B_2^p(\theta_0, 2\eta_0)$.

Part (b): For the least eigenvalue of the empirical Hessian in $B_2^p(\theta_0, \eta_1)$, we have

$$\begin{aligned} \inf_{||\theta-\theta_0||_2 \leq \eta_1} \lambda_{\min}(\nabla^2 \widehat{R}_n(\theta)) &\geq \inf_{||\theta-\theta_0||_2 \leq \eta_1} \lambda_{\min}(\nabla^2 R(\theta)) - \sup_{\theta \in B^p(0,\eta_1)} ||\nabla^2 \hat{R}_n(\theta) - \nabla^2 R(\theta)||_{op} \\ &\geq \kappa - \kappa/2 = \kappa/2 > 0. \end{aligned} \tag{3.29}$$

This lead to the conclusion that, $\widehat{R}_n(\theta)$ is strong convex inside the ball $B_2^p(\theta_0, \eta_1)$.

Part(c): When $2\eta_0 < \eta_1$, by strong convexity of $\widehat{R}_n(\theta)$ in $B_2^p(\theta_0, \eta_1)$, there exists a unique local minimizer, which is in $B_2^p(\theta_0, 2\eta_0)$. We denote the unique local minimizer as $\widehat{\theta}_n$.

By Theorem 3.2.1, there is a unique stationary point of the population risk function $R(\theta)$ in the ball $B_2^p(\theta_0, \eta_0)$. Suppose θ^* is the unique stationary point of $R(\theta)$. By Taylor expansion of $\widehat{R}_n(\theta)$ at the point θ^*, there exists a $\tilde{\theta}$ in $B^p(\theta_0, 2\eta_0)$, such that

$$\widehat{R}_n(\widehat{\theta}_n) = \widehat{R}_n(\theta^*) + \langle \widehat{\theta}_n - \theta^*, \nabla \widehat{R}_n(\theta^*) \rangle + \frac{1}{2}(\widehat{\theta}_n - \theta^*)' \nabla^2 \widehat{R}_n(\tilde{\theta})(\widehat{\theta}_n - \theta^*) \leq \widehat{R}_n(\theta^*) \tag{3.30}$$

Since by equation (3.29), the least eigenvalue of $\nabla^2 \widehat{R}_n(\tilde{\theta})$ is greater than $\kappa/2$, which lead to

$$\frac{\kappa}{4}||\widehat{\theta}_n - \theta^*||_2^2 \leq \langle \theta^* - \widehat{\theta}_n, \nabla \widehat{R}_n(\theta^*) \rangle \leq ||\theta^* - \widehat{\theta}_n||_2 ||\nabla \widehat{R}_n(\theta^*)||_2, \tag{3.31}$$

which yield

$$||\widehat{\theta}_n - \theta^*||_2 \leq \frac{4}{\kappa}||\nabla \widehat{R}_n(\theta^*)||_2. \tag{3.32}$$

By Theorem 3.2.1, $||\theta_0 - \theta^*||_2 < \eta_0$, combined with equation (3.32) and the uniform convergency theorem in Lemma 3.2.1 yield

$$||\widehat{\theta}_n - \theta_0||_2 \leq \eta_0 + \frac{4\tau}{\kappa}\sqrt{\frac{C * p \log n}{n}}. \tag{3.33}$$

□

Proof of lemma 3.3.1: From the Theorem 3 in Mei, Bai, and Montanari, 2018, the uniform convergency theorem of our Lemma 3.3.1 holds if Assumption 4, 5 in Mei, Bai, and Montanari, 2018 hold under the contaminated model with outliers. Here we will show under our assumption 3.2.1 and 3.3.1, there exist constants T_0 and L_0 such that

a For all $\theta \in B_2^p(r)$, $Y \in \mathbb{R}, X \in \mathbb{R}^p$, $||\nabla_\theta \rho(Y - \langle X, \theta \rangle)||_\infty \leq T_0 M$

b There exist functions $h_1 : \mathbb{R} \times \mathbb{R}^{p+1} \to \mathbb{R}$, and $h_2 : \mathbb{R}^{p+1} \to \mathbb{R}^p$, such that

$$\langle \nabla_\theta \rho(Y - \langle X, \theta \rangle), \theta - \theta_0 \rangle = h_1(\langle \theta - \theta_0, h_2(Y, X) \rangle), Y, X). \tag{3.34}$$

In addition, $h_1(t, Y, X)$ is $L_0 M$- Lipschitz to its first argument t, $h_1(0, Y, X) = 0$, and $h_2(Y, X)$ is mean-zero and τ^2-sub-Gaussian.

Part (a). The gradient of the loss is

$$\nabla_\theta \rho(Y - \langle X, \theta \rangle) = -\psi(Y - \langle X, \theta \rangle)X. \tag{3.35}$$

By assumption 3.2.1, we have $| - \psi(Y - \langle X, \theta \rangle)| \leq L_\psi$. By assumption 3.3.1, we have $||X||_\infty \leq M\tau$. Therefore, (a) is satisfied with parameter $T_0 = L_\psi \tau$.

Part (b). Note

$$\langle \nabla_\theta \rho(Y - \langle X, \theta \rangle), \theta - \theta_0 \rangle = -\psi(Y - \langle X, \theta \rangle)\langle X, \theta - \theta_0 \rangle. \tag{3.36}$$

We take $h_2(Y, X) = X$, $t = \langle X, \theta - \theta_0 \rangle$ and $h_1(t, Y, X) = -\psi(Y - t - \langle X, \theta_0 \rangle)t$. Clearly, we have $h_1(0, Y, X) = 0$ and $h_2(Y, X)$ is mean 0 and τ^2-sub-Gaussian. Furthermore, note $|t| \leq 2rM\tau$, we have

$$\begin{aligned} |\frac{\partial}{\partial t} h_1(t, Y, X)| &= |\psi'(Y - t - \langle X, \theta_0 \rangle)t - \psi(Y - t - \langle X, \theta_0 \rangle)| && (3.37) \\ &\leq 2ML_\psi r\tau + L_\psi && (3.38) \\ &\leq (2L_\psi r\tau + L_\psi)M. && (3.39) \end{aligned}$$

Therefore, $h_1(t, X, Y)$ is at most $(2L_\psi r\tau + L_\psi)M$-Lipschitz in its first argument t. By part (a) and part (b), we can see assumption 4, 5 are satisfied under the gross error model, which prove the uniform convergency theorem in our Lemma 3.3.1. □

Proof of theorem 3.3.1: We decompose the proof into four technical lemmas. First, in Lemma 3.7.1, we prove there cannot be any stationary points of the regularized empirical risk $\hat{L}_n$ in (3.10) outside the region $\mathbb{A}$, which is a cone with $\mathbb{A} = \{\theta_0 + \Delta : ||\Delta_{S_0^c}||_1 \leq 3||\Delta_{S_0}||_1\}$. Then in Lemma 3.7.2, we show there cannot be any stationary points outside the region $B_2^p(\theta_0, r_s)$ where r_s is the statistical radius which is not less than η_0 in Theorem 3.2.1. In Lemma 3.7.3, we argue that all stationary points should have support size less or equal to $cs_0 \log p$. Finally, in Lemma 3.7.4, we show there cannot be two stationary points in $B_2^p(\theta_0, \eta_1) \cap \mathbb{A}$. Note $\hat{L}_n(\theta)$ is a continuous function, which indicates the existence of the global minimizer. Therefore, we can conclude there is and only is one unique stationary point of the regularized empirical risk $\hat{L}_n$ as long as $r_s < \eta_1$.

To start with those lemmas, we define the subgradient of $\hat{L}_n$ at θ as:

$$\partial \hat{L}_n(\theta) = \{\nabla R_n(\theta) + \lambda_n \nu : \nu \in \partial ||\theta||_1\}. \tag{3.40}$$

Therefore, the optimality condition implies that θ is a stationary point of $\hat{L}_n$ if and only if $\mathbf{0} \in \partial \hat{L}_n(\theta)$. To simplify notations, all constants in the following lemmas are dependent on $(\rho, L_\psi, \tau^2, r, \gamma, \pi)$ but independent on δ, s_0, n, p, M.

Lemma 3.7.1. *Let $S_0 = supp(\theta_0)$ and $s_0 = |S_0|$. Define a cone $\mathbb{A} = \{\theta_0 + \Delta : ||\Delta_{S_0^c}||_1 \leq 3||\Delta_{S_0}||_1\} \subseteq \mathbb{R}^p$. For any $\pi > 0$, there exist constants C_0, C_1 such that letting $\lambda_n \geq C_0 M\sqrt{\frac{\log p}{n}} + \delta\frac{C_1}{\sqrt{s_0}}$, with probability at least $1-\pi$, $\hat{L}_n(\theta)$ has no stationary points in $B_2^p(0,r) \cap \mathbb{A}^c$:*

$$\langle z(\theta), \theta - \theta_0 \rangle > 0, \quad \forall \theta \in B_2^p(0,r) \cap \mathbb{A}^c, z(\theta) \in \partial \hat{L}_n(\theta) \tag{3.41}$$

Proof. For any $z(\theta) \in \partial \hat{L}_n(\theta)$, it can be written as $z(\theta) = \nabla \hat{R}_n(\theta) + \lambda_n \nu(\theta)$, where $\nu(\theta) \in \partial ||\theta||_1$. Therefore, we have

$$\langle z(\theta), \theta - \theta_0 \rangle = \langle \nabla R(\theta), \theta - \theta_0 \rangle + \langle \nabla \hat{R}_n(\theta) - \nabla R(\theta), \theta - \theta_0 \rangle + \lambda_n \langle \nu(\theta), \theta - \theta_0 \rangle \tag{3.42}$$

Note by (3.20) we have

$$\langle \theta - \theta_0, \nabla R(\theta) \rangle \geq (1-\delta)\frac{3}{4} H(\frac{8\tau r}{3}\sqrt{\frac{c_2}{\gamma}})||\theta - \theta_0||_2^2 \tau^2 \gamma - \delta L_\psi ||\theta - \theta_0||_2 \tau. \tag{3.43}$$

By lemma 3.3.1, for any $\pi > 0$, there exists a constant C_π such that

$$\mathbf{P}(\sup_{0<||\theta||_2<r} \frac{|\langle \nabla \hat{R}_n(\theta) - \nabla R(\theta), \theta - \theta_0 \rangle|}{||\theta - \theta_0||_1} \leq C_\pi M\sqrt{\frac{\log p}{n}}) > 1 - \pi. \tag{3.44}$$

Letting $\Delta = \theta - \theta_0$, we have

$$\langle \nu(\theta), \theta - \theta_0 \rangle = \langle \nu(\theta)_{S_0^c}, \Delta_{S_0^c} \rangle + \langle \nu(\theta)_{S_0}, \Delta_{S_0} \rangle \geq ||\Delta_{S_0^c}||_1 - ||\Delta_{S_0}||_1 \tag{3.45}$$

Plugging (3.43),(3.44),(3.45) into (4.5) yields

$$\begin{aligned}\langle z(\theta), \theta - \theta_0 \rangle &\geq (1-\delta)\frac{3}{4}H(\frac{8\tau r}{3}\sqrt{\frac{c_2}{\gamma}})||\theta-\theta_0||_2^2\tau^2\gamma - \delta L_\psi||\theta-\theta_0||_2\tau \quad (3.46)\\ &- C_\pi M\sqrt{\frac{\log p}{n}}(||\Delta_{S_0^c}||_1 + ||\Delta_{S_0}||_1) + \lambda_n(||\Delta_{S_0^c}||_1 - ||\Delta_{S_0}||_1) \quad (3.47)\end{aligned}$$

Let $\lambda_n \geq 2C_\pi M\sqrt{\frac{\log p}{n}} + C_2$, we have

$$\begin{aligned}\langle z(\theta), \theta - \theta_0 \rangle &\geq (1-\delta)\frac{3}{4}H(\frac{8\tau r}{3}\sqrt{\frac{c_2}{\gamma}})||\theta-\theta_0||_2^2\tau^2\gamma - \delta L_\psi||\theta-\theta_0||_2\tau\\ &+ C_\pi M\sqrt{\frac{\log p}{n}}(||\Delta_{S_0^c}||_1 - 3||\Delta_{S_0}||_1) + C_2(||\Delta_{S_0^c}||_1 - ||\Delta_{S_0}||_1) \quad (3.48)\end{aligned}$$

Next, we will find the lower bound of $||\Delta_{S_0^c}||_1 - ||\Delta_{S_0}||_1$ under the constraint of $||\Delta_{S_0^c}||_1 - 3||\Delta_{S_0}||_1 \geq 0$. Note by Cauchy inequality, we have

$$||\Delta||_2^2 \geq \frac{||\Delta_{S_0^c}||_1^2}{p-s_0} + \frac{||\Delta_{S_0}||_1^2}{s_0} \quad (3.49)$$

Therefore, under the constraint of $||\Delta_{S_0^c}||_1 - 3||\Delta_{S_0}||_1 \geq 0$, the minimal value of $||\Delta_{S_0^c}||_1 - ||\Delta_{S_0}||_1$ is obtained when $||\Delta_{S_0^c}||_1 - 3||\Delta_{S_0}||_1 = 0$ and $||\Delta||_2^2 = \frac{||\Delta_{S_0^c}||_1^2}{p-s_0} + \frac{||\Delta_{S_0}||_1^2}{s_0}$. By solving the two equations yield

$$\begin{aligned}||\Delta_{S_0^c}||_1 &= 3\sqrt{\frac{(p-s_0)s_0}{8s_0+p}}||\Delta||_2 \quad (3.50)\\ ||\Delta_{S_0}||_1 &= \sqrt{\frac{(p-s_0)s_0}{8s_0+p}}||\Delta||_2 \quad (3.51)\end{aligned}$$

and $||\Delta_{S_0^c}||_1 - ||\Delta_{S_0}||_1 \geq 2\sqrt{\frac{(p-s_0)s_0}{8s_0+p}}||\Delta||_2$. Combined with (3.48), setting $C_1 = \frac{L_\psi\tau}{2}$ and $C_2 = C_1\frac{\delta}{\sqrt{s_0}}$ yield $2\sqrt{\frac{(p-s_0)s_0}{8s_0+p}}C_2 \geq \delta L_\psi\tau$, which implies $\langle z(\theta), \theta-\theta_0\rangle > 0$, as long as $\theta \in \mathbb{A}^c$, i.e., $||\Delta_{S_0^c}||_1 - 3||\Delta_{S_0}||_1 > 0$. □

Lemma 3.7.2. *For any $\pi > 0$, $\theta \in \mathbb{A}$, $z(\theta) \in \partial\hat{L}_n(\theta)$, there exist constants C_0, C_1 such*

that with probability at least $1-\pi$,

$$\langle z(\theta), \theta - \theta_0 \rangle > 0 \tag{3.52}$$

as long as $||\theta - \theta_0||_2 > r_s$, *where*

$$r_s = \frac{\delta}{1-\delta} C_0 + \frac{4\sqrt{s_0}}{1-\delta}(M\sqrt{\frac{\log p}{n}} + \lambda_n)C_1. \tag{3.53}$$

Proof. Since for any $\theta \in \mathbb{A}$, we have $||\theta - \theta_0||_1 \leq 4\sqrt{s_0}||\theta - \theta_0||_2$. Combining with (4.5) yields

$$\begin{aligned}
\langle z(\theta), \theta - \theta_0 \rangle &\geq \langle \nabla R(\theta), \theta - \theta_0 \rangle - C_\pi M\sqrt{\frac{\log p}{n}}||\theta - \theta_0||_1 - \lambda_n||\theta - \theta_1||_1 \\
&\geq (1-\delta)\frac{3}{4}H(\frac{8\tau r}{3}\sqrt{\frac{c_2}{\gamma}})||\theta - \theta_0||_2^2\tau^2\gamma - \delta L_\psi||\theta - \theta_0||_2\tau \\
&\quad -(C_\pi M\sqrt{\frac{\log p}{n}} + \lambda_n)4\sqrt{s_0}||\theta - \theta_0||_2,
\end{aligned} \tag{3.54}$$

which is greater than 0 as long as

$$||\theta - \theta_0||_2 \geq \frac{\delta L_\psi + (C_\pi M\sqrt{\frac{\log p}{n}} + \lambda_n)4\sqrt{s_0}}{(1-\delta)\frac{3}{4}H(\frac{8\tau r}{3}\sqrt{\frac{c_2}{\gamma}})\tau\gamma} := r_s. \tag{3.55}$$

Taking $C_0 = \frac{L_\psi}{\frac{3}{4}H(\frac{8\tau r}{3}\sqrt{\frac{c_2}{\gamma}})\tau\gamma}$ and $C_1 = \frac{\max(1, C_\pi)}{\frac{3}{4}H(\frac{8\tau r}{3}\sqrt{\frac{c_2}{\gamma}})\tau\gamma}$ give the result of r_s in equation (3.53). □

Lemma 3.7.3. *If* $\delta \leq 1/2$, *for any* π, *there exist constants* C_0, C_1, C *such that letting* $\lambda_n \geq C_0 M\sqrt{(\log p)/n} + \delta C_1/\sqrt{s_0}$, *with probability at least* $(1-\pi)$, *any stationary points of* $\hat{L}_n(\theta)$ *in* $B_2^p(\theta_0, r_s) \cap \mathbb{A}$ *has support size* $|S(\hat{\theta})| \leq C s_0 \log p$.

Proof. Let $\hat{\theta} \in B_2^p(\theta_0, r_s) \cap \mathbb{A}$ be a stationary point of $\hat{L}_n(\theta)$ in (3.10). Then we have

$$\nabla R_n(\hat{\theta}) + \lambda_n \nu(\hat{\theta}) = 0, \tag{3.56}$$

where $\nu(\hat{\theta}) \in ||\hat{\theta}||_1$. Thus, we have

$$\left(\nabla R_n(\hat{\theta})\right)_j = \pm\lambda_n, \quad \forall j \in S(\hat{\theta}) \tag{3.57}$$

Note $|\psi(y_i - \langle x_i, \theta_0\rangle)| \leq L_\psi$ and $\langle x_i, e_j\rangle$ is τ^2-subgaussian with mean 0. Then there exists an absolute constant c_0 such that $\psi(y_i - \langle x_i, \theta_0\rangle)\langle x_i, e_j\rangle$ is $c_0 L_\psi^2 \tau^2$-subgaussian, see Lemma 1(d) in Mei, Bai, and Montanari, 2018. Thus we have $\frac{1}{n}\sum_{i=1}^n \psi(y_i - \langle x_i, \theta_0\rangle)\langle x_i, e_j\rangle$ is $c_0 L_\psi^2 \tau^2/n$-subgaussian with mean $\langle \nabla R(\theta_0), e_j\rangle$. Moreover, note $|\langle \nabla R(\theta_0), e_j\rangle| = |\delta \mathbf{E}_g \psi(y_i - \langle x_i, \theta_0\rangle)\langle x_i, e_j\rangle| \leq \delta L_\psi \mathbf{E}|\langle x_i, e_j\rangle| \leq \delta L_\psi \tau$, we have for any $t > 0$,

$$\begin{aligned} & \mathbf{P}(|\frac{1}{n}\sum_{i=1}^n \psi(y_i - \langle x_i, \theta_0\rangle)\langle x_i, e_j\rangle| \geq t + \delta L_\psi \tau) \\ \leq\ & \mathbf{P}(|\frac{1}{n}\sum_{i=1}^n \psi(y_i - \langle x_i, \theta_0\rangle)\langle x_i, e_j\rangle - \langle \nabla R(\theta_0), e_j\rangle| \geq t) \\ \leq\ & 2\exp(-\frac{t^2 n}{2c_0 L_\psi^2 \tau^2}). \end{aligned} \tag{3.58}$$

Thus, we can get

$$\begin{aligned} \mathbf{P}\left(||\nabla R_n(\theta_0)||_\infty > t + \delta L_\psi \tau\right) &\leq p \max_{1\leq j\leq p} \mathbf{P}\left(|\frac{1}{n}\sum_{i=1}^n \psi(y_i - \langle x_i, \theta_0\rangle)\langle x_i, e_j\rangle| > t + \delta L_\psi \tau\right) \\ &\leq 2p\exp(-\frac{t^2 n}{2c_0 L_\psi^2 \tau^2}). \end{aligned} \tag{3.59}$$

Thus, a choice of $t = L_\psi \tau \sqrt{\frac{2c_0(\log p + \log 6/\pi)}{n}}$ and $C = \sqrt{c_0 \log 6/\pi}$ will guarantee that

$$\mathbf{P}\left(||\nabla \hat{R}_n(\theta_0)||_\infty > L_\psi \tau (C\sqrt{\frac{\log p}{n}} + \delta)\right) \leq \pi/3 \tag{3.60}$$

Let $\lambda_n \geq 2L_\psi \tau (C\sqrt{\frac{\log p}{n}} + \delta)$, we have the event $(||\nabla R_n(\theta_0)||_\infty < \lambda_n/2)$ happens with the

probability at least $1 - \pi/3$. Under this event, combing with (3.57) yields

$$\lambda_n/2 \leq \left| \left(\nabla R_n(\theta_0) - \nabla R_n(\hat{\theta}) \right)_j \right|, \quad \forall j \in S(\hat{\theta}). \tag{3.61}$$

Squaring and summing over $j \in S(\hat{\theta})$, we have

$$\begin{aligned}
\lambda_n^2 |S(\hat{\theta})| &\leq 4 \left\| \left(\nabla \hat{R}_n(\theta_0) - \nabla \hat{R}_n(\hat{\theta}) \right)_{S(\hat{\theta})} \right\|_2^2 && (3.62) \\
&= 4 \left\| \left(\frac{1}{n} \sum_{i=1}^n \left(\psi(y_i - \langle \theta_0, x_i \rangle) - \psi(y_i - \langle \hat{\theta}, x_i \rangle) \right) x_i \right)_{S(\hat{\theta})} \right\|_2^2 && (3.63) \\
&= 4 \left\| \left(\frac{1}{n} \sum_{i=1}^n \left(\psi'(y_i - \langle \beta_i, x_i \rangle) \right) \langle \theta_0 - \hat{\theta}, x_i \rangle x_i \right)_{S(\hat{\theta})} \right\|_2^2 && (3.64) \\
&\leq 4L_\psi^2 \left\| \left(\frac{1}{n} \sum_{i=1}^n \langle \theta_0 - \hat{\theta}, x_i \rangle x_i \right)_{S(\hat{\theta})} \right\|_2^2 && (3.65)
\end{aligned}$$

where β_i are located on the line between θ_0 and $\hat{\theta}$ obtained by intermediate value theorem. Moreover, by Minkowski inequality and Cauchy-Schwarz inequality yield

$$\begin{aligned}
\left\| \left(\frac{1}{n} \sum_{i=1}^n \langle \theta_0 - \hat{\theta}, x_i \rangle x_i \right)_{S(\hat{\theta})} \right\|_2 &\leq \frac{1}{n} \sum_{i=1}^n |\langle \theta_0 - \hat{\theta}, x_i \rangle| \left\| (x_i)_{S(\hat{\theta})} \right\|_2 && (3.66) \\
&\leq \frac{1}{n} \left((\sum_{i=1}^n |\langle \theta_0 - \hat{\theta}, x_i \rangle|^2)(\sum_{i=1}^n \| (x_i)_{S(\hat{\theta})} \|_2^2) \right)^{1/2}
\end{aligned}$$

Due to the restricted smoothness property of the sub-Gaussian random variables Mei, Bai, and Montanari, 2018, there exists a constant c_1 depending on π such that with probability at least $1 - \pi/3$, as $n \geq c_1 s_0 \log p$, we have

$$\sup_{\theta \in \mathbb{A}} \frac{\frac{1}{n}(\sum_{i=1}^n |\langle \theta_0 - \theta, x_i \rangle|^2)}{||\theta - \theta_0||_2^2} \leq 3\tau^2. \tag{3.67}$$

Therefore, with probability at least $1 - \pi/3$, we have

$$\sup_{\theta \in \mathbb{A} \cap B^p(\theta_0, r_s)} \frac{1}{n}(\sum_{i=1}^{n} |\langle \theta_0 - \hat{\theta}, x_i \rangle|^2) \leq 3\tau^2 \sup_{\theta \in \mathbb{A} \cap B^p(\theta_0, r_s)} ||\theta - \theta_0||_2^2 \leq 3\tau^2 r_s^2. \tag{3.68}$$

Moreover, by Lemma 13 in Mei, Bai, and Montanari, 2018, for any π, there exists constant c_2 depending on π such that

$$\mathbf{P}(\frac{1}{n}\sum_{i=1}^{n} \| (x_i)_{S(\hat{\theta})} \|_2^2 > c_2 \tau^2 \log p) \leq \pi/3. \tag{3.69}$$

By (3.60,3.68,3.69), as well as (3.66), at least $1 - \pi$,

$$\begin{aligned} \lambda_n^2 |S(\hat{\theta})| &\leq 4L_\psi^2 3\tau^2 r_s^2 c_2 \tau^2 \log p & (3.70) \\ &= C r_s^2 \log p & (3.71) \end{aligned}$$

By equation (3.53) we have

$$r_s^2 \leq C_0(\frac{\delta}{1-\delta})^2 + \frac{s_0}{(1-\delta)^2}(M^2 \frac{\log p}{n} + \lambda_n^2) C_1 \tag{3.72}$$

Taking $\lambda_n \geq C_2 M \sqrt{(\log p)/n} + C_3 \delta / \sqrt{s_0}$ gives us

$$\begin{aligned} |S(\hat{\theta})| &\leq (C_4 \frac{s_0}{(1-\delta)^2} + s_0 C_5) \log p & (3.73) \\ &= C s_0 \log p & (3.74) \end{aligned}$$

□

Lemma 3.7.4. *For any positive constants C_0 and π, letting $r_0 = C_0 s_0 \log p$, there exist constant C_1 such that when $n \geq C_1 s_0 \log^2 p$,*

$$\mathbf{P}(\sup_{\theta \in B_2^p(\theta_0, r) \cap B_0^p(0, r_0)} \sup_{\nu \in B_2^p(0,1) \cap B_0^p(0, r_0)} \langle \nu, (\nabla^2 \hat{R}_n(\theta) - \nabla^2 R(\theta)) \nu \rangle \leq \kappa/2) \geq 1 - \pi. \tag{3.75}$$

Moreover, the regularized empirical risk $\hat{L}_n(\theta)$ in (3.10) cannot have two stationary points in the region $B_2^p(\theta_0, \eta_1) \cap B_0^p(0, r_0/2)$.

Proof. According to (3.23), we have

$$\inf_{\theta \in B_2^p(\theta_0, \eta_1)} \lambda_{\min}(\nabla^2 R(\theta)) \geq \kappa. \tag{3.76}$$

By lemma 3.3.1, there exists constant C such that when $n \geq C s_0 \log^2 p$,

$$\mathbf{P}\left(\inf_{\theta \in B_2^p(\theta_0, \eta_1) \cap B_0^p(0, r_0)} \inf_{\nu \in B_2^p(0,1) \cap B_0^p(0, r_0)} \langle \nu, (\nabla^2 \hat{R}_n(\theta)) \nu \rangle \geq \kappa/2 \right) \leq \pi. \tag{3.77}$$

Suppose θ_1, θ_2 are two distinct stationary points of $\hat{L}_n(\theta)$ in $B_2^p(\theta_0, \eta_1) \cap B_0^p(0, r_0/2)$. Define $u = \frac{\theta_2 - \theta_1}{||\theta_1 - \theta_2||_2}$. Since θ_1 and θ_2 are $r_0/2$-sparse, u is r_0 sparse, as well as $\theta_1 + tu$ for any $t \in \mathbb{R}$. Therefore,

$$\begin{aligned} \langle \nabla \hat{R}_n(\theta_2), u \rangle &= \langle \nabla \hat{R}_n(\theta_1), u \rangle + \int_0^{||\theta_1 - \theta_2||_2} \langle u, \nabla^2 \hat{R}_n(\theta_1 + tu) u \rangle dt \\ &\geq \langle \nabla \hat{R}_n(\theta_1), u \rangle + \frac{\kappa}{2} ||\theta_2 - \theta_1||_2. \end{aligned} \tag{3.78}$$

Note the regularization term $\lambda_n ||\theta||_1$ is convex, we have for any subgradients $\nu(\theta_1) \in \partial ||\theta_1||_1$, $\nu(\theta_2) \in \partial ||\theta_2||_1$,

$$\lambda_n \langle \nu(\theta_2), u \rangle \geq \lambda_n \langle \nu(\theta_1), u \rangle. \tag{3.79}$$

Adding (3.78) with (3.79) gives

$$\langle \nabla \hat{R}_n(\theta_2) + \lambda_n \nu(\theta_2), u \rangle \geq \langle \nabla \hat{R}_n(\theta_1) + \lambda_n \nu(\theta_1), u \rangle + \frac{\kappa}{2} ||\theta_2 - \theta_1||_2, \tag{3.80}$$

which is contradict with the assumption that θ_1 and θ_2 are two distinct stationary points of $\hat{L}_n(\theta)$. □

Proof of Theorem 3.3.1. Now we are ready to prove Theorem 3.3.1. By Lemma 3.7.1 and Lemma 3.7.2, as $n \geq Cs_0 \log p$, letting $\lambda_n \geq C_0 M \sqrt{\frac{\log p}{n}} + \delta \frac{C_1}{\sqrt{s_0}}$, all stationary points of $L_n(\theta)$ are in $B_2^p(\theta_0, r_s) \cap \mathbb{A} \cap B_0^p(C_1 s_0 \log p)$, where r_s is defined in (3.53), $\mathbb{A}$ is the cone defined in Lemma 3.7.1. This proves Theorem 3.3.1(a). Moreover, by Lemma 3.7.3, Lemma 3.7.4, as $n \geq C_2 s_0 \log^2 p$, $\hat{L}_n(\theta)$ cannot have two distinct stationary points in $B_2^p(\theta_0, \eta_1) \cap \mathbb{A} \cap B_0^p(C_1 s_0 \log p)$. Thus, as long as $\eta_1 \geq r_s$, there is only one unique stationary point of the regularized empirical risk function $\hat{L}_n(\theta)$, which is the corresponding regularized M-estimator of (3.10). This proves Theorem 3.3.1 (b).

Proof of Corollary 3.4.1: Huber's loss function is defined by

$$\rho_\alpha(t) = \begin{cases} \frac{1}{2}t^2, & \text{if } |t| \leq \alpha \\ \alpha(|t| - \alpha/2), & \text{if } |t| > \alpha. \end{cases} \tag{3.81}$$

the corresponding score function would be

$$\psi_\alpha(t) = \rho'_\alpha(t) = \begin{cases} t, & \text{if } |t| \leq \alpha \\ sign(t)\alpha, & \text{if } |t| > \alpha. \end{cases} \tag{3.82}$$

Note for any $\alpha > 0$, all of $\psi(t)$, $\psi'(t)$ and $\psi''(t)$ are bounded. Specifically, we have $|\psi_\alpha(t)| \leq \alpha$, $|\psi'(t)| = |\psi''(t)| = 0$. Therefore, the assumptions in Theorem 3.2.1 and Theorem 3.2.2 are satisfied. It is suffice to find the explicit expression of η_0 and η_1 in equation (3.21) and (3.22). Since $|\psi'(t)| = |\psi''(t)| = 0$, it is easy to see $\eta_1 = +\infty$, which implies the Huber's estimator has nice computational tractability, regardless the choice of tuning parameter α and the percentage of outliers δ. Moreover, to find the explicit expression of η_0, according to Assumption 3.4.1, we have $c_2 = 3, \gamma = 1$. Thus, we can calculate

$$
\begin{aligned}
h(z) &= \int_{-\infty}^{+\infty} \psi_\alpha(z+\epsilon) f_0(\epsilon) d\epsilon = \int_{-\infty}^{\infty} \psi_\alpha(t) f_0(t-z) dt \\
&= \int_0^\alpha t \left[f_0(t-z) - f_0(t+z) \right] dt + \alpha \int_\alpha^{+\infty} \left[f_0(t-z) - f_0(t+z) \right] dt \\
&\geq \int_0^\alpha t \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{t^2+z^2}{2\sigma^2}} \left(\frac{tz}{\sigma^2} \right) dt + \alpha \int_\alpha^{+\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{t^2+z^2}{2\sigma^2}} \left(\frac{tz}{\sigma^2} \right) dt \\
&\geq \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{\alpha^2+z^2}{2\sigma^2}} \int_0^\alpha t \left(\frac{tz}{\sigma^2} \right) dt + \frac{z\alpha}{\sigma^2} e^{-\frac{z^2}{2\sigma^2}} \int_\alpha^{+\infty} t \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{t^2}{2\sigma^2}} dt \\
&= \frac{z\alpha^3}{3\sqrt{2\pi}\sigma^3} e^{-\frac{z^2+\alpha^2}{2\sigma^2}} + \frac{z\alpha}{\sqrt{2\pi}\sigma} e^{-\frac{z^2+\alpha^2}{2\sigma^2}}
\end{aligned}
$$

Therefore we have $H(s) = (\frac{\alpha^3}{3\sqrt{2\pi}\sigma^3} + \frac{\alpha}{\sqrt{2\pi}\sigma}) e^{-\frac{s^2+\alpha^2}{2\sigma^2}}$. By equation (3.21) in the proof of Theorem 3.2.1 yields

$$
\begin{aligned}
\eta_0(\delta, \alpha) &= \frac{\delta L_\psi}{(1-\delta)\frac{3}{4} H(\frac{8\tau r}{3}\sqrt{\frac{c_2}{\gamma}})\tau\gamma} && (3.83) \\
&= \frac{\delta}{1-\delta} \frac{4\sqrt{2\pi}\sigma^3}{(\alpha^2+3\sigma^2)\tau} e^{\frac{\alpha^2+22\tau^2 r^2}{2\sigma^2}}, && (3.84)
\end{aligned}
$$

which complete the proof. □

Proof of Corollary 3.4.2: When the loss function is defined by $\rho_\alpha(t) = \frac{1-e^{-\alpha t^2/2}}{\alpha}$, the corresponding score function would be $\psi_\alpha(t) = \rho'_\alpha(t) = te^{-\alpha t^2/2}$. Moreover, we can get $\psi'_\alpha(t) = e^{-\alpha t^2/2}(1-\alpha t^2)$ and $\psi''_\alpha(t) = e^{-\alpha t^2/2}\alpha(\alpha t^2 - 3)$. Note for any $\alpha > 0$, all of $\psi_\alpha(t)$, $\psi'_\alpha(t)$ and $\psi''_\alpha(t)$ are bounded.

$$
\begin{aligned}
|\psi_\alpha(t)| &\leq \sqrt{\frac{e}{\alpha}} \\
|\psi'_\alpha(t)| &\leq \max\{1, 2e^{-1.5}\} = 1 \\
|\psi''_\alpha(t)| &\leq \max\{e^{-(3+\sqrt{6})/2}\sqrt{(18+6\sqrt{6})\alpha}, e^{-(3-\sqrt{6})/2}\sqrt{(18-6\sqrt{6})\alpha}\} \leq 1.5\sqrt{\alpha}.
\end{aligned}
$$

Therefore, the Assumption 3.2.1 is satisfied. It is suffice to find the explicit expression

of η_0 and η_1 in equation (3.21) and (3.22). In order to have an accurate expression, we will use the individual bound of $\psi_\alpha(t), \psi'_\alpha(t), \psi''_\alpha(t)$ instead of the universal bound L_ψ. Specifically, according to Assumption 3.4.2, x_i is τ^2-sub-Gaussian, $c_2 = 3, \gamma = 1/3$. Thus, we can calculate $h(z) = \int_{-\infty}^{+\infty} \psi_\alpha(z+\epsilon) f_0(\epsilon) d\epsilon = \frac{z}{(1+\alpha\sigma^2)^{3/2}} e^{-\frac{\alpha z^2}{2(1+\alpha\sigma^2)}}$ and $H(s) = \frac{1}{(1+\alpha\sigma^2)^{3/2}} e^{-\frac{\alpha s^2}{2(1+\alpha\sigma^2)}}$. By equation (3.21) in the proof of Theorem 3.2.1 yields

$$\begin{aligned} \eta_0(\delta, \alpha) &= \frac{\delta L_\psi}{(1-\delta)\frac{3}{4} H(\frac{8\tau r}{3}\sqrt{\frac{c_2}{\gamma}})\tau\gamma} && (3.85) \\ &= \frac{\delta}{1-\delta}\sqrt{\frac{e}{\alpha}} \frac{4(1+\alpha\sigma^2)^{3/2}}{\tau} e^{\frac{32\alpha r^2 \tau^2}{3(1+\alpha\sigma^2)}} && (3.86) \end{aligned}$$

Similarly, we can calculate $h'(0) = E_{f_0}\psi'_\alpha(\epsilon) = \frac{1}{(1+\alpha\sigma^2)^{3/2}}$. Note $|\psi'_\alpha(t)| \leq 1, |\psi''_\alpha(t)| \leq 1.5\sqrt{\alpha}$, by equation (3.22) in the proof of Theorem 3.2.1 yields

$$\begin{aligned} \eta_1(\delta, \alpha) &= \frac{(1-\delta)h'(0)\tau^2 - \delta}{2\sqrt{3} \times 1.5\sqrt{\alpha}\tau} && (3.87) \\ &= \frac{1}{3\sqrt{3\alpha}(1+\alpha\sigma^2)^{3/2}\tau} \left[\tau^2 - \delta(\tau^2 + (1+\alpha\sigma^2)^{3/2})\right]. && (3.88) \end{aligned}$$

According to equation (3.55) in the proof of Theorem 3.3.1, we have with high probability, all stationary points of the empirical risk function $\hat{L}_n(\theta)$ in (3.17) are inside the ball $B_2^p(\theta_0, r_s)$, where

$$\begin{aligned} r_s &= \eta_0 + \frac{12 C_\pi \tau \sqrt{(s_0 \log p)/n} + 2\tau\delta L_\psi}{(1-\delta)\frac{3}{4} H(\frac{8\tau r}{3}\sqrt{\frac{c_2}{\gamma}})\tau\gamma} && (3.89) \\ &= (1+2\tau)\eta_0 + \frac{16 C_\pi \tau \sqrt{(s_0 \log p)/n}}{(1-\delta) H(\frac{8\tau r}{3}\sqrt{\frac{c_2}{\gamma}})\tau\gamma}. && (3.90) \end{aligned}$$

Therefore, as $n >> s_0 \log p$, we have $r_s \approx (1+2\tau)\eta_0$, which completes the proof. □

CHAPTER 4

APPLIED RESEARCH IN NONLINEAR PROFILE MONITORING

4.1 Introduction

With the rapid development of advanced sensing technologies, rich and complex real-time profile or curve data are available in many processes in biomedical sciences, manufacturing and engineering. For instance, physiologic monitoring systems generated real-time profile conditions of a patient in intensive care units. In modern manufacturing, profile data are generated to provide valuable information about the quality or reliability performance of the process or product. In these applications, it is often desirable to utilize the observed profile data to develop efficient methodologies for process monitoring and fault diagnosing.

A concrete motivating example of profile data in this chapter is from a progressive forming process with five die stations including preforming, blanking, initial forming, forming, and trimming, see Figure 4.1 for illustration. Ideally, when the process is in control, a work piece should pass through these five stations. However, a missing part problem, which means that the work piece is not settled in the right die station but is conveyed to the downstream stations, may occur in this process (Lei, Zhang, and Jin, 2010; Zhou, Liu, Zhang, Zhang, and Shi, 2016). Such a fault often leads to unfinished or nonconforming products and/or severe die damage. The tonnage signal measured by the press tonnage sensor, which is the summation of all stamping forces, contains rich process information of forming operations and widely used for monitoring the forming process. Figure 4.2 shows the tonnage profiles collected under normal condition and five faulty conditions corresponding to missing operations occurring in each of the five die stations. It is clear from the figure that each profile is highly nonlinear, since the observed forces at different segments correspond

[1]The materials in this chapter were published in *New Frontiers in Biostatistics and Bioinformatics*, 2018.

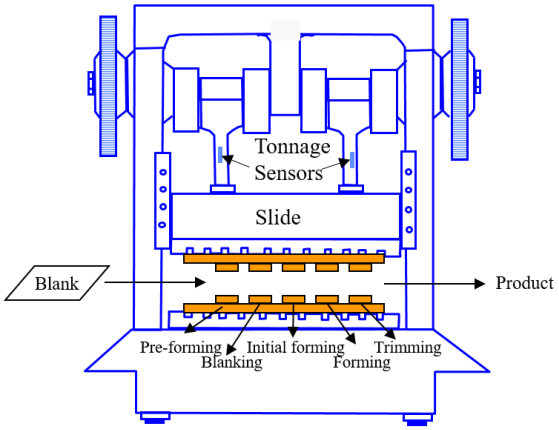

Figure 4.1: Illustration of a progressive forming process.

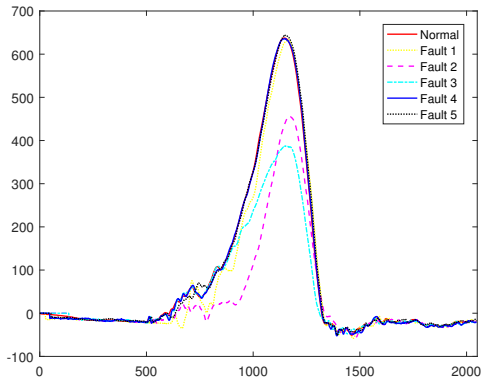

Figure 4.2: Six profile samples from a forming process: one is in-control, normal sample and the other five are out-of-control, fault samples.

to different stages of the operation within one production cycle. In addition, the difference between normal profiles and fault profiles are also nonlinear. For some particular faults, i.e., Fault 4, profiles are quite overlapping with the normal profiles. Under a high-production rate environment, it is highly desirable but challenging to effectively online monitor these profiles and detect those different types of unknown but subtle changes quickly.

In the profile monitoring literature, much research has been done for monitoring linear profiles, see, for example, Kang and Albin (2000), Chang and Gan (2006), Zou, Zhou, Wang, and Tsung (2007), Zou, Tsung, and Wang (2007), and Kazemzadeh, Noorossana, and Amiri (2008). However, in many real-world applications including those profiles in Figure 4.2, the form of the profile data are too complicated to be expressed as a linear or parametric function. Several nonlinear profile monitoring procedures have been developed in the literature based on nonparametric regression techniques such as smoothing splines (Gardner, Lu, Gyurcsik, Wortman, Hornung, Heinisch, Rying, Rao, Davis, and Mozumder, 1997; Chang and Yadama, 2010), Fourier analysis (Chen and Nembhard, 2011), local kernel regression (Qiu, Zou, and Wang, 2010; Zou, Qiu, and Hawkins, 2009) and functional principal components analysis (FPCA) (Hall, Poskitt, and Presnell, 2001; Paynabar, Zou,

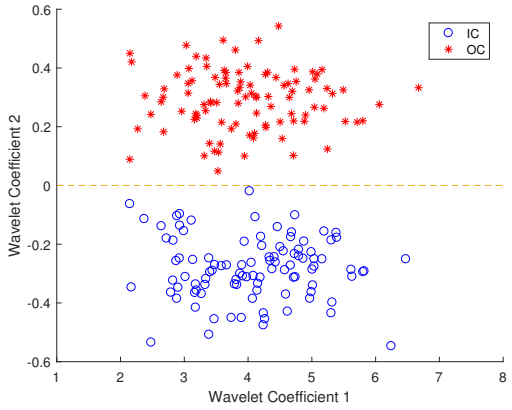

Figure 4.3: A simulated data set in the 2-dimensional wavelet domain, where blue circles indicate IC observations and red stars indicate OC observations. The mean shift is along the second wavelet coefficient, and the change is undetectable if using the first wavelet coefficient

and Qiu, 2016). However all these approaches tend to monitor a smooth, in-control profile, and thus may loss information about local structures such as jumps or cusps. Moreover, all these approaches are based on monitoring the changes of selected model coefficients, while it will be difficult to interpret their meanings back to the original profiles.

In this chapter, we propose to monitor nonlinear profiles based on the discrete wavelet transform (DWT). Besides a useful dimensional reduction tool, wavelet-based approaches have other advantages: the multi-resolution decomposition of the wavelets could be useful to locate the anomaly of the profile, and fast computational algorithms of the DWT are available (Mallat, 1989). Indeed, DWT has been applied to detect and diagnose process faults in the offline context, see Fan (1996) and Jin and Shi (1999). In the online monitoring context, many existing methods follow the suggestions of Donoho and Johnstone (1994) to first conduct wavelet shrinkage for dimension reduction under the in-control state, and then monitor the changes on the selected wavelet coefficients for the out-of-control state, see Hotelling T^2 control chart (Jeong, Lu, and Wang, 2006; Zhou, Sun, and Shi, 2006), and the CUSUM-type control chart (Lee, Hur, Kim, and Wilson, 2012). However, one will lose detection power if the change of the out-of-control state is on the wavelet coefficients that are not selected under the in-control state. To illustrate the importance of the out-of-control

state on the wavelet coefficients selection, we provide a simple two-dimensional example in Figure 4.3. As can be seen in this figure, the magnitude of wavelet coefficient 2 is very small compared with wavelet coefficient 1. However, if we just select wavelet coefficient 1 based on the in-control estimation, it would be difficult to detect the out-of-control samples since the changes occurred on the wavelet coefficient 2. To address this issue, it was proposed in Chicken, Pignatiello Jr, and Simpson (2009) to use all wavelet coefficients to conduct a likelihood ratio test. However, as we will show later in the simulation and case study, their methods are based on some asymptotic approximated likelihood ratio statistics, therefore may lose some detection power especially when the changed wavelet coefficients are sparse. Moreover, their method is not scalable and requires a lot of memory to store past observations.

In this chapter, we propose to first construct the local adaptive CUSUM statistics as in Lorden and Pollak (2008) and Liu, Zhang, and Mei (2019) for monitoring all wavelet coefficients by the hard-shrinkage estimation of the mean of in-control coefficients. Then we use the order-shrinkage to select those wavelet coefficients that are involved in the change significantly. Thus, from the methodology point of view, our proposed methodologies are analogous to those off-line statistical methods such as (adaptive) truncation, soft-, hard- and order-thresholding, see Neyman (1937), Donoho and Johnstone (1994), Fan and Lin (1998), and Kim and Akritas (2010). However, our motivation here is different and our application to profile monitoring is new.

The remainder of this chapter is as follows. In Section 4.2, we present problem formulation and background information of wavelet transform. In Section 4.3, we develop our proposed schemes for online nonlinear profile monitoring. In Section 4.4, a case study about monitoring tonnage signature is presented. In Section 4.5, a simulation study about monitoring the Mallet's piecewise smooth function is conducted.

4.2 Problem formulation and wavelet background

In this section, we will first present the mathematical formulation of the profile monitoring problem based on an additive change point model. Then we give a brief review of wavelet transformation that will be used for our proposed profile monitoring procedure.

Assume we obverse p-dimensional profile data, $y_1, y_2, \cdots$, sequentially from a process. Each profile y_k consists of p coordinates $y_k(x_i)$, for $i = 1, 2, ..., p$, with x_i equispaced over the interval $[0, 1]$, and can be thought of as the realization of a profile function $y_k(x)$. In the profile monitoring problem, we assume that the profile functions $y_k(x)$'s are from the additive change-point model:

$$y_k(x) = \begin{cases} f_0(x) + \epsilon_k(x), & \text{for } k = 1, 2, .., \nu \\ f_1(x) + \epsilon_k(x), & \text{for } k = \nu + 1, ... \end{cases} \tag{4.1}$$

where $f_0(\cdot)$ and $f_1(\cdot)$ are the mean functions that need be estimated from the data, and $\epsilon_k(x)$'s are the random noise which are assumed to be normally distributed with mean 0 that are independent across different time k. The problem is to utilize the observed profile data $y_k(x_i)$'s to detect the unknown change-time ν as quickly as possible when it occurs.

Since our proposed methods are based on monitoring the coefficients of the wavelet transformations of $y_k(x)$'s, let us provide a brief review of wavelet transformation of profile data. For any square-integrable function $f(x)$ on $\mathbb{R}$, it can be written as an (infinite) linear combinations of wavelet basis functions:

$$f(x) = \sum_{k \in \mathbb{Z}} c_{j_0}^k \phi_{j_0 k}(x) + \sum_{j=j_0}^{\infty} \sum_{k \in \mathbb{Z}} d_j^k \psi_{jk}(x). \tag{4.2}$$

Here the sets of two bases, $\phi_{jk}(x)$'s and $\psi_{jk}(x)$'s, are known as scaling and wavelet basis functions respectively, and are generated from two parent wavelets: one is the father wavelet $\phi(x)$ that characterizes basic wavelet scale, and the other is the mother wavelet

$\psi(x)$ that characterizes basic wavelet shape. Mathematically, $\phi_{jk}(x) = 2^{j/2}\phi(2^j x - k)$ and $\psi_{jk}(x) = 2^{j/2}\psi(2^j x - k)$, and the decomposed coefficients $c_{j_0}^k$ and d_j^k are called the scaling and detail coefficients, which represent the low-frequency and high-frequency components of original function $f(x)$.

When the observed data are discrete and dyadic, i.e., $\mathbf{y} = (y(x_1), y(x_2), ..., y(x_p))^T$ with p a dyadic integer, $p = 2^J$, discrete wavelet transform (DWT) can be used to determine the wavelet coefficients $\mathbf{c}$ fast and efficiently. The matrix form of DWT is represented as $\mathbf{c} = W\mathbf{y}$, where W is orthonormal wavelet transformation matrix (Mallat, 1999), which depends on the selected orthogonal wavelet basis. A large families of choices for wavelet basis functions are available for use, see for example Daubechies (1992). Also see Mallat (1999) for an efficient algorithm to implement DWT. In this chapter, the Haar transform is chosen as one way of DWT because Haar coefficients have an explicit interpretation of the changes in the profile observations. Also see Jin and Shi (2001) and Zhou, Sun, and Shi (2006) as examples of applying Haar transform to monitor profile samples.

For the observed p-dimension profile, $y = (y(x_1), ..., y(x_p))$, we consider the Haar transformation with wavelet basis functions:

$$\phi_{00}(x) = 1, x \in [0, 1] \tag{4.3}$$

$$\psi_{km}(x) = \begin{cases} 2^{\frac{k-1}{2}}, & \frac{m-1}{2^{k-1}} < x < \frac{m-1/2}{2^{k-1}} \\ -2^{\frac{k-1}{2}}, & \frac{m-1/2}{2^{k-1}} < x < \frac{m}{2^{k-1}} \\ 0, & \text{elsewhere} \end{cases} \tag{4.4}$$

where k represents the scale of Haar transform and $m = 1, 2, ..., 2^{k-1}$.

For simplicity, we assume $p = 2^J$ (otherwise we can add new extra zero coordinations to the original profile if needed). When Haar transform is chosen, the wavelet coefficients $\mathbf{c} = (c(1), c(2), ..., c(p))^T$ are often written as $(c_0^0, c_1^1, c_2^1, c_2^2, ..., c_J^1, ..., c_J^{2^{J-1}})^T$, which represent the Haar coefficients for different levels from 0 to J.

For any new observed p-dimension profile, $y = (y(x_1), ..., y(x_p))$, the explicit expres-

sion of these Haar coefficients are given by

$$
\begin{aligned}
c_0^0 &= 2^{-\frac{J}{2}} \sum_{\ell=1}^{2^J} y(x_\ell), \\
c_k^m &= 2^{\frac{J-k-1}{2}} \{ s[(m-1)2^{J-k+1}+1, (m-\frac{1}{2})2^{J-k+1}] - s[(m-\frac{1}{2})2^{J-k+1}+1, m2^{J-k+1}] \}, \\
&= 2^{-\frac{J-k+1}{2}} \{ \sum_{\ell=(m-1)2^{J-k+1}+1}^{(m-\frac{1}{2})2^{J-k+1}} y(x_\ell) - \sum_{\ell=(m-\frac{1}{2})2^{J-k+1}+1}^{m2^{J-k+1}} y(x_\ell) \} \qquad (4.5)
\end{aligned}
$$

for $k = 1, ..., J$; $m = 1, 2, ..., 2^{k-1}$ and $s[i, j]$ is defined by $s[i, j] = \frac{1}{j-i+1} \sum_{\ell=i}^{j} y(x_\ell)$. In other words, the Haar coefficient c_0^0 is proportional to the mean of all data and the other coefficients c_k^m are proportional to the mean difference of two adjacent intervals of length 2^{J-k}.

4.3 Our proposed method

At the high-level, our proposed profile monitoring method is based on monitoring the mean shifts on wavelet coefficients of nonlinear profiles $y_k(x)$'s. First, we use the in-control profiles from the historical training data to estimate the pre-change distributions of the wavelet coefficients. Second, we construct local monitoring statistics for each wavelet coefficient by recursively estimating the post-change mean of the wavelet coefficients. Third, we construct global monitoring procedure based on the information of the first several largest monitoring statistics.

It is necessary to emphasize that in the literature, wavelets are usually used for dimension reduction to select significant features and filter out noise Donoho and Johnstone (1994). Here our proposed method is constructing efficient monitoring statistic for each wavelet coefficients and then perform dimension reduction on the monitoring statistics. There are two technical challenges that need special attention. The first one is that we do not know which wavelet coefficients will be affected under the out-of-control state, and the second one is that we do not know what are the changed magnitudes or the post-

change distributions for those affected wavelet coefficients. To address these two challenges, we propose a computationally efficient algorithm that can monitor a large number of wavelet coefficients simultaneously in parallel based on local recursive CUSUM procedures, and then combine these local procedures together to raise a global alarm using the order-thresholding transformation in Liu, Zhang, and Mei (2019) to filter out those unaffected Haar coefficients. The recursive CUSUM procedure is to adaptively update the estimates of the post-change means, and it was first proposed in Lorden and Pollak (2008) for detecting a normal mean shift from 0 to some unknown, positive values. Here we extend it to the wavelet context when one wants to detect both positive and negative mean shifts of the wavelet coefficients.

For the purpose of demonstration, in the remaining of the chapter, we consider Haar coefficients as an example since they can easily be calculated and interpreted. Furthermore, they can capture the local changes on the profile efficiently.

For better presentation of our proposed nonlinear profile monitoring methods, we split this section into four subsections. Subsection 4.3.1 focuses on estimating the in-control means of Haar coefficients, and subsection 4.3.2 discusses how to recursively estimate possible mean shifts of Haar coefficients and constructs local monitoring statistics for each wavelet coefficient. Subsection 4.3.3 derives our proposed monitoring method and subsection 4.3.4 discusses how to choose tuning parameters.

4.3.1 In-control estimation

In our case study and in many real-world applications, it is reasonable to assume that some in-control profiles are available for learning the process variables. Without loss of generality, assume that there are m in-control profiles before online monitoring, and denote $\mathbf{c}_\ell$ as the vector of Haar coefficients of the ℓ^{th} profile $y_\ell(x)$ under the in-control status for $\ell = -m+1, \cdots, -1, 0$. If we denote $\mathbf{c}^{(ic)}$ as the mean vector of Haar coefficients under

the in-control state, then Haar coefficients under the in-control state are assumed as

$$\mathbf{c}_\ell = \mathbf{c}^{(ic)} + \mathbf{e}_\ell, \qquad \text{where} \qquad \mathbf{e}_\ell \sim N(\mathbf{0}, \Sigma_p). \tag{4.6}$$

for $\ell = -m+1, \cdots, -1, 0$. In other words, when there are no changes, the Haar coefficients $\mathbf{c}_\ell$ are i.i.d. multivariate normally distributed with in-control mean $\mathbf{c}^{(ic)}$ and diagonal covariance matrix $\Sigma_p = \text{diag}(\sigma_1^2, ..., \sigma_p^2)$.

It is well-known that the sample mean based on the in-control Haar coefficients $\mathbf{c}_\ell$ is not always a good estimator for $\mathbf{c}^{(ic)}$ when the dimension p is large (James and Stein, 1961). In the offline wavelet context, it is often assumed that the in-control p-dimensional mean vector of the Haar coefficients, $\mathbf{c}^{(ic)} = (c_1^{(ic)}, \cdots, c_p^{(ic)})$, has a sparsity structure and applying shrinkage techniques to filter out noise and obtain an accurate estimation (Donoho and Johnstone, 1994; Donoho and Johnstone, 1995). In this chapter, we follow the literature and apply hard shrinkage on the sample mean of in-control Haar coefficients. Specifically, let $\bar{\mathbf{c}}$ be the sample mean of m in-control Haar coefficient vectors, i.e.,

$$\bar{\mathbf{c}} = \frac{1}{m} \sum_{\ell=-m+1}^{0} \mathbf{c}_\ell.$$

Then the estimator of $\mathbf{c}^{(ic)} = (c_1^{(ic)}, \cdots, c_p^{(ic)})$ is

$$\hat{c}_i^{(ic)} = \begin{cases} \bar{c}_i^{(ic)}, & \text{if } |\bar{c}_i^{(ic)}| > \rho_1 \hat{\sigma}_i \\ 0, & \text{if } |\bar{c}_i^{(ic)}| \leq \rho_1 \hat{\sigma}_i \end{cases} \tag{4.7}$$

where $\hat{\sigma}_i$ is the sample standard deviation of the i-th Haar coefficient, and ρ_1 is a crucial tuning parameter to control the sparsity of the mean vector $\mathbf{c}^{(ic)}$. The choice of ρ_1 will be discussed in details later.

4.3.2 Out-of-control estimation and local statistics

In the profile monitoring context, the p-dimensional mean vector of the Haar coefficients is assumed to shift from the in-control value $\mathbf{c}^{(ic)}$ to an out-of-control value $\mathbf{c}^{(oc)} = (c_1^{(oc)}, \cdots, c_p^{(oc)})$. The difficulty is that one generally has limited knowledge about the out-of-control or fault samples in online profile monitoring, and thus one may not be able to accurately estimate the out-of-control mean $\mathbf{c}^{(oc)}$ even if we also put the sparsity constraints on $\mathbf{c}^{(oc)}$. For that reason, it makes more sense in online profile monitoring to assume that the difference vector $c^{(oc)} - c^{(ic)}$, instead of $c^{(oc)}$ itself, is sparse. To be more concrete, below we assume that only a few components of $\mathbf{c}^{(oc)} - \mathbf{c}^{(ic)}$ are non-zero, and $|c_i^{(oc)} - c_i^{(ic)}|/\sigma_i > \rho_2$ if the i-th component is affected, for some constant $\rho_2 > 0$, where σ_i is the standard deviation in (4.6).

Note that the change may affect those components with in-control value $c_i^{(ic)} = 0$, and thus one cannot simply monitor those non-zero components under the in-control state. Also, since we do not know which Haar coefficients will have mean shifts and do not know what the magnitudes of mean shift are, one intuitive idea is to adaptively and accurately estimate the post-change mean $\mathbf{c}^{(oc)}$ as we collect data for online monitoring under the sparsity assumption of $\mathbf{c}^{(oc)} - \mathbf{c}^{(ic)}$. Unfortunately, such an approach is generally computationally expensive and infeasible for online monitoring. Here we observe that the focus of profile monitoring is not necessarily on the accurate estimation of $\mathbf{c}^{(oc)}$, but on accurately raising a global alarm when there is a change. Hence, we propose a different approach that first locally monitors each component for a possible significant local mean shift, and then apply the order-thresholding technique to raise a global alarm under the sparse assumption that only a few local components are affected by the change.

When monitoring online profiles y_k's, at each time k, we first use (4.5) to derive the corresponding p-dimension Haar coefficients $\mathbf{c}_k$, and then standardize each of p components

by

$$X_{i,k} = \frac{c_k(i) - \hat{c}_i^{(ic)}}{\hat{\sigma}_i}, \tag{4.8}$$

for $i = 1, \cdots, p$, where $\{\hat{c}_i^{(ic)}, \hat{\sigma}_i\}_{i=1,\cdots,p}$ are estimators of the in-control mean $\mathbf{c}^{(ic)}$ and standard deviation σ in (4.7) based on in-control samples.

By (4.7), rigorously speaking, the normalized coefficients $X_{i,k}$ might not be i.i.d. $N(0,1)$ unless the tuning parameter $\rho_1 = 0$. In the context of online profile monitoring, the tuning parameter ρ_1 will often be small, and thus it is not bad to assume that the $X_{i,k}$'s satisfy the normality assumption from the practical viewpoint. Hence, the profile monitoring problem is reduced to the problem of monitoring the possible mean shifts of p-dimensional multivariate normal random vectors $\mathbf{X}_k = (X_{1,k}, \cdots, X_{p,k})$, where the means of some components may shift from 0 to some positive or negative value with magnitude of at least $\rho_2 > 0$.

If we know the exact post-change mean μ_i for the i-th component that is affected by the change, it is straightforward to develop an efficient local detection scheme, since one essentially faces the problem of testing the hypotheses in the change-point model where $X_{i,1}, \cdots, X_{i,\nu-1}$ are i.i.d. $f_0(x) =$ pdf of $N(0,1)$ and $X_{i,\nu}, \cdots, X_{i,n}$ are i.i.d. $f_1(x) =$ pdf of $N(\mu_i, 1)$. At each time k, we repeatedly test the null hypothesis $H_0 : \nu = \infty$ (no change) against the alternative hypothesis $H_1 : \nu = 1, 2, \cdots$ (a change occurs at some finite time), see Lorden, 1971. Thus the log generalized likelihood ratio statistic at time k becomes

$$W_{i,k}^* = \max_{1 \le \nu \le k} \frac{\prod_{\ell=1}^{\nu} f_0(X_{i,\ell}) \prod_{\ell=\nu+1}^{k} f_1(X_{i,\ell})}{\prod_{\ell=1}^{k} f_0(X_{i,\ell})}, \tag{4.9}$$

which can be recursively computed for normal distributions as

$$W_{i,k}^* = \max\left(W_{i,k-1}^* + \mu_i X_{i,k} - \frac{1}{2}(\mu_i)^2, 0 \right), \tag{4.10}$$

for $k = 1, \cdots$, with the initial value $W^*_{i,k=0} = 0$. In the literature, the statistic $W^*_{i,k}$ in (4.10) was first defined by Page (1954), and is called cumulative sum (CUSUM) statistics and enjoys theoretical optimality (Lorden, 1971; Moustakides, 1986).

In our context of profile monitoring, we do not know the value of the post-change mean μ_i except that $|\mu_i| \geq \rho_2$, thus we cannot use the CUSUM $W^*_{i,n}$ in (4.10) directly. One natural idea is to estimate μ_i from observed data, and then plug-in the estimated $\hat{\mu}_i$ into the CUSUM statistics in (4.10). For that purpose, at time k, denote by $\hat{\nu}_k$ the largest $\ell \leq k-1$ such that $W^*_{i,\ell} = 0$. Then the generalized likelihood ratio properties suggest that $\hat{\nu}_k$ is actually the maximum likelihood estimate of the change-point ν at time k, and thus one would expect that the data between time $[\hat{\nu}_k, k]$ would likely come from the post-change distributions, which allows us to provide a reasonable estimate of the post-change mean $\hat{\mu}_i$ at time k. This idea was first rigorously investigated in Lorden and Pollak, 2008 for detecting positive mean shifts of normal distributions, and here we aim to detect either positive or negative mean shifts. Specifically, at time k, for the i-th standardized Haar coefficients $X_{i,k}$'s, we define $\hat{\mu}^{(1)}_{i,k}$ and $\hat{\mu}^{(2)}_{i,k}$ as the estimates of the post-change mean of $X_{i,k}$ when restricted to the positive and negative values, respectively, under the assumption that $|\mu_i| \geq \rho_2$, with the explicit expressions as:

$$\hat{\mu}^{(1)}_{i,k} = \max\Big(\rho_2, \frac{s + S^{(1)}_{i,k}}{t + T^{(1)}_{i,k}}\Big) > 0, \qquad \hat{\mu}^{(2)}_{i,k} = \min\Big(-\rho_2, \frac{-s + S^{(2)}_{i,k}}{t + T^{(2)}_{i,k}}\Big) < 0, \tag{4.11}$$

and for $j = 1, 2$ and for any k, the sequences $(S^{(j)}_{i,k}, T^{(j)}_{i,k})$ are defined recursively

$$\begin{pmatrix} S^{(j)}_{i,k} \\ T^{(j)}_{i,k} \end{pmatrix} = \begin{cases} \begin{pmatrix} S^{(j)}_{i,k-1} + X_{i,k-1} \\ T^{(j)}_{i,k-1} + 1 \end{pmatrix} & \text{if } W^{(j)}_{i,k-1} > 0 \\ \begin{pmatrix} 0 \\ 0 \end{pmatrix} & \text{if } W^{(j)}_{i,k-1} = 0 \end{cases}. \tag{4.12}$$

Roughly speaking, for each estimate $\hat{\mu}^{(j)}_{i,k}$, if $\hat{\nu}^{(j)}_k$ is the candidate change-point, then $T^{(j)}_{i,k}$

denotes the time steps between $\hat{\nu}_k^{(j)}$ and k, whereas $S_{i,k}^{(j)}$ is the summation of all observations in the interval $[\hat{\nu}_k^{(j)}, k]$. The constants s and t in (4.11) are pre-specified, non-negative constants, and s/t can be thought of as a prior estimate of the post-change mean.

By plugging the adaptive estimations $\hat{\mu}_{i,k}^{(j)}$ of the post-change mean μ_i in the CUSUM statistics in (4.10), we can derive the local monitoring adaptive CUSUM statistics by

$$W_{i,k} = \max\left(W_{i,k}^{(1)}, W_{i,k}^{(2)}\right), \tag{4.13}$$

where $W_{i,k}^{(1)}$ and $W_{i,k}^{(2)}$ are the local detection statistics for detecting positive and negative mean shifts:

$$\begin{aligned} W_{i,k}^{(1)} &= \max\left(W_{i,k-1}^{(1)} + \hat{\mu}_{i,k}^{(1)} X_{i,k} - \frac{1}{2}(\hat{\mu}_{i,k}^{(1)})^2, 0\right), \\ W_{i,k}^{(2)} &= \max\left(W_{i,k-1}^{(2)} + \hat{\mu}_{i,k}^{(2)} X_{i,k} - \frac{1}{2}(\hat{\mu}_{i,k}^{(2)})^2, 0\right). \end{aligned} \tag{4.14}$$

4.3.3 Global online monitoring procedure

At time k, we have p local detection statistics $W_{i,k}$'s for $i = 1, \cdots, p$, one for monitoring each specific Haar coefficient locally. In general, the larger values of the $W_{i,k}$'s, the more likely the Haar coefficient is affected. Since we don't know which Haar coefficients are affected by the change, we follow Liu, Zhang, and Mei (2019) to raise a global alarm based on the largest r values of the $W_{i,k}$'s. This allows us to filter out those non-affected Haar coefficients, and provides the list of candidate affected Haar coefficients.

Specifically, at each time k, we order p local detection statistics $W_{i,k}$'s for p Haar coefficients, say, $W_{(1),k} \geq W_{(2),k} \geq ... \geq W_{(p),k}$ are order statistics of $W_{i,k}$'s. Then our proposed profile monitoring scheme $N(b, r)$ is to raise an alarm at first time when the summation of the top r statistics $W_{(1),k}, ..., W_{(r),k}$ exceed some pre-defined threshold b, i.e.,

$$N(b, r) = \inf\{k : \sum_{i=1}^{r} W_{(i),k} \geq b\}, \tag{4.15}$$

where r is the tuning parameter that is determined by the sparsity of the post-change, b is the pre-specified constant to control false alarm.

In summary, our proposed profile monitoring scheme $N(b, r)$ in (4.15) is based on monitoring Haar coefficients. We use recursive CUSUM procedures, which can adaptively estimate unknown changes, to monitor each Haar coefficient individually, and use order-thresholding to address the sparse post-change scenario when only a few Haar coefficients are affected by the change.

It is important to emphasize that our proposed procedure $N(b, r)$ is robust in the sense that it can detect a wide range of possible changes on the profiles without requiring any knowledge on the potential failure pattern. Additionally, by the recursive formulas in (4.12) and (4.14), for a new coming profile, our proposed procedure only involves a computational complexity of order $O(p)$ to update local detection statistics for p Haar coefficients, as well as additional order of $O(p\log(p))$ to sort these p local detection statistics. Thus at each fixed time step, the overall computational complexity of our proposed methodology is of order $O(p\log(p))$. Meanwhile, for the GLR procedure in Chicken, Pignatiello Jr, and Simpson (2009), the computational complexity is of order $O(t^2p^2)$ at time step t, which can be reduced to the order of $O(K^2p^2)$ if one only uses a fixed window size of K latest observations to make decisions instead of all t observations, where K often needs to be at least of order $O(\log(p))$ to be statistically efficient. Hence, as compared to the GLR procedure, our proposed procedure can be easily implemented recursively and thus is scalable when online monitoring high-dimension profile data over a long time period.

Algorithm 1 Implementation of our proposed procedure $N(b, r)$ in (4.15)

Initial parameters: ρ_1, ρ_2, s, t, and r.

In-control estimation: Using a set of m in-control p-dimensional profile samples $\mathbf{y_1}, .., \mathbf{y_m}$, perform the following steps.

Step 1: get the Haar coefficients $\mathbf{c_1}, .., \mathbf{c_m}$ by equation (4.5).

Step 2: get the estimation of standard deviation of the i^{th} Haar coefficient $\hat{\sigma}_i$.

Step 3: get $\hat{\mathbf{c}}^{(ic)}$ by equation (4.7) with the threshold ρ_1.

Online monitoring:

initialize $k = 0$, and set all initial observations $X_i = 0$ and all $S_i^{(j)} = T_i^{(j)} = W_i^{(j)} = 0$, for $i = 1, \ldots, p$ and $j = 1, 2$.

While the scheme $N(b, r)$ has not raised an alarm

do 1. Update $(S_i^{(j)}, T_i^{(j)})$ via (4.12).

2. Compute the intermediate variables $\hat{\mu}_i^{(j)}$ from (4.11) which are the estimates of the post-change means.

3. Input new p-dimensional profile $\mathbf{y}$, using the estimated in-control mean $\hat{\mathbf{c}}^{(ic)}$ and standard deviation $\hat{\sigma}$ to get the updated standardized p components $\{X_1, ..., X_p\}$ by (4.8).

4. For $i = 1, \ldots, p$, recompute the local monitoring statistics $W_i^{(j)}$ in (4.14) and W_i in (4.13).

5. Get the order statistics of $\{W_1, ..., W(p)\}$ denoted by $W_{(1)} \geq W_{(2)} \geq ... \geq W_{(p)}$

6. Compute the global monitoring statistics

$$G = \sum_{i=1}^{r} W_{(i)}$$

if $G \geq b$ **terminate:** Raising an alarm at time k and declaring that a change has occurred;

end the while loop

4.3.4 Parameter settings

For our proposed monitoring procedure $N(b, r)$, there are two global parameters, r and b, and four local parameters, ρ_1, ρ_2, s, t. Optimal choices of these parameters will depend on the specific applications and contexts, and below we will discuss how to set the reasonable values of those parameters based on our extensive numerical experiences.

Let us first discuss the choices of two global parameters, r and b. The optimal choice of r that maximizes the detection power of the proposed procedure $N(b, r)$ is the number

of truly changed Haar coefficients, which is often unknown. Based on our extensive simulations Liu, Zhang, and Mei (2019), when monitoring hundreds or thousands of Gaussian data streams simultaneously with a unknown number of affected local streams, the value $r \in [5, 10]$ often can reach a good balance on the detection power and the robustness to detect a wide range of possible shifts. Hence, in the case study and simulation study, we choose $r = 8$. As for the global parameter b, it controls when to stop the monitoring procedure and is often chosen to satisfy the pre-specified false alarm constraints. A standard approach in the literature is to choose b by repeatedly sampling in-control measurements either from in-control training data or from Monte Carlo in-control models, so that the monitoring procedure $N(b, r)$ will satisfy false alarm constraint.

Next, the local parameter ρ_1 in (4.7) essentially conducts a dimension reduction for in-control profiles. A good choice of the ρ_1 will depend on the characteristics of in-control profile data in specific applications, and in general the cutoff threshold ρ_1 should be chosen balance the bias-variance tradeoff of estimation of the in-control mean profile. Much theoretical research has been done on how to choose ρ_1 for the single profile (Donoho and Johnstone, 1994; Donoho and Johnstone, 1998). These existing approaches focus more on the wavelet coefficient or mean profile estimation in the context of de-noising while the main objective in our context is to detect the changes of wavelet coefficients. Since we will conduct another dimension reduction at the layer of local detection statistics, it is often better to be conservative to choose a small constant $\rho_1 > 0$ value so as to keep more Haar coefficients from the in-control profiles. Also automatic or tuning-free approaches have been developed to choose the cutoff threshold such as ρ_1 adaptively in other contexts, see Zou and Qiu (2009) and Zou, Wang, Zi, and Jiang (2015). However, such approaches are often computationally expensive, and it is unclear how to extend them to multiple profiles monitoring while keeping the proposed procedure to be scalable. In our simulation and case study, we found out that a simple choice of $\rho_1 = 0.15$ will yield significantly better results as compared with the existing methods in the literature. It remains an open problem

to derive the optimal choice of ρ_1 under the general setting so that our propose procedures are efficient in both computational and statistical viewpoints.

Finally, the local parameter ρ_2 represents the interested-smallest magnitude of mean shift of wavelet coefficients to be detected. In practice, it can be set based on the engineering domain knowledge to ensure production yield. In this chapter, we set $\rho_2 = 0.25$. In addition, the local parameters, s and t in (4.11), are related to the prior distribution of the unknown post-change mean μ_i, so that the corresponding estimators of μ_i is a Bayes estimator and will be more robust than using the sample mean directly. In this chapter, we follow Lorden and Pollak (2008) to choose $s = 1$ and $t = 4$.

4.4 Case study

In this section, we apply our proposed wavelet-based methodology to a real progressive forming manufacturing process dataset in Lei, Zhang, and Jin (2010) that includes 307 normal profiles and 5 different groups of fault profiles. Each group contains 69 samples which are collected under the faults due to missing part occurring in one of these five operations respectively. Additionally, there are $p = 2^{11} = 2048$ measurement points in each profile.

The original research on Lei, Zhang, and Jin (2010) focuses on the offline classification of normal and fault profile samples, while our research mainly emphasizes on the fast online detection. We will compare the performance of our proposed monitoring procedure with the other two common used procedures to illustrate the efficiency of our scheme. First one is the Hotelling's T^2 control chart based on selected wavelet coefficients (Zhou, Sun, and Shi, 2006). The second one is based on the asymptotic maximum-likelihood test in Chicken, Pignatiello Jr, and Simpson (2009). Specifically, we consider the following three procedures:

- Our proposed method $N(b, r)$ in (4.15);

- Hotellings T^2 control chart based on the first r out of p wavelet coefficients:

$$T(b,r) = \inf\{j \geq 1 : w_j \geq b\}.$$

where

$$w_j = \sum_{i=1}^{r} \left(\frac{c_j(i) - \hat{c}_i^{(ic)}}{\hat{\sigma}_i^2}\right)^2$$

- The method in Chicken, Pignatiello Jr, and Simpson (2009), where the generalized likelihood ratio test was used on all p wavelet coefficients:

$$M^*(b) = \inf\left\{n \geq 1 : \max_{1\leq i<n}\left\{\left[\frac{\sum_{j=i+1}^{n}\tilde{w}_j}{n-i} - \frac{\sum_{j=1}^{i}\tilde{w}_j}{i}\right] * \sum_{j=i+1}^{n}\left(\frac{w_j}{p} - 1\right)\right\} \geq b\right\}.$$

where

$$\begin{aligned}\tilde{w}_j &= \frac{1}{\hat{\sigma}^2}\sum_{i=1}^{p}\{\max(0, |c_j(i) - \hat{c}_i^{(ic)}| - \lambda)\}^2 \\ \lambda &= \sqrt{2\frac{\log p}{p}}\hat{\sigma}.\end{aligned}$$

In order to have a fair comparison, r is chosen as 8 for our proposed method $N(b,r)$ in (4.15) and the Hotellings T^2 control chart $T(b,r)$ in (4.16).

To evaluate the detection efficiency of those methods, we first find the appropriate values of the global threshold b such that the average run length of each scheme is 200 when the samples are collected by sampling from the 307 normal profiles with replacement. Then, using the obtained global threshold value b, we simulate the detection delay when the samples are sequentially collected by sampling from the 69 fault profiles. All Monte Carlo simulations are based on 500 repetitions. The results of detection delay and standard error are summarized in Table 4.1.

From Table 4.1, we can see all of these three methods can detect the change of Fault 1, 2, 3 and 5 very fast (on average, just need one sample to detect such change). It is necessary

Table 4.1: A comparison of the detection delays of 3 methods with in-control average run length equal to 200 based on 500 repetitions in Monte Carlo simulations. The standard errors of the detection delays are reported in the bracket.

Method	Fault 1	Fault 2	Fault 3	Fault 4	Fault 5
N(b=73,r=8)	1(0)	1(0)	1(0)	1.51(0.03)	1.01(0.01)
T(b=23.33,r=8)	1(0)	1(0)	1(0)	17.71(0.78)	1(0)
$M^*(b = 600)$	1(0)	1(0)	1(0)	4.47(0.13)	1.22(0.02)

to emphasize that although as shown in Figure 4.2, the difference between normal profile and the Fault 4 profile is very subtle, our proposed method can detect the Fault 4 change much faster than the other two methods.

4.5 Simulation study

In this section, we present the simulation study results to illustrate the efficiency of our proposed procedure. We follow the nonlinear profile monitoring literature to consider the in-control mean profile as the Mallet's piecewise smooth function in Mallat (1999) , see Figure 4.4. This testbed curve is a complicated function with several non-differentiable points and difficult patterns, including several transient jumps, therefore cannot easily be modeled by parametric models or other non-parametric models and has been popularly used in much research to evaluate the performance of nonlinear profile monitoring procedures, see Jeong, Lu, and Wang (2006), Chicken, Pignatiello Jr, and Simpson (2009), and Lee, Hur, Kim, and Wilson (2012).

The out-of-control mean profile follows the same setup in the previous literature (Lee, Hur, Kim, and Wilson, 2012) and assumes a local mean shift on some intervals. Specifically, the out-of-control mean profiles are designed as $f_1(x) = f_0(x) + \mu I_\delta(x)$ where the shift magnitude $\mu \in \{0.25, 0.5, 1\}$ and three different changed intervals: (1) $\delta = [0, 1]$, which is referred as Global shift; (2) $\delta = [\frac{73}{512}, \frac{76}{512}] \cup [\frac{288}{512}, \frac{296}{512}]$, which is referred as Local shift I and (3) $\delta = [\frac{3}{512}, \frac{15}{512}] \cup [\frac{344}{512}, \frac{347}{512}]$, which is referred as Local shift II.

Based on the mean profiles, we generate in-control and out-of-control sample profiles,

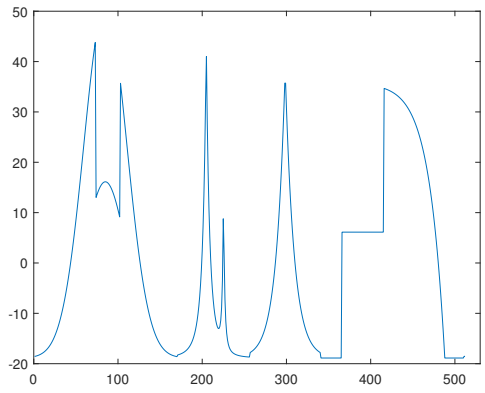

Figure 4.4: Mallat's piecewise smooth function.

which consist of a realization of $p = 512$ pairs $(x_i, y(x_i))$ with $x_1, ..., x_p$ equal spaced on $[0, 1]$ and $y(x_i) = f_0(x_i) + \epsilon(x_i)$ as in-control sample profile and $y(x_i) = f_1(x_i) + \epsilon(x_i)$ as out-of-control sample profile, where $\epsilon(x_i)$ is i.i.d standard normally distributed $N(0, 1)$.

We will compare the performance of our proposed method $N(b, r = 8)$ in (4.15) with the same two methods in the previous section: the method $M^*(b)$ in (4.16) and the method $T(b, r = 8)$ in (4.16). In this simulation study, we still set $\rho_1 = 0.15, \rho_2 = 0.25, s = 1, t = 4$ for our proposed scheme.

Specifically, based on 1000 Monte Carlo simulations, we keep the in-control average run length of those schemes as 200 and compare the detection delay under the Global shift, Local shift I and Local shift II with different magnitudes of mean shift. The results are summarized in Table 4.2.

From Table 4.2 we can see that (1) our proposed method $N(b, r)$ yields the smallest detection delay for detecting local shifts compared with the other two methods $M^*(b)$ and $T(b, r)$; (2) a competitive results for detecting the global shifts under different magnitudes of shifts. This implies our proposed wavelet-based monitoring procedure is more robust to the unknown changes.

Table 4.2: A comparison of the detection delays of 3 methods with in-control average run length equal to 200 based on 1000 repetitions in Monte Carlo simulations. The standard errors of the detection delays are reported in the bracket

Method	μ	Global shift	Local shift I	Local shift II
$N(b=51, r=8)$	0.25	2.59(0.01)	92.38(0.52)	67.41(0.42)
	0.5	1(0.01)	31.63(0.18)	22.17(0.14)
	1	1(0.00)	9.46(0.05)	6.53(0.04)
T(b=21.7,r=8)	0.25	1.03(0.01)	151.82(4.68)	253.57(7.15)
	0.5	1.00(0)	144.38(4.39)	100.59(2.99)
	1	1.00(0)	79.08(2.58)	24.81(0.74)
$M^*(b=10.1)$	0.25	8.26(0.18)	157.40(4.81)	151.55(4.73)
	0.5	1.29(0.02)	125.24(4.09)	106.31(3.58)
	1	1.00(0)	35.97(0.87)	24.55(0.55)

4.6 Conclusions

In this chapter, we develop a new scalable scheme for monitoring nonlinear profiles with unknown post-change distribution. This chapter makes three methodological contributions. First, we propose to use all wavelet coefficients to monitor the process, while the prior literature of nonlinear profile monitoring is dominated by analyzing and using just significant coefficients. Second, we propose to use two shrinkage techniques to filter out the noise introduced by using all wavelet coefficients. One is using hard shrinkage to estimate the in-control mean coefficients. The other one is to build monitoring procedure only focusing on the information of a few coefficients, which have higher likelihood to be changed. Third, we propose to utilize a recent developed adaptive-CUSUM procedure in Liu, Zhang, and Mei (2019) to efficiently monitor the standardized wavelet coefficients without knowing the information about the post-change.

There is plenty of room for improving our proposed scheme for monitoring nonlinear profiles, calling for further research. First, this chapter mainly focuses on the detection of mean shift of the normal distributed profile. Although there are many applications of our proposed scheme, it is also necessary to work on the detection procedures for more generally distributed profiles. Second, this chapter makes an independence assumption on

the noise distribution in (4.1). It will be useful to develop a more robust method that can handle different correlation structure of the profile data.

CHAPTER 5
APPLIED RESEARCH IN MODELING OF PAPERS' CITATION TRAJECTORIES

5.1 Introduction

Science is a skewed world where a small number of publications receive a disproportionate amount of citations. What do citation trajectories of the most cited papers look like? Do they follow the typical citation trajectory documented in the literature, specifically, the annual citation counts of a paper rise to a peak in the first few years after publication and then slowly fade away over time? Figure 5.1 plots annual citations of the top 10 most cited papers published in the American Physical Society (APS) journals, and their annual citations are counted in the Web of Science (WoS) from the year of publication to 2016. Among them the youngest was published in 1999, and the oldest 1964. Correspondingly, the length of their observed citation trajectories range from 18 to 53 years. In addition to their exceptionally large number of citations, a remarkable observation is that most of them (at least seven out of ten) do not even show any sign that their annual citations are about to peak and will start to decline in the near future. We refer to this phenomenon of continual rise in annual citations without decline as evergreens, which clearly violates the typical pattern of citation trajectory. Although we cannot predict whether these papers will remain highly cited in the future, the fact that they have not yet become obsolete after up to 53 years calls for attention, especially considering that the majority of papers reach their citation peak around the 3rd or 5th year after publication and that most bibliometric analyses examine citations in a relatively short time window.

The objective of this chapter is to better understand evergreens in particular and pat-

[1]The materials in this chapter were published in *Journal of Informetrics*, 2017.

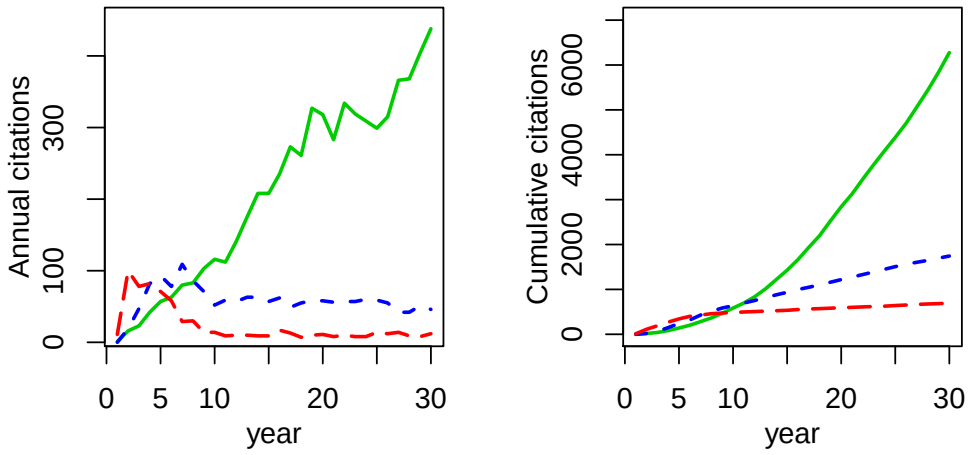

Figure 5.1: three selected papers.

terns of citation trajectory in general. Moreover, do evergreens constitute a general type of citation trajectory, or are they so rare that they cannot be captured in any statistical cluster analysis? To this end, we develop a functional data analysis (FDA) method to analyze the 30-year citation trajectories of a sample of publications published in 1980 in APS journals. Our FDA method integrates functional principal components analysis, Poisson regression, and K-means clustering. More specifically, we model the citation trajectories of individual publications by a small number of common basis functions and paper-specific coefficients on these basis functions. For each paper, its 30-dimensional vector of citations can be characterized by its coefficients on the common basis functions, which subsequently serve as inputs for the K-means clustering, to uncover general types of citation trajectories. Results of our cluster analysis provide strong evidence that evergreens exit as a general class of citation trajectory. In addition, we cannot predict whether a paper will become an evergreen by some ex ante paper features such as the number of authors and references. The remainder of this paper is organized as follows. We begin with a brief review of previous cluster analyses of citation trajectories and the functional data analysis, followed by a description of our dataset. Next, our proposed model and method is presented, with the emphasis on how to combine functional principal component analysis, Poisson regression,

and K-means clustering algorithm for modeling and clustering citation trajectories. Then we report the empirical results of our proposed model and method to the real citation data set. Implications of our findings are also discussed.

5.2 Prior literature

5.2.1 Clustering citation trajectories

Citation ageing is a long-standing research topic, and different patterns of citation trajectories have been documented in the bibliometrics literature (Aversa, 1985; Avramescu, 1979; Baumgartner and Leydesdorff, 2014; Garfield, 1980; Glänzel and Schoepflin, 1995; Line, 1993; Redner, 2005; Wang, 2014). Aversa (1985) conducted probably the first rigorous statistical analysis of citation trajectories, investigating 9-year citation trajectories of 400 highly cited papers published in 1972 and applying the K-means clustering algorithm to the normalized annual citation counts (i.e., annual citations divided by total citations in the whole studied time period). Aversa (1985) identified two clusters: delayed rise - slow decline and early rise - rapid decline. Costas, Leeuwen, and Raan (2010) analyzed about 30 million documents in WoS published between 1980 and 2008. Following Prices observation, documented in his personal communication to Aversa (1985) , Costas, Leeuwen, and Raan (2010) classified papers into three categories: 50% papers as normal documents, 25% as delayed documents, and 25% as flashes-in-the-pan. However, these three clusters are defined based on a single real-valued summary statistics of individual papers, Year 50%, defined as the year when a paper has cumulated half of its total citations up to year 2008. Moreover, there are no statistical justification on the proportion of these three clusters. More recently, Colavizza and Franceschet (2016) examined about half million papers published in APS journals and applied the spectral clustering method on the normalized annual citations received by these papers within the APS database. The three identified general types of citation trajectories are middle-of-the-roads, sprinters, and marathoners. Middle-of-the-roads papers display an average citation ageing pattern, and can be viewed

as corresponding to normal documents. Sprinters has an early and high peak and a fast decline, which can be viewed as flashes-in-the-pan. Marathoners represent fast or slow-rise, moderately peaked histories, followed by a slow decline, or absence of decline, or even a constant rise in received citations over time and therefore can correspond to delayed documents or evergreens. The phenomenon of evergreens, which were emphasized by Avramescu (1979) and Price (see Aversa (1985)), were not identified by clustering analyses in Aversa (1985) and Costas, Leeuwen, and Raan (2010), while marathoners in some specifications in Colavizza and Franceschet (2016) also display a continually increasing annual citation curve. One possible explanation is that these later cluster analyses focus on general types, while evergreens are rather outliers and therefore cannot be identified in statistical cluster analysis.

5.2.2 Functional data analysis

Functional data analysis (FDA) is a recent new development in the field of statistics and has a tremendous growth over the past decades (Besse and Ramsay, 1986; Rice and Silverman, 1991; Hoover, Rice, Wu, and Yang, 1998; Ramsay and Silverman, 1997; Yao, Müller, and Wang, 2005; Hall, Müller, and Wang, 2006; Leng and Müller, 2006; Hadjipantelis, Aston, and Evans, 2012). FDA might be particularly useful for bibliometric analysis for two reasons: First, FDA is a non-parametric method and therefore is useful for analyzing bibliometric data for which the underlying distribution is often unclear. Second, FDA analyzes high-dimensional data, such as curves and shapes, which are of particular interest to bibliometric studies. Using regression analysis as an analogy, while traditional regression analysis only allows one real-valued dependent variable, FDA allows both dependent and independent variables to be multidimensional. Most FDA methods deal with continuous data, but paper citations to be analyzed in this study are discrete count data. There are only a few FDA studies dealing with count variables (Linde, 2009; Serban, Staicu, and Carroll, 2013; Wu, Müller, and Zhang, 2013), and our proposed method is different from

these limited existing FDA methods for Poisson data. Specifically, we propose to adapt the methods in Rice and Silverman (1991) from Gaussian distributed data to Poisson count data by exploring the close relationship between Poisson and Gaussian distributions.

5.3 Data

The data used for this study are research papers published in 1980 in the American Physical Society (APS) journals, specifically six journals which were active in 1980: Physical Review A, B, C, D, Physical Review Letter, and Reviews of Modern Physics. APS journal paper citations trajectories have been extensively studied in prior literature (Colavizza and Franceschet, 2016,Redner, 2005,Wang, Song, and Barabási, 2013). We only include original research papers labeled as article and exclude other document types such as review or note. There are a total of 4023 research papers, and their cumulative citations in the first 30 years after publication, i.e., between 1980 and 2009, are retrieved from the Web of Science (WoS). Since a sufficient amount of citations are required for reliable modeling of the citation trajectories (Aversa, 1985,Colavizza and Franceschet, 2016,Wang, Song, and Barabási, 2013), we decide to focus on papers with at least 30 citations in the first 30 years after publication. The resulting dataset consists of 1699 papers. For a robustness test, we also analyzed the top 400 cited papers and obtained similar clustering results.

There is considerable variation in individual papers citation trajectories in our dataset. Figure 5.2 plots the citation trajectories of four selected papers . Three of them loosely resemble the three general types labeled by Costas, Leeuwen, and Raan (2010) as flash-in-the-pan (red curve), normal document (blue curve), and delayed document (purple curve). The normal document (blue) follows the typical citation aging pattern, where the citations gradually increase and then decrease over time. The flash-in-the-pan (red) has relatively faster citation rising and declining processes, while the delayed document (purple) has relatively slower citation rising and declining process. All these three types follow the typical pattern of citation trajectory, although they vary in the general speed of citation ageing.

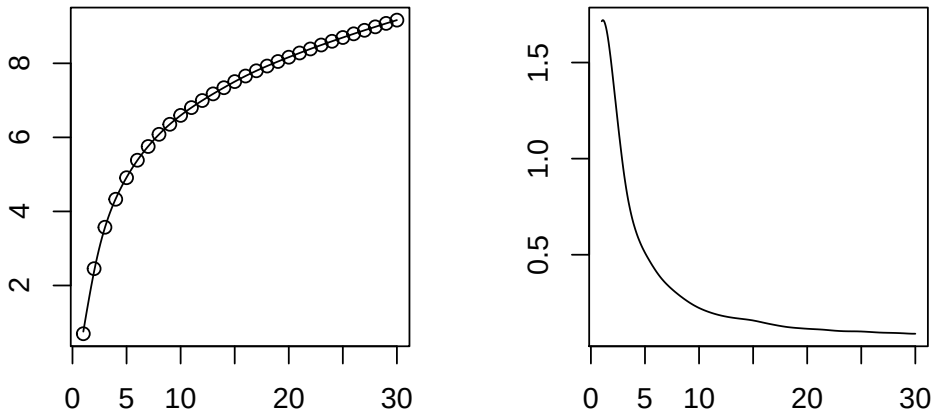

Figure 5.2: mean and first derivative function.

However, the green curve has a continually rising annual citation curve without a declining stage during the first 30 years after being published. In particular, the green curve in Figure 5.1 illustrates the annual and cumulative citations of the paper in Physical Review Letters entitled Ground State of the Electron Gas by a Stochastic Method coauthored by Ceperley and Alderby, which was the most cited paper up to year 2009 among all papers published in 1980 in APS journals. In addition, Figure 5.2 suggests the uncertainty in paper citations and the difficulty of using short-term citations to predict long-term citations. Specifically, while evergreen papers would eventually become extremely highly cited, their citations in the first few years are not necessarily very large. Moreover, the citation distribution is highly skewed. In terms total number of citations in the period of 30 years, 10 papers (0.6%) in our sample have citations greater than 1000, 50 papers (2.9%) greater than 400, and 1244 papers (73.2%) fewer than 100. This implies that the distribution of papers across different types of citation trajectories might also be uneven.

5.4 Methodology

The objective of this paper is to empirically uncover general types of citation trajectories based on the observed paper citation time series and examine whether evergreens constitute

a general type of citation trajectory. The main idea is to use functional principal component analysis and Poisson regression to model citation trajectories, which allows us to conduct dimension reduction, that is, to characterize the vector of 30-years citation counts of a paper by a much smaller number of parameters derived from our model. Subsequently these parameters can be used as the inputs for the K-means cluster analysis for uncovering general types of citation trajectories.

5.4.1 Functional Poisson regression model

We develop a nonparametric model for the cumulative citations based on functional Poisson regression. The nonparametric approach does not impose any theoretical assumptions on the mechanisms underlying the citation process but lets the data speak for themselves. We adopt this explorative approach in order to better understand divergent citation patterns in real life.

For each paper $i = 1, \cdots, N$, denote by $X_i(j)$ the cumulative number of citations for the i-th paper in year j after being published, where $j = 1, \cdots, T$. We choose to use the cumulative annual citation counts instead of the annual citation counts because the former is smoother and thus easier to model. For notational convenience, we denote $X_i(0) = 0$, and assume that the observed cumulative citation $X_i(j)$'s are the realization of a counting process $X_i(t)$ for $0 \leq t \leq T$.

For each paper $i = 1, \cdots, N$, $x_i(j)$ denotes the observed cumulative number of citations for the i-th paper in year j after being published, where j is a discrete time variable (i.e., year) and $j = 1, \cdots, T$. For notational convenience, we denote $x_i(0) = 0$. Our proposed functional Poisson regression model assumes that the observed cumulative citation $x_i(j)$s are the realization of a counting process $X_i(t)$ for the continuous time variable t and $0tT$.

$$X_i(t) \sim \text{Poisson}(\mu_i(t)), \tag{5.1}$$

for $t \geq 0$, where the mean function $\mu_i(t)$ satisfies

$$\sqrt{\mu_i(t)} = \eta(t) + \sum_{\nu=1}^{\ell} \xi_{i,\nu}\phi_\nu(t) \tag{5.2}$$

and the (offset) functions $\eta(t)$ and the basis functions $\phi_\nu(t)$'s are smooth functions of t that are the same to all papers. Their estimations will be further discussed later.

In Equation (2) we adopt a square-root transformation for the mean function $\mu_i(t)$. Note that for Poisson regression, or more generally Generalized Linear Models, there are two popular transformation for the mean $\mu_i(t)$ of the count data: one is the log-transformation $\log(\mu_i(t))$, and the other is the square-root transformation $\sqrt{\mu_i(t)}$ (Nelder and Wedderburn, 1972). For any given basis functions $\phi_\nu(t)s$, both transformation strategies have been widely used in the statistics literature, and which transformation is better depends on the specific application and dataset. In the context of this study, the square root transformation is preferable. In the functional principal component analysis for deriving basis functions (as will be explained later), we will approximate the Poisson distribution of citation counts by a Gaussian distribution via a square-root transformation, which allows us to take advantage of the rich literature of FDA for Gaussian distributed data (Rice and Silverman, 1991; Ramsay and Silverman, 2005). Therefore, taking the square-root transformation strategy here for the Poisson regression matches the square-root transformation in functional principal component analysis and accordingly yields a better fit to the data. In addition, in standard principal component analysis, the number l of basis functions is assumed to be relatively small, while the retained basis functions should be able to explain most information of the original data. Under our model in Equation (1) and (2), the goal is essentially to find an estimate $\hat{\mu}_i(t)$ that is a smooth version of $X_i(t)s$ with certain correlation structure. In addition, it is also useful to think our proposed model as a dimension reduction, representing the T-dimensional cumulative citations of a paper as a ℓ-dimensional vector of coefficients $\xi_{i,\nu}s$. Subsequently, the problem of identifying general citation patterns can

then be reduced to the cluster analysis of the ℓ-dimensional vector of coefficients $\xi_{i,\nu}s$.

5.4.2 Model parameter estimation

When fitting the functional Poisson regression model in Equation (1) and (2) to the observed cumulative citations $x_i(j)$s of the N papers, we need to estimate two kinds of unknown quantities: the common basis functions $\eta(t)$ and $\psi_\nu(t)$s which are the same for all papers, and the paper-specific coefficients $\xi_{i,\nu}$s which are tailored for each paper individually. Clearly they are closely related, and there are no unique estimation methods. Here we propose to estimate them by using the functional principal component analysis method and Poisson regression, respectively. Regarding the estimation of the common basis functions $\eta(t)$ and $\psi_\nu(t)$s in Equation (2), intuitively one should use information across all the observed N papers. From the functional decomposition viewpoint, these basis functions can be any set of orthogonal bases, although some bases are more efficient than others. In the functional data analysis literature, the estimation of these basis functions has been well-studied for Gaussian distributed data, e.g., Rice and Silverman (1991) and Ramsay and Silverman (2005). Here we propose to adapt these prior methods to Poisson count data by exploring the close relationship between Poisson and Gaussian distributions. For a Poisson random variable X with a large mean $\mu > 0$, a well-known fact is $\sqrt{X} \sim N(\sqrt{\mu}, 0.5^2)$ (Thacker and Bromiley, 2001). Note that the variance of $\sqrt{X}$ is approximately constant, and thus the square-root transformation of Poisson data is often referred to as the variance-stabilizing transformation in the statistical literature (Anscombe, 1948). Brown, Carter, Low, and Zhang (2004) also used the square-root transformation to establish the global asymptotic equivalence between Poisson process and Gaussian process. In this paper we consider the square-root transformation of the count variable, $\sqrt{(X_i(t))}$, so that the bases $\eta(t)$ and $\psi_\nu(t)$s in Equation (2) can be estimated by applying the rich functional data analysis literature to "approximate Gaussian" data $\sqrt{(X_i(t))}$, e.g., Rice and Silverman (1991) and Ramsay and Silverman (2005). Specifically, the square-root transformation of the ob-

served citation counts for each paper can be modeled as being independent realizations of a stochastic process $Y(t) = \sqrt{X(t)}$, with mean $E(Y(t)) = \eta(t)$ and covariance function$\gamma(s,t) = cov(X(s), X(t))$. We assume that there is an orthogonal expansion (in the L_2 sense) of $\gamma(s,t)$ in terms of eigenfunctions

$$\gamma(s,t) = \sum_{\nu=1}^{\infty} \lambda_\nu \psi_\nu(s)\psi_\nu(t). \tag{5.3}$$

According to the Karhunen-Love expansion theorem, a random citation curve $Y_i(t) = \sqrt{(X_i(t))}$ may then be expressed as

$$\sqrt{(X_i(t))} = \eta(t) + \sum_{\nu=1}^{\infty} \xi_{i,\nu}\psi_\nu(t), \tag{5.4}$$

where the $\xi_{i,\nu}$s are uncorrelated random variables with mean 0 and variance $Var(\xi_{i,\nu}) = \lambda_\nu$ (Rice and Silverman, 1991;Hall, Müller, and Wang, 2006). Therefore, functions $\eta(t)$ and $\psi_\nu(t)$s in Equation (4) are close related to the mean function and correlation function of the stochastic process $Y(t) = \sqrt{(X(t))}$, and we will use them as the basis functions in Equation (2). The basis functions $\eta(t)$ and $\psi_\nu(t)$s in Equation (4) for Gaussian data can be estimated by spline smoothing and functional principal component analysis methods in Rice and Silverman (1991) and Ramsay and Silverman (2005). After estimating the common basis functions $\eta(t)$ and $\psi_\nu(t)$s, the next step is to estimate the coefficients $\xi_{i,\nu}$s in the standard Poisson regression model from observed raw citations $x_i(j)$s. This can be done by maximum likelihood estimation for Poisson regression, which is implemented in many statistical packages. In our analysis, the estimation of the coefficients $\xi_{i,\nu}$s is done on a Windows 8 Laptop with Intel i7-4510U CPU 2.0GHz by using the glm() function in the free statistical software R (version 3.1.1).

5.4.3 Cluster analysis

Given that the N papers and their corresponding cumulative citation curves $x_i(j)$s can be represented as N points in the ℓ-dimensional space of coefficients $(\xi_{i,1}, ..., \xi_{i,\ell})$, we propose to conduct cluster analysis by applying the K-means clustering algorithm to the induced ℓ-dimensional coefficient space. In addition, in this ℓ-dimensional coefficient space, the coefficients $\xi_{i,\nu}$s in Equation (2) correspond to different basis functions $\psi_\nu(t)$ and vary considerably in scale. Therefore, we first standardize coefficients $\xi_{i,\nu}$s by

$$\tilde{\xi}_{i,\nu}) = (\xi_{i,\nu} - \mu_\nu)/s_\nu \tag{5.5}$$

where μ_ν and s_ν are respectively the mean and standard derivation of the fitted N coefficient values $(\xi_{1,\nu}, ..., \xi_{N,\nu})$, for each principal component $\nu = 1, ..., \ell$. Subsequently, we define the distance between papers in term of citation trajectories as the Euclid distance of the standardized coefficients $(\tilde{\xi}_{i,1}, \cdot, \tilde{\xi}_{i,\ell}))$ in the ℓ-dimensional space, based on which we use the K-means clustering algorithm to cluster papers into K different groups. Given the explorative nature of this study, we experiment and compare clustering results for $K = 2, 3, 4, 5,$ and 6 clusters.

5.4.4 Summary of methodology

Our proposed functional Poisson regression model for clustering paper citation trajectories can be summarized as follows.

- Given T-years cumulative citation trajectories of N papers $x_i(j)$ for $i = 1, 2, , N$, and $j = 1, 2, , T$, first derive the square-root transformed data, $y_i(j) = \sqrt{(x_i(j))}$.
- Estimate the mean functions $\eta(t)$ and eigenfunctions $\psi_\nu(t)$s of the transformed data $y_i(j)$, using functional principal component analysis.
- Determine ℓ, the number of eigenfunctions $\psi_\nu(t)$s to retain.

- For each individual paper i, use the mean functions $\eta(t)$ and ℓ eigenfunctions $\psi_\nu(t)$s as basis functions and fit a Poisson regression model to its observed cumulative citation trajectory $(x_i(1), x_i(2), , x_i(T))$. This yields, for each individual paper, the estimated coefficients $(\xi_{i,1}, \xi_{i,2}, , \xi_{i,\ell})$. Accordingly, the T-dimension vector of cumulative citations for paper $i, (x_i(1), x_i(2), , x_i(T))$, can be represented by its ℓ-dimensional vector of coefficients, $(\xi_{i,1}, \xi_{i,2}, , \xi_{i,\ell})$.

- Standardize each coefficient $\xi_{i,\nu}$ by $\tilde{\xi}_{i,\nu} = (\xi_{i,\nu} - \mu_\nu)/s_\nu$, where μ_ν and s_ν are the mean and standard derivation of the N fitted coefficient values $(\xi_{1,\nu}, \xi_{2,\nu}, , \xi_{N,\nu})$, for each principal component $\nu = 1, \cdots, \ell$.

- Apply the K-means clustering algorithm to the standardized coefficients $\tilde{\xi}_{i,\nu}$ to group N papers into K clusters.

5.5 Results

This section reports the numerical results of applying our proposed model and method to our sampled 1699 APS papers.

5.5.1 Estimating basis functions

The basis functions $\eta(t)$ and $\psi_\nu(t)$s play an important role in our proposed model and method, and they are estimated in R (version 3.1.1) using the codes of Ramsay, Hooker, and Graves (2009).

Figure 5.3 plots the estimated mean curve $\eta(t)$ and its first derivative $\eta'(t)$. Here $\eta(t)$ and $\eta'(t)$ are closely related to the average cumulative citations and average annual citations over time, respectively. The estimated first derivative $\eta'(t)$ is positive but decreases over time. This is consistent with the "typical citation pattern that the annual citations generally are the largest in early years and subsequently decline slowly.

Figure 5.4 plots the estimated smoothing versions of the first four eigenfunctions $\psi_\nu(t)$s.

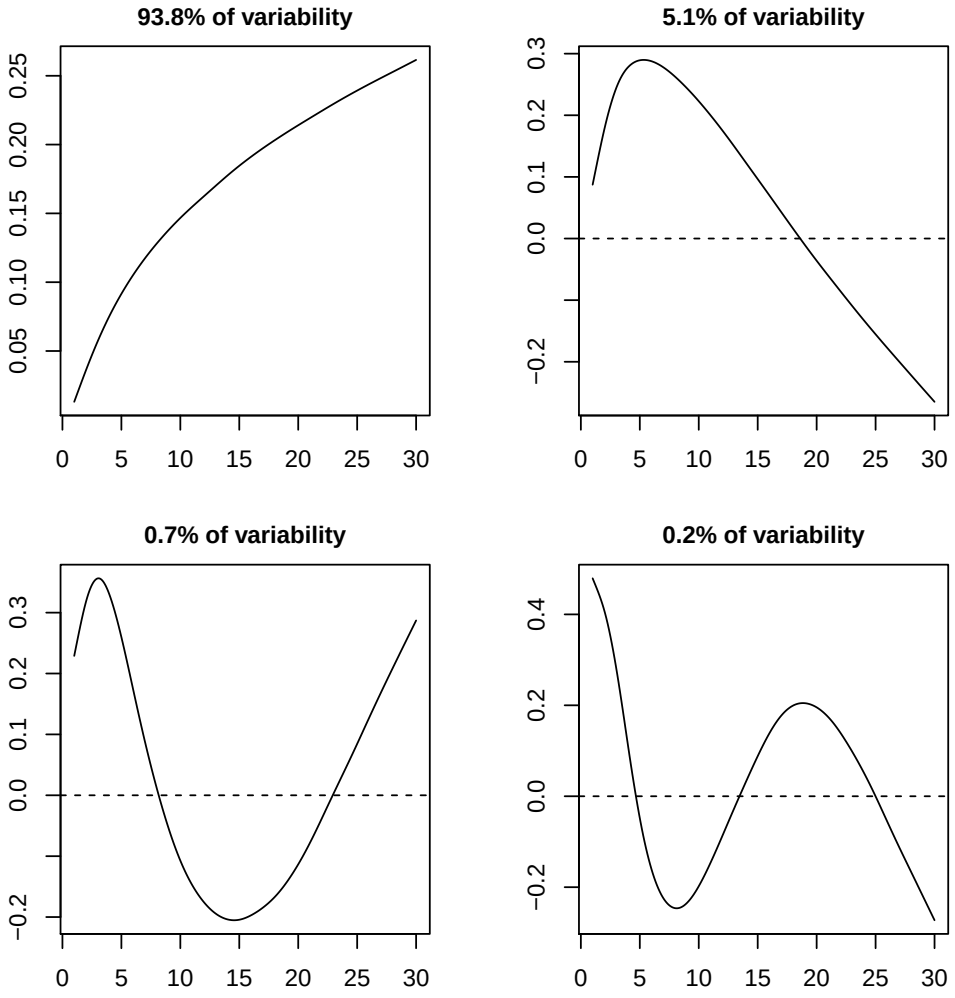

Figure 5.3: four eigenfunctions.

They correspond to the four largest eigenvalues of 299.86, 16.39, 2.17 and 0.65, and these four eigenfunctions account for 93.8%, 5.1%, 0.7% and 0.2% of the total variability, respectively. The shape of these eigenfunctions indicates how a papers cumulative citation trajectory might deviate from the mean curve $\eta(t)$. Specifically, the first smoothed eigenfunction $\hat{\psi}_1(t)$ is positive and monotonically increasing. Therefore, if a paper has a positive coefficient on $\hat{\psi}_1(t)$, then this paper will have more citations than an average paper (i.e., the mean curve) across all years, and more importantly its advantage over an average paper magnifies over time. This observation is consistent with the well-known cumulative advantage or preferential attachment phenomenon in citations. The second smoothed

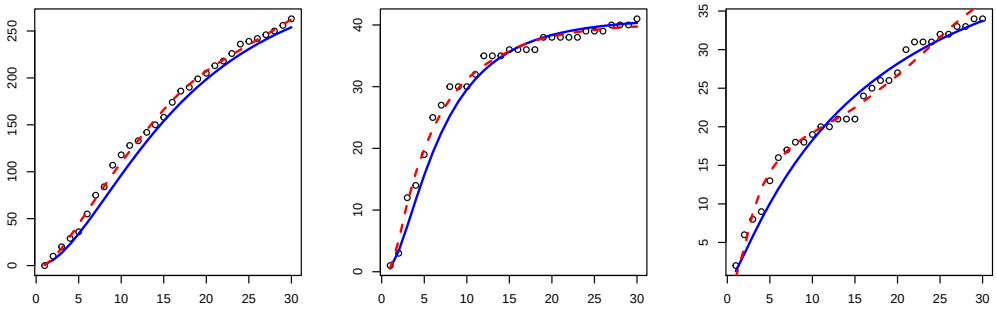

Figure 5.4: fitted results for three papers.

eigenfunction $\hat{\psi}_2(t)$ is positive in early years but negative in late years. If a paper has a positive coefficient on $\hat{\psi}_2(t)$, then this paper would have relatively more citations in early years than an average paper but fewer citations in later years, displaying a relatively fast citation ageing process. The third and fourth smoothed eigenfunctions, $\hat{\psi}_3(t)$ and $\hat{\psi}_4(t)$, capture more fine-grained fluctuation in citation trajectories over time. Furthermore, they both exhibit a periodic pattern, suggesting that the highly or less cited feature can be cyclic.

5.5.2 Determining the number of eigenfunctions

A critical step of our analysis is to decide how many eigenfunctions to retain, for which there is still no standard procedure in the FDA literature (Wang, Chiou, and Mueller, 2015). The rule of thumb is to choose a reasonably small number ℓ of eigenfunctions that not only explain high percentage (e.g., 95% or 99%) of total variation but also have a good fit to the observed data. Therefore, we take into account both the total explained variability and the goodness of fit. In terms of explained variability, the first one, two and three eigenfunctions together account for 93.8%, 98.9% and 99.6% of the total variability, respectively. According to the rule of thumb, that is, 95% or 99% of total variation to retain, we can choose $\ell = 2$ or 3.

We then examine the goodness of fit. Figure 5.5 evaluates the goodness of fit for the first $\ell = 2, 3, 4, 5$ basis functions using the mean square error (MSE) criterion. More precisely, results in Figure 5.5 are based on 10-fold cross validation: We randomly partition the 1699

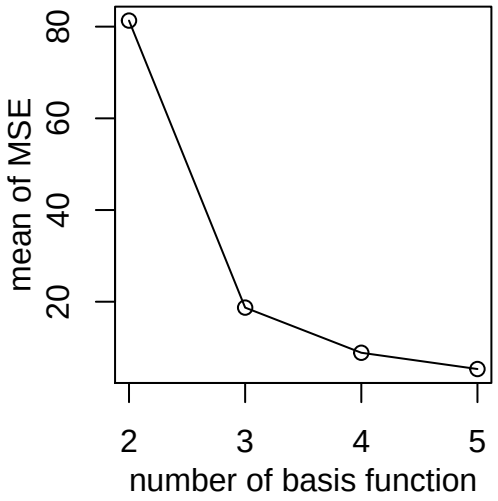

Figure 5.5: cross-validation.

papers into 10 subgroups, where 9 subgroups have 170 papers and the 10th subgroup has 169 papers. For papers in each subgroup, we fit the functional Poisson regression model using the $\ell = 2, 3, 4, 5$ basis functions estimated from all papers in the other subgroups. Then we calculate the error of the fit, that is, the difference between fitted and observed values. The average mean standard squares error (MSE) for the 1699 papers by using different number of basis functions are plotted in Figure 5.5. Based on this graph, we can adopt a strategy similar to the Cattells scree test, that is, search for the elbow point. It seems that the goodness of fit improves considerably when increasing ℓ from 2 to 3, while further increase in ℓ only improve the goodness of fit marginally. Therefore, we choose $\ell = 3$, partly because increasing ℓ from 2 to 3 brings the largest improvement in fitting performance and partly because the first three eigenfunctions contain 99.6% variability, which is sufficiently high.

5.5.3 Fitting individual paper models

Based on the estimated basis functions, we fit our proposed functional Poisson regression model to each individual paper in the dataset, following the procedure as described in Sec-

tion 5.4.2. For evaluating the fitness of our model, we compare our model with a recently developed parametric model for individual papers citations, in Wang, Song, and Barabási (2013) (hereafter the WSB model). Wang, Song, and Barabási (2013) model paper citations by a Poisson process, specifically, the expected cumulative number of citations of the i-th paper in year $t(t \geq 0)$ is

$$\Lambda_i(t) = m\left(\exp\left\{\lambda_i \Phi\Big(\frac{\log(t) - \mu_i}{\sigma_i}\Big)\right\} - 1\right), \tag{5.6}$$

(6) where $\Phi(t)$ is the cumulative density function of the standard normal $N(0,1)$ random variable, λ_i, μ_i, and σ_i are three paper-specific parameters that describe the citation trajectory of the i-th paper, and parameter m is a global constant for the average citations of all papers and is set at 30 in Wang, Song, and Barabási (2013). For fitting individual paper models, the natural choice is to use the estimated basis functions $\eta(t)$ and $\psi_\nu(t)$s in Section 5.5.1 directly to derive the estimated coefficients $\xi_{i,\nu}$s in Equation (5.2) for each individual paper (as will be implemented in the next subsection for clustering analysis). However, using this approach for comparing model fitting performance is to some extent unfair to the WSB model, because our functional Poisson regression model would have used the same dataset twice: One at the population level for estimating basis functions and the other at the individual paper level for estimating paper-specific coefficients on the basis functions. However, the WSB model uses the data only once. Therefore, for a relatively fair comparison of model fitting, we use to the same 10-fold cross-validation as discussed in Section 5.5.2. More precisely, we randomly partition the 1699 papers into 10 subgroups, where 9 subgroups have 170 papers and the 10th subgroup has 169 papers. For papers in each subgroup, we fit the functional Poisson regression model using the $\ell = 3$ basis functions estimated from all papers in the other subgroups and the WSB model separately. Then we calculate the mean squares error (MSE) of the fit by our model and WSB model.

To assess the goodness of fit, we compare the distribution of residuals. In addition,

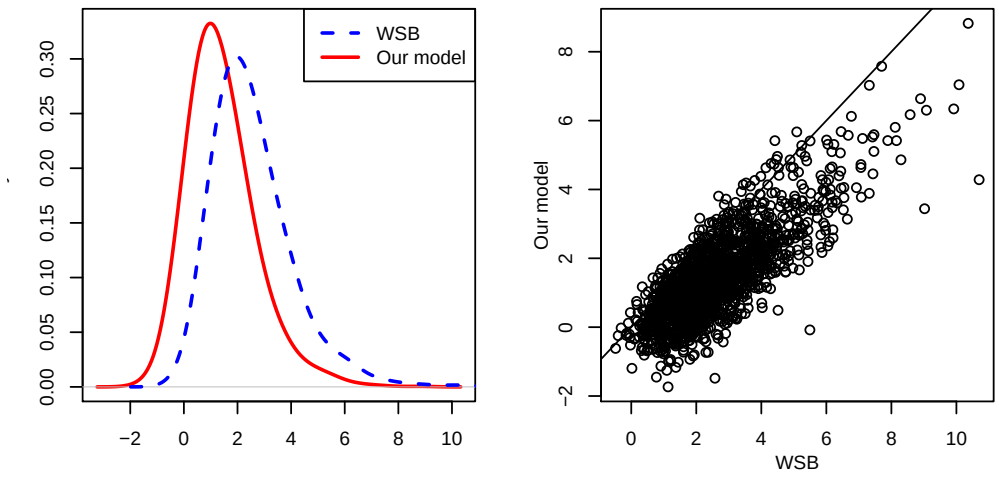

Figure 5.6: Kernel and scatter plots to compare fitting results with wsb model.

we plot log MSEs instead of MSEs at the original scale, considering that the distribution of MSEs is highly skew. Figure 5.6 left panel plots the kernel densities of log MSEs. Our functional Poison regression model clearly has smaller MSEs, and the Wilcoxon sum rank test further suggests that the MSEs of our proposed functional Poisson regression model are stochastically smaller than those of the WSB model. In addition, Figure 5.6 right panel reports a scatter plot of log MSEs, which suggests that our proposed model fits most papers (i.e., points below the diagonal line) better than the WSB model. It is important to note that this comparison is still to some extent unfair to the WSB model. We used a separate training set for estimating our basis functions, although this training set does not overlap with the testing set, it still shares some similarity with the testing set, for example, both sets are physics papers published in 1980. In addition, the WSB model is developed for predicting long-term citations, while the goal of our model is to have a parsimonious characterization of citation trajectories with satisfactory goodness of fit. Therefore, WSB would avoid overfitting, while our model would intentionally over-fit the data to certain degree. For the same reason, we opted for the original WSB model documented in Wang, Song, and Barabási (2013) for this comparison, instead of the WSB-with-prior model documented in Shen, Wang, Song, and Barabási (2014). The WSB-with-

prior model incorporates a conjugate prior and thereby reduces the number of estimated parameters, for avoiding overfitting. Compared with the original WSB model, the WSB-with-prior model has a lower fitting power but a higher prediction power. In summary, based on the comparison results, we do not claim that our model is superior to the WSB model, but only conclude that our model does fit the data well.

5.5.4 Clustering paper trajectories

Using the estimated basis functions $\eta(t)$ and the first three $\psi_\nu(t)$s from the whole sample of 1699 papers as reported in Section 5.5.1, we estimate coefficients $\xi_{i,\nu}$s in Equation (5.2) for each of the 1699 papers. These estimated coefficients $\xi_{i,\nu}$s are in turn used as inputs for the K-means clustering analysis. Given the explorative nature of this clustering analysis, we experiment with different number of clusters, ranging from two to six.

We first report results for four clusters. To illustrate characteristics of the identified four clusters, or general types of citation trajectories, we find the centers of each cluster in the 3-dimensional standard coefficients spaces and then convert them back into the original paper citations space to derive four central curves in terms of cumulative and annual citations (Figure 5.7). The number of observations in each cluster is as follows: red (972 papers, that is, 57.2% of the whole sample of 1699 papers), blue (454 papers, 26.7%), purple (228 papers, 13.4%), and green (45 papers, 2.7%). Both the red and blue curves in Figure 5.7 are consistent with previous clustering studies (Aversa, 1985; Costas, Leeuwen, and Raan, 2010; Colavizza and Franceschet, 2016)), in the sense that the speed of citation aging is slow for some papers while relatively fast for others. However, the year of citation peak seems to be the same for both the red and blue curves, while the only difference is about the scale of the peak. Therefore, both red and blue curves might belong to the category of normal documents as labeled by Costas, Leeuwen, and Raan (2010). We name the red curve as normal I and the blue curve as normal II. The purple curve, compared with both the red and blue ones, display a slower rising process, as well as a slower declining process after the

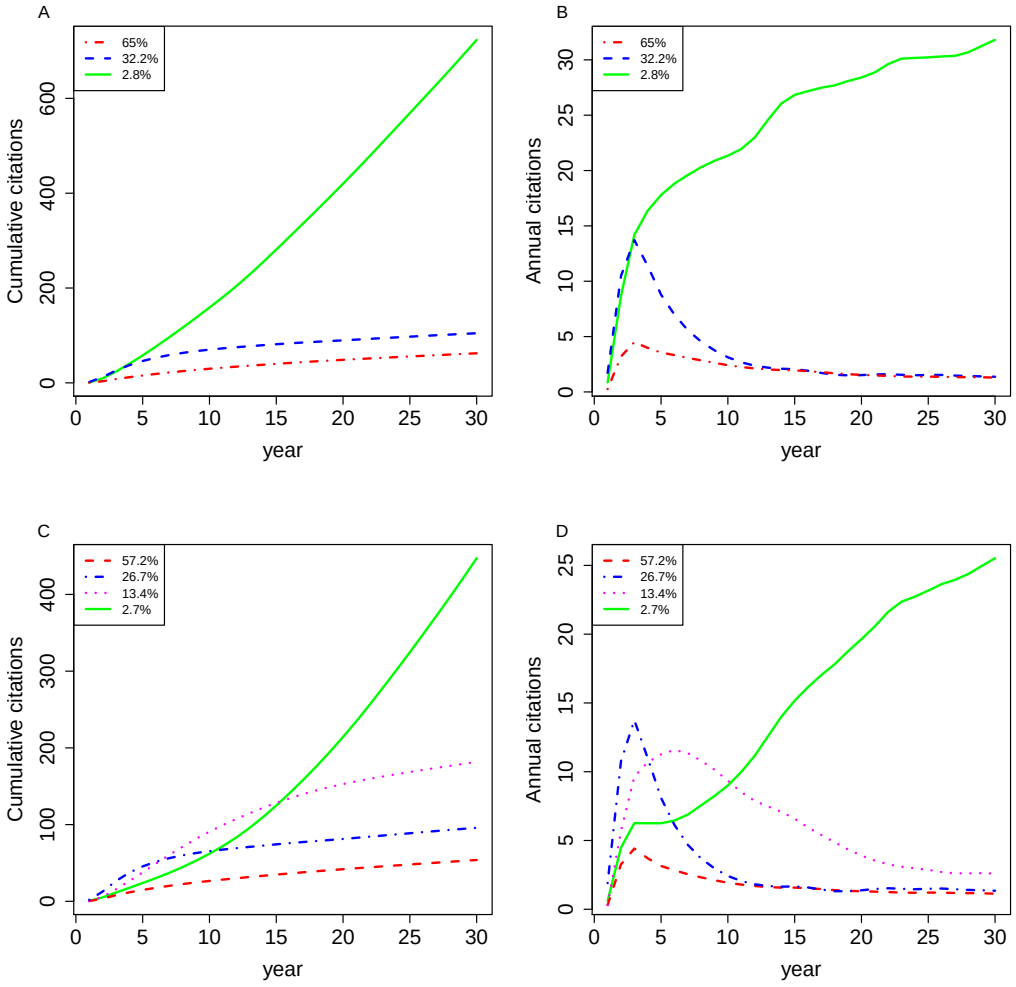

Figure 5.7: clustering based on our method with K=3 and 4.

citation peak. The timing of its citation peak is later than the red and blue ones. The scale of its citation peak is lower than the blue one but higher than the red one. In addition, its total number of 30-year citations is larger than the red and blue ones. The purple curve corresponds to the delayed documents, as labeled by Costas, Leeuwen, and Raan (2010). The most interesting curve in Figure 5.7 is the green one, which clearly demonstrates a continual rise in annual citations without declining within the 30-years period after publication. We refer to this type of papers as evergreens, which were emphasized by Price (see Aversa (1985)) and Avramescu (1979) but were not identified by later cluster analyses (Aversa, 1985; Costas, Leeuwen, and Raan, 2010). Marathoners in some specifications in Colav-

izza and Franceschet (2016) also display a continually increasing annual citation curve. These evergreens appear to have fewer citations than the normal II and delayed documents in the first few years after publications but clearly much more citations in the long run. Furthermore, all other types (i.e., normal I, normal II, and delayed documents) still follow a "typical citation trajectory, where a papers annual citations rise to its peak shortly after publication and then slowly decline, although some types reach the citation peak higher or faster than others. However, evergreens clearly violate this "typical pattern, at least within the 30-year time window, which is much longer than the citation time window adopted in most bibliometric analyses.

Results for other choices of K are reported in Figure 5.8. On the one hand, decreasing K would miss some types of citation trajectories. For example, the three-cluster result (Figure 5.8A3) misses delayed documents, and the two-cluster result (Figure 5.8A2) additionally misses evergreens. On the other hand, increasing K from 4 to 5 or 6 does not uncover new types which are sufficiently distinct from the identified four types, and additional clusters in Figure 5.8A5-6 locate in a continuous space from fast to slow ageing, following the "typical pattern. In order to better evaluate the performance of our proposed clustering approach, we compare our proposed clustering method, which clusters citation trajectories based on the ℓ-dimensional vector of standardized paper-specific coefficients $\tilde{\xi}^*_\nu$s, with two alternative approaches, specifically, clustering based on (a) the T-dimensional vector of the raw annual citations (raw annual method) and (b) the T-dimensional vector of the proportion of annual citations (proportion method, i.e., normalized annual citations, the number of annual citations in each year divided by the number of total citations over the T years). For the comparison of clustering results we focus on two aspects: the shape of the central curves and the distribution of papers across clusters. Clustering results using the proportion method for $K = 2, \cdots, 6$ are reported in Figure 5.8B2-6. Compared with our proposed method, the proportion method clusters papers more evenly across different clusters. In terms of the shape of the central curves, using K = 4 (Figure 5.8B4) as an example, all

four curves seem to reach their peak around the same time (while the purple line is a bit later than the others, and the green curve has an initial local peak at around the same time, followed by a decline and then start rise again), although they display very different speed of citation declining. In addition, the speed of citation declining seems to be positively associated with the scale of the peak. For example, the blue curve has the highest peak and also the fastest citation decline after the peak. It is difficult to interpret the clusters. Maybe the red, purple, and blue curves can be labeled as delayed document, normal document, and flash-in-the-pan respectively, according to their speed of rising and declining, but the red one does not seem to have a later peak than the others. In addition, it is unclear how to interpret the green curve, they seem to have a continual rise in annual citations (if we ignore the decline following the first local peak), similar to our identified evergreens. However, different from evergreens, the number of annual citations of the green type in Figure 5.8B4 is a small constant. Furthermore, most papers in the green cluster have very limited number of total citations. One possible explanation is that this alternative approach uses the proportion of annual citations, which is very sensitive when a paper has a relatively small number of total citations. Central curves of annual citations resulting from the clustering method based on raw annual citations are plotted in Figure 5.8C2-6. The clustering result is dominated by the scale of citations, but does not reveal distinct features between different clusters in terms of the shape of citation curves. Take 4-cluster results (Figure 5.8C4) as an example, 94.2% papers (red) have a moderate number of citations, 5.1% papers (purple) have even fewer citations, 0.6% papers (blue) have considerably more citations, and 0.1% papers (green) are extremely highly cited. Except the green curve, all others show a similar shape in the citation curve, and the difference between them is the scale of citations. Although this alternative approach also successfully identify a small number of evergreen papers (i.e., the green curve), it misses a number of true evergreen papers, specifically, papers that are classified as evergreens by our proposed method but not by this alternative approach actually also exhibit a pattern of continual rise in annual citations. Thus, we con-

clude that clustering using raw annual citations is over-dominated by the scale of citations and is inadequate for capturing nuanced difference in the shape of citation trajectories.

5.6 Discussion

This paper proposes a nonparametric functional Poisson regression model to describe citation trajectories of individual papers and combines our model with the K-means clustering algorithm for cluster analysis, using the coefficients of the eigenfunctions in our model. Results suggest the existence of evergreens as a general type of citation trajectories. This paper makes two methodological contributions. First, we develop a functional data analysis method for discrete count data, by combining principal component analysis and Poisson regression, while the prior literature of functional data analysis is dominated by analyzing continuous data. Second, this paper also demonstrates the usefulness of the functional data analysis for bibliometric studies. Because it is a nonparametric approach and is designed for analyzing high-dimensional data, the functional data analysis can be a powerful tool for bibliometric analysis.

5.6.1 Limitations and future research

This study has several limitations. First, constrained by data availability, we cannot claim whether our observed evergreen papers will remain being (highly) cited in the future or will eventually become obsolete. Although the latter is very plausible, the former is not entirely impossible. Larivière, Archambault, and Gingras (2008) show that researchers have been relying on an increasingly old body of literature since mid-1960s, so it is still possible that some classic pieces will never experience obsolesce or obliteration by incorporation, that is, becoming commonly known and integrated into the daily work in the field that it is no longer explicitly cited (Merton, 1973). Although we cannot draw a conclusive inference on the fate of our identified evergreen papers, the finding that a considerable number of papers assemble characteristics of evergreens in a 30-year time period is still very relevant for sci-

ence and bibliometric studies, since most studies and evaluations use a shorter time window and assume a the typical citation trajectory. Second, this study uses a sample of journal articles in one field (i.e., physics) and one year (i.e., 1980), and accordingly has a limitation in terms of generalizability. Third, although our method can single out evergreens, it does not identify sleeping beauties in science (Ke, Ferrara, Radicchi, and Flammini, 2015; Raan, 2004). This is probably because sleeping beauties are very rare and therefore are difficult to identify in large scale statistical analyses (Colavizza and Franceschet, 2016). There is plenty of room for improving our functional data analysis method for citation data, calling for further research. From the functional smoothing viewpoint, the cumulative citation curve must be non-decreasing. While our proposed fitting method yields non-decreasing fitted curves numerically for the cumulative citations of all 1699 papers in our dataset, it is important to develop a better estimation method that guarantees the non-decreasing property theoretically, e.g., using the monotone smoothing method developed in Ramsay (1998). From the cluster analysis viewpoint, we conduct unsupervised learning in our dataset and rely on prior literature and our domain knowledge on paper citation behavior, for assessing the classification results of different approaches. It will be useful to develop a more objective criterion for evaluating results of cluster analysis. In addition, we have some interesting observations that evergreens is negatively correlated with the number of pages and authors, more research is required for better understanding what determines the citation trajectory of a paper. The regression model using readily available paper feature for predicting evergreens has very poor performance, we would need to investigate what kind of intrinsic paper quality might predict whether a paper becomes an evergreen in science.

5.6.2 Implications

Results of this chapter have three important implications for bibliometric studies and research evaluations. First, our findings demonstrate that papers with similar citations in the short run may have completely different citation patterns in the long run. Compared

with normal documents, delayed documents and evergreens receive fewer citations in the short run but more citations in the long run. This serves as a warning about the bias in the use of short-time-window citation counts in research evaluations. Second, the observation of evergreens calls for more research on the "endurance" of citation impact, in addition to the aspect of "delay" emphasized in prior literature. Phenomena of scientific prematurity (Stent, 1972), delayed recognition (Garfield, 1980), and sleeping beauties (Raan, 2004) have been extensively studied in previous literature, which focus on the long time lag before a scientific contribution makes notifiable impact. On the other hand, Evergreens, similar as the term of marathoners in Colavizza and Franceschet (2016), reminds the other important but understudied aspect of citation trajectoryendurance. Third, evergreens also have implications for parametric models of citation trajectories. There is a strong interest in modelling citation trajectories, partly because it is a challenging scientific problem and partly because of the policy interest in predicting long-term citations. In a recent report published in Science, Wang, Song, and Barabási (2013) proposed a parametric nonhomogeneous Poisson process to model the citation trajectory of individual papers. Although this model is elegant from the pure mathematical viewpoint, its predictive power is unsatisfactory, especially for those highly cited ones (Wang, Mei, and Hicks, 2014). One possible explanation is that it assumes a "typical" citation trajectory, while evergreens, which are highly cited, do not follow this pattern. Our nonparametric analysis and identified general types of citation trajectory questions this assumption and shed light on future parametric modeling of citation trajectories.

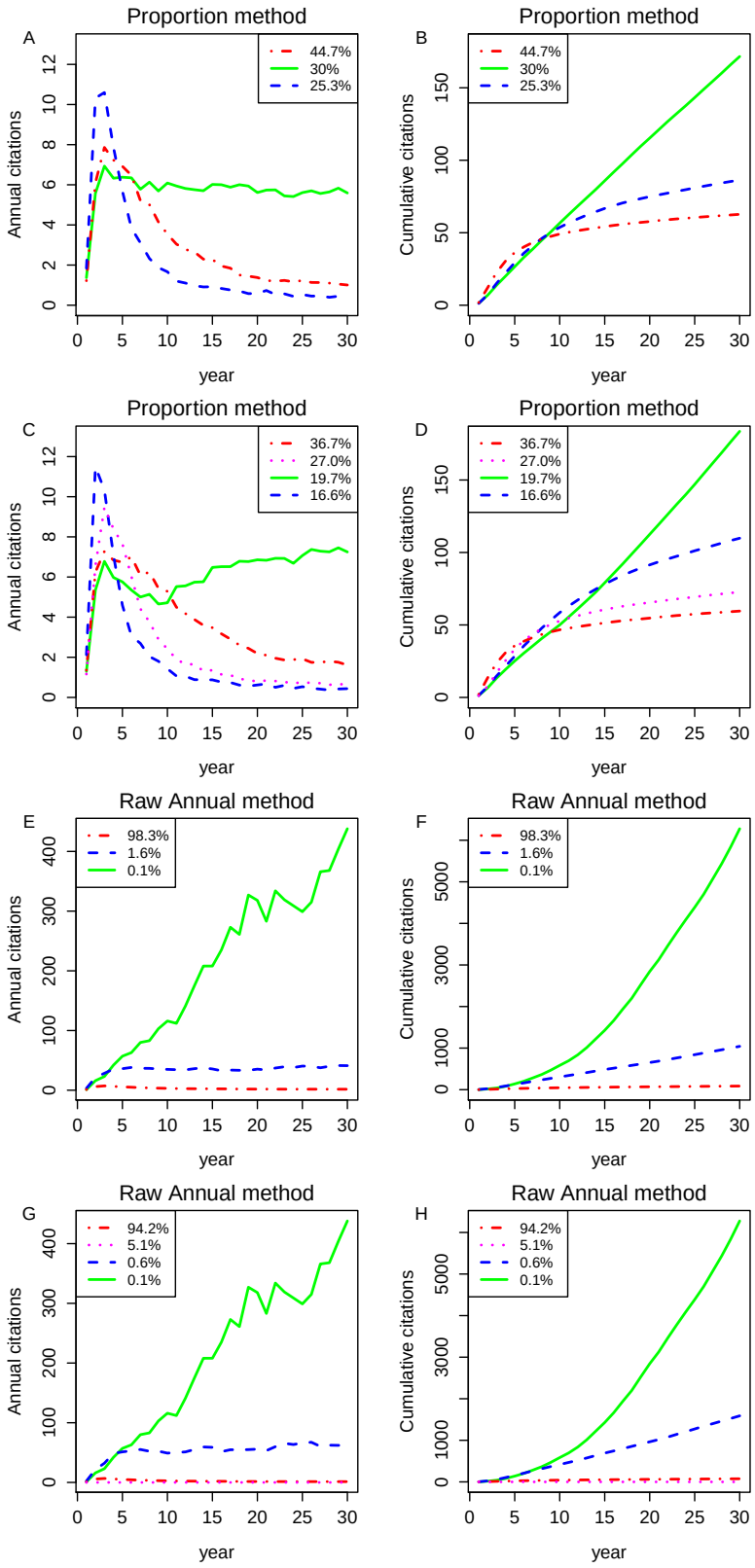

Figure 5.8: Clustering based on raw annual method and proportion method with K=3 and 4.

CHAPTER 6
CONCLUSIONS AND FUTURE RESEARCH

6.1 Summary of original contributions

This thesis contributes to the area of sequential change-point detection, robust sparse learning and monitoring of high-dimensional streaming data, and functional data analysis from both theoretical point of view and applied point of view. The original contributions of this thesis include the following aspects:

- **Robust change-point detection.** In the first chapter, we develop a new L_α-CUSUM local detection statistic, which is more robust than the classical CUSUM statistics. Moreover, on the global level, we combine those local L_α-CUSUM statistic together by soft-shrinkage transformation. We show he resulting global monitoring scheme enjoy nice theoretical statistical efficiency and robustness for monitoring high-dimensional data streams. Moreover, we propose a new concept called false alarm breakdown point, which can measure the robustness of any online monitoring procedure and show our proposed robust schemes indeed have positive breakdown point.

- **Detection delay analysis.** In the second chapter, we conduct detection delay analysis on some families of communication-efficient monitoring schemes under both classical low-dimensional regime where the number of data streams K is fixed and modern high-dimensional regime where the number of data streams K goes to ∞. Our theoretical results provide statistical foundation that using appropriate shrinkage can help increase communication efficiency in the large-scale sensor network while still keeping good detection efficiency.

- **Tractability of robust M-estimator.** In the third chapter, we investigate the robust-

ness properties and computational tractability of general (non-convex) M-estimator in both classical low-dimensional regime and modern high-dimensional regime. Our results reveal the M-estimator in general can achieve the minimax estimation error rate and has only one unique stationary point when the proportion of outliers is small. Therefore, we explain the reason why the M-estimator can be computed efficiently and can be widely used.

- **Nonlinear profile monitoring** In the fourth chapter, we proposed a novel profile monitoring procedure by combining the Wavelet technique, two-side adaptive CUSUM statistics, and order-shrinkage technique. We show our proposed method has good detection performance compared with other methods in literature by simulations and real data case study.

- **Modeling of papers' citation trajectories** In the fifth chapter, we propose a new functional Poisson regression model to fit and learn individual paper's citation trajectory. We show our model can not only fit papers' citation data well, but also be used for clustering papers into different citation patterns. We conduct careful interpretation on our classification results and demonstrate they can provide useful implication for bibliometric studies and research evaluations

6.2 Future research

My future research plan involves the development of modeling and monitoring methodologies for complex systems. My research agenda is not limited to develop data-driven statistical models but also to build up theoretical foundations of the developed methodologies. A more detailed discussion on my future research topics are provided below:

- **Robust feature extraction for image or matrix type data.** With the rapid development of advanced sensing techniques, high-quality image or video types data are much cheaper to get while reflecting more information of the complex systems. One

may lose a lot of spatial information if we simply treat the image or matrix type data as a vector. Therefore, we propose to decompose it by the 2-dimensional tensor basis and work on penalized M-estimator in tensor regression to extract important features. Moreover, we would like to investigate whether the good computation properties can be guaranteed or not in terms of tensor regression. After extracting those important features, we can apply our robust monitoring schemes in Chapter 1 directly to do change-point or anomaly detection for the streaming image data.

- **Robust sequential online decision: adaptive sampling and sequential estimation.** Another extension to my current research is to develop a robust sequential decision-making framework for high-dimensional data streams in term of adaptive sampling, in which we need to dynamically change the location of sensors to capture more useful information. Furthermore, I would like to investigate more robust adaptive estimation methodologies with desirable theoretical properties that can recursively update model parameters based on new observed data.

REFERENCES

[1] A. Alfons, C. Croux, and S. Gelper, "Sparse least trimmed squares regression for analyzing high-dimensional large data sets," *The Annals of Applied Statistics*, vol. 7, no. 1, pp. 226–248, 2013.

[2] A. Anandkumar, R. Ge, D. Hsu, S. M. Kakade, and M. Telgarsky, "Tensor decompositions for learning latent variable models," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 2773–2832, 2014.

[3] D. F. Andrews, P. J. Bickel, F. R. Hampel, P. J. Huber, W. H. Rogers, and J. W.Tukey, *Robust estimates of location: Survey and advances*. Princeton University Press, 1972.

[4] F. J. Anscombe, "The transformation of poisson, binomial and negative-binomial data," *Biometrika*, vol. 35, no. 3, pp. 246–254, 1948.

[5] S. Appadwedula, V. V. Veeravalli, and D. L. Jones, "Energy-efficient detection in sensor networks," *IEEE Journal on Selected areas in communications*, vol. 23, no. 4, pp. 693–702, 2005.

[6] E. Aversa, "Citation patterns of highly cited papers and their relationship to literature aging: A study of the working literature," *Scientometrics*, vol. 7, no. 3-6, pp. 383–389, 1985.

[7] A. Avramescu, "Actuality and obsolescence of scientific literature," *Journal of the American Society for Information Science*, vol. 30, no. 5, pp. 296–303, 1979.

[8] Z. Bai, C. R. Rao, and Y Wu, "M-estimation of multivariate linear regression parameters under a convex discrepancy function," *Statistica Sinica*, vol. 2, no. 1, pp. 237–254, 1992.

[9] S. Banerjee and G. Fellouris, "Decentralized sequential change detection with ordered cusums," in *2016 IEEE International Symposium on Information Theory (ISIT)*, IEEE, 2016, pp. 36–40.

[10] T. Banerjee and V. V. Veeravalli, "Data-efficient quickest change detection in sensor networks," *IEEE Transactions on Signal Processing*, vol. 63, no. 14, pp. 3727–3735, 2015.

[11] M. Basseville and I. V. Nikiforov, *Detection of abrupt changes: Theory and application*. Prentice Hall Englewood Cliffs, 1993, vol. 104.

[12] A. Basu, I. R. Harris, N. L. Hjort, and M. C. Jones, "Robust and efficient estimation by minimising a density power divergence," *Biometrika*, vol. 85, no. 3, pp. 549–559, 1998.

[13] S. E. Baumgartner and L. Leydesdorff, "Group-based trajectory modeling (gbtm) of citations in scholarly literature: Dynamic qualities of "transient" and "sticky knowledge claims"," *Journal of the Association for Information Science and Technology*, vol. 65, no. 4, pp. 797–811, 2014.

[14] A. E. Beaton and J. W. Tukey, "The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data," *Technometrics*, vol. 16, no. 2, pp. 147–185, 1974.

[15] Y. Bengio, "Learning deep architectures for ai," *Foundations and trends® in Machine Learning*, vol. 2, no. 1, pp. 1–127, 2009.

[16] P. Besse and J. O. Ramsay, "Principal components analysis of sampled functions," *Psychometrika*, vol. 51, no. 2, pp. 285–311, 1986.

[17] S. Boucheron, G. Lugosi, and P. Massart, *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

[18] T. Brooks, S. Pope, and M. Marcolini, "Uci machine learning repository," 2014.

[19] L. D. Brown, A. V. Carter, M. G. Low, and C. H. Zhang, "Equivalence theory for density estimation, poisson processes and gaussian white noise with drift," *The Annals of Statistics*, vol. 32, no. 5, pp. 2074–2097, 2004.

[20] E. Candes and T. Tao, "The dantzig selector: Statistical estimation when p is much larger than n," *The Annals of Statistics*, vol. 35, no. 6, pp. 2313–2351, 2007.

[21] E. J. Candes, "Modern statistical estimation via oracle inequalities," *Acta numerica*, vol. 15, pp. 257–325, 2006.

[22] E. J. Candes, X. Li, and M. Soltanolkotabi, "Phase retrieval from coded diffraction patterns," *Applied and Computational Harmonic Analysis*, vol. 39, no. 2, pp. 277–299, 2015.

[23] E. Cantoni and E. Ronchetti, "Robust inference for generalized linear models," *Journal of the American Statistical Association*, vol. 96, no. 455, pp. 1022–1030, 2001.

[24] H. P. Chan *et al.*, "Optimal sequential detection in multi-stream data," *The Annals of Statistics*, vol. 45, no. 6, pp. 2736–2763, 2017.

[25] L. Chang, S. Roberts, and A. Welsh, "Robust lasso regression using tukey's biweight criterion," *Technometrics*, vol. 60, no. 1, pp. 36–47, 2018.

[26] S. I. Chang and S. Yadama, "Statistical process control for monitoring non-linear profiles using wavelet filtering and b-spline approximation," *International Journal of Production Research*, vol. 48, no. 4, pp. 1049–1068, 2010.

[27] T.-C. Chang and F.-F. Gan, "Monitoring linearity of measurement gauges," *Journal of Statistical Computation and Simulation*, vol. 76, no. 10, pp. 889–911, 2006.

[28] M. Chen, C. Gao, and Z. Ren, "A general decision theory for hubers epsilon-contamination model," *Electronic Journal of Statistics*, vol. 10, no. 2, pp. 3752–3774, 2016.

[29] S. Chen and H. B. Nembhard, "A high-dimensional control chart for profile monitoring," *Quality and Reliability Engineering International*, vol. 27, no. 4, pp. 451–464, 2011.

[30] E. Chicken, J. J. Pignatiello Jr, and J. R. Simpson, "Statistical process monitoring of nonlinear profiles using wavelets," *Journal of Quality Technology*, vol. 41, no. 2, p. 198, 2009.

[31] G. Colavizza and M. Franceschet, "Clustering citation histories in the physical review," *Journal of Informetrics*, vol. 10, no. 4, pp. 1037–1051, 2016.

[32] R. Costas, T. N. van Leeuwen, and A. F. van Raan, "Is scientific literature subject to a sell-by-date? a general methodology to analyze the durability of scientific documents," *Journal of the American Society for Information Science and Technology*, vol. 61, no. 2, pp. 329–339, 2010.

[33] I. Daubechies, *Ten lectures on wavelets*. SIAM, 1992.

[34] J. E. Dennis Jr and R. E. Welsch, "Techniques for nonlinear least squares and robust regression," *Communications in Statistics-Simulation and Computation*, vol. 7, no. 4, pp. 345–359, 1978.

[35] F. Desobry, M. Davy, and C. Doncarli, "An online kernel change detection algorithm," *IEEE Transactions on Signal Processing*, vol. 53, no. 8, pp. 2961–2974, 2005.

[36] D. L. Donoho and P. J. Huber, "The notion of breakdown point," *A festschrift for Erich L. Lehmann*, pp. 157–184, 1983.

[37] D. L. Donoho and I. M. Johnstone, "Ideal spatial adaptation by wavelet shrinkage," *Biometrika*, vol. 81, no. 3, pp. 425–455, 1994.

[38] ——, "Adapting to unknown smoothness via wavelet shrinkage," *Journal of the American Statistical Association*, vol. 90, no. 432, pp. 1200–1224, 1995.

[39] ——, "Minimax estimation via wavelet shrinkage," *The Annals of Statistics*, vol. 26, no. 3, pp. 879–921, 1998.

[40] B. Efron, *Large-scale inference: Empirical bayes methods for estimation, testing, and prediction*. Cambridge University Press, 2012, vol. 1.

[41] N. El Karoui, D. Bean, P. J. Bickel, C. Lim, and B. Yu, "On robust regression with high-dimensional predictors," *Proceedings of the National Academy of Sciences*, vol. 110, no. 36, pp. 14 557–14 562, 2013.

[42] J. Fan, "Test of significance based on wavelet thresholding and neyman's truncation," *Journal of the American Statistical Association*, vol. 91, no. 434, pp. 674–688, 1996.

[43] J. Fan and S.-K. Lin, "Test of significance when data are curves," *Journal of the American Statistical Association*, vol. 93, no. 443, pp. 1007–1021, 1998.

[44] D. Ferrari and Y. Yang, "Maximum lq-likelihood estimation," *The Annals of Statistics*, vol. 38, no. 2, pp. 753–783, 2010.

[45] C.-D. Fuh and Y. Mei, "Quickest change detection and kullback-leibler divergence for two-state hidden markov models," *IEEE Transactions on Signal Processing*, vol. 63, no. 18, pp. 4866–4878, 2015.

[46] M. M. Gardner, J.-C. Lu, R. S. Gyurcsik, J. J. Wortman, B. E. Hornung, H. H. Heinisch, E. A. Rying, S. Rao, J. C. Davis, and P. K. Mozumder, "Equipment fault detection using spatial signatures," *IEEE Transactions on Components, Packaging, and Manufacturing Technology: Part C*, vol. 20, no. 4, pp. 295–304, 1997.

[47] E. Garfield, "Premature discovery or delayed recognition-why," *Current Contents*, vol. 21, pp. 5–10, 1980.

[48] W. Glänzel and U. Schoepflin, "A bibliometric study on ageing and reception processes of scientific literature," *Journal of Information Science*, vol. 21, no. 1, pp. 37–53, 1995.

[49] J. Glaz, J. I. Naus, S. Wallenstein, S. Wallenstein, and J. I. Naus, *Scan statistics*. Springer, 2001.

[50] L. Gordon and M. Pollak, "An efficient sequential nonparametric scheme for detecting a change of distribution," *The Annals of Statistics*, vol. 22, pp. 763–804, 1994.

[51] ——, "A robust surveillance scheme for stochastically ordered alternatives," *The Annals of Statistics*, vol. 23, pp. 1350–1375, 1995.

[52] P. Z. Hadjipantelis, J. A. Aston, and J. P. Evans, "Characterizing fundamental frequency in mandarin: A functional principal component approach utilizing mixed effect models," *The Journal of the Acoustical Society of America*, vol. 131, no. 6, pp. 4651–4664, 2012.

[53] P. Hall, H. G. Müller, and J. L. Wang, "Properties of principal component methods for functional and longitudinal data analysis," *The Annals of Statistics*, vol. 34, pp. 1493–1517, 2006.

[54] P. Hall, D. S. Poskitt, and B. Presnell, "A functional dataanalytic approach to signal discrimination," *Technometrics*, vol. 43, no. 1, pp. 1–9, 2001.

[55] F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel, *Robust statistics: The approach based on influence functions*. John Wiley & Sons, 2011.

[56] F. R. Hampel, "Contributions to the theory of robust estimation," PhD thesis, University of California Berkeley, 1968.

[57] S. Heritier and E. Ronchetti, "Robust bounded-influence tests in general parametric models," *Journal of the American Statistical Association*, vol. 89, no. 427, pp. 897–904, 1994.

[58] D. R. Hoover, J. A. Rice, C. O. Wu, and L. P. Yang, "Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data," *Biometrika*, vol. 85, no. 4, pp. 809–822, 1998.

[59] P. J. Huber, "Robust estimation of a location parameter," *The Annals of Mathematical Statistics*, vol. 35, no. 1, pp. 73–101, 1964.

[60] ——, "A robust version of the probability ratio test," *The Annals of Mathematical Statistics*, vol. 36, no. 6, pp. 1753–1758, 1965.

[61] P. J. Huber and E Ronchetti, *Robust statistics*, 2nd ed. New York: Wiley, 2009.

[62] W. James and C. Stein, "Estimation with quadratic loss," in *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, vol. 1, 1961, pp. 361–379.

[63] M. K. Jeong, J.-C. Lu, and N. Wang, "Wavelet-based spc procedure for complicated functional data," *International Journal of Production Research*, vol. 44, no. 4, pp. 729–744, 2006.

[64] J. Jin and J. Shi, "Feature-preserving data compression of stamping tonnage information using wavelets," *Technometrics*, vol. 41, no. 4, pp. 327–339, 1999.

[65] ——, "Automatic feature extraction of waveform signals for in-process diagnostic performance improvement," *Journal of Intelligent Manufacturing*, vol. 12, no. 3, pp. 257–268, 2001.

[66] L. Kang and S. L. Albin, "On-line monitoring when the process yields a linear profile," *Journal of quality Technology*, vol. 32, no. 4, p. 418, 2000.

[67] R. B. Kazemzadeh, R. Noorossana, and A. Amiri, "Phase i monitoring of polynomial profiles," *Communications in Statistics–Theory and Methods*, vol. 37, no. 10, pp. 1671–1686, 2008.

[68] Q. Ke, E. Ferrara, F. Radicchi, and A. Flammini, "Defining and identifying sleeping beauties in science," *Proceedings of the National Academy of Sciences*, vol. 112, no. 24, pp. 7426–7431, 2015.

[69] J Kiefer and J Sacks, "Asymptotically optimum sequential inference and design," *The Annals of Mathematical Statistics*, pp. 705–750, 1963.

[70] M. H. Kim, M. G. Akritas, *et al.*, "Order thresholding," *The Annals of Statistics*, vol. 38, no. 4, pp. 2314–2350, 2010.

[71] W. S. Krasker and R. E. Welsch, "Efficient bounded-influence regression estimation," *Journal of the American statistical Association*, vol. 77, no. 379, pp. 595–604, 1982.

[72] M. Kulldorff, "Prospective time periodic geographical disease surveillance using a scan statistic," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, vol. 164, no. 1, pp. 61–72, 2001.

[73] T. L. Lai, "Sequential changepoint detection in quality control and dynamical systems," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 613–658, 1995.

[74] ——, "Sequential analysis: Some classical problems and new challenges," *Statistica Sinica*, pp. 303–351, 2001.

[75] S. Lambert-Lacroix and L. Zwald, "Robust regression through the hubers criterion and adaptive lasso penalty," *Electronic Journal of Statistics*, vol. 5, pp. 1015–1053, 2011.

[76] V. Larivière, É. Archambault, and Y. Gingras, "Long-term variations in the aging of scientific literature: From exponential growth to steady-state science (1900–2004),"

Journal of the American Society for Information Science and technology, vol. 59, no. 2, pp. 288–296, 2008.

[77] J. Lee, Y. Hur, S.-H. Kim, and J. R. Wilson, "Monitoring nonlinear profiles using a wavelet-based distribution-free cusum chart," *International Journal of Production Research*, vol. 50, no. 22, pp. 6574–6594, 2012.

[78] E. L. Lehmann and G. Casella, *Theory of point estimation*. Springer Science & Business Media, 2006.

[79] Y. Lei, Z. Zhang, and J. Jin, "Automatic tonnage monitoring for missing part detection in multi-operation forging processes," *Journal of Manufacturing Science and Engineering*, vol. 132, no. 5, p. 051 010, 2010.

[80] X. Leng and H. G. Müller, "Classification using functional data analysis for temporal gene expression data," *Bioinformatics*, vol. 22, no. 1, pp. 68–76, 2006.

[81] G. Li, H. Peng, and L. Zhu, "Nonconcave penalized m-estimation with a diverging number of parameters," *Statistica Sinica*, vol. 21, pp. 391–419, 2011.

[82] A. van der Linde, "A bayesian latent variable approach to functional principal components analysis with binary and count data," *Advances in Statistical Analysis*, vol. 93, no. 3, pp. 307–333, 2009.

[83] M. B. Line, "Changes in the use of literature with time-obsolescence revisited," *Library Trends*, vol. 41, no. 4, pp. 665–684, 1993.

[84] K. Liu, Y. Mei, and J. Shi, "An adaptive sampling strategy for online high-dimensional process monitoring," *Technometrics*, vol. 57, no. 3, pp. 305–319, 2015.

[85] K. Liu, R. Zhang, and Y. Mei, "Scalable sum-shrinkage schemes for distributed monitoring large-scale data streams," *Statistica Sinica*, vol. 29, pp. 1–22, 2019.

[86] P.-L. Loh, "Statistical consistency and asymptotic normality for high-dimensional robust m-estimators," *The Annals of Statistics*, vol. 45, no. 2, pp. 866–896, 2017.

[87] G. Lorden, "Procedures for reacting to a change in distribution," *The Annals of Mathematical Statistics*, vol. 42, no. 6, pp. 1897–1908, 1971.

[88] G. Lorden and M. Pollak, "Sequential change-point detection procedures that are nearly optimal and computationally simple," *Sequential Analysis*, vol. 27, no. 4, pp. 476–512, 2008.

[89] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online dictionary learning for sparse coding," in *Proceedings of the 26th annual international conference on machine learning*, ACM, 2009, pp. 689–696.

[90] S. Mallat, *A wavelet tour of signal processing*. Academic press, 1999.

[91] S. G. Mallat, "Multifrequency channel decompositions of images and wavelet models," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 12, pp. 2091–2110, 1989.

[92] R. A. Maronna and V. J. Yohai, "Asymptotic behavior of general m-estimates for regression and scale with random carriers," *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, vol. 58, no. 1, pp. 7–20, 1981.

[93] S. Mei, Y. Bai, A. Montanari, *et al.*, "The landscape of empirical risk for nonconvex losses," *The Annals of Statistics*, vol. 46, no. 6A, pp. 2747–2774, 2018.

[94] Y. Mei, "Efficient scalable schemes for monitoring a large number of data streams," *Biometrika*, vol. 97, no. 2, pp. 419–433, 2010.

[95] ——, "Quickest detection in censoring sensor networks," in *2011 IEEE International Symposium on Information Theory Proceedings*, IEEE, 2011, pp. 2148–2152.

[96] Y. Mei, "Information bounds and quickest change detection in decentralized decision systems," *IEEE Transactions on Information theory*, vol. 51, no. 7, pp. 2669–2681, 2005.

[97] R. K. Merton, *The sociology of science: Theoretical and empirical investigations*. University of Chicago press, 1973.

[98] I. Mizera and C. H. Müller, "Breakdown points and variation exponents of robust m-estimators in linear models," *The Annals of Statistics*, vol. 27, no. 4, pp. 1164–1177, 1999.

[99] G. V. Moustakides, "Optimal stopping times for detecting changes in distributions," *The Annals of Statistics*, vol. 14, no. 4, pp. 1379–1387, 1986.

[100] J. A. Nelder and R. W. Wedderburn, "Generalized linear models," *Journal of the Royal Statistical Society: Series A (General)*, vol. 135, no. 3, pp. 370–384, 1972.

[101] J. Neyman, "Smooth test for goodness of fit," *Scandinavian Actuarial Journal*, vol. 20, no. 3-4, pp. 149–199, 1937.

[102] E. S. Page, "Continuous inspection schemes," *Biometrika*, vol. 41, no. 1/2, pp. 100–115, 1954.

[103] K. Paynabar, C. Zou, and P. Qiu, "A change-point approach for phase-i analysis in multivariate profile monitoring and diagnosis," *Technometrics*, vol. 58, no. 2, pp. 191–204, 2016.

[104] M. Pollak *et al.*, "Optimal detection of a change in distribution," *The Annals of Statistics*, vol. 13, no. 1, pp. 206–227, 1985.

[105] ——, "Average run lengths of an optimal method of detecting a change in distribution," *The Annals of Statistics*, vol. 15, no. 2, pp. 749–779, 1987.

[106] H. V. Poor and O. Hadjiliadis, *Quickest detection*. Cambridge University Press Cambridge, 2009, vol. 40.

[107] Y. Qin and C. E. Priebe, "Robust hypothesis testing via lq-likelihood," *Statistica Sinica*, vol. 27, no. 4, pp. 1793–1813, 2017.

[108] P. Qiu, C. Zou, and Z. Wang, "Nonparametric profile monitoring by mixed effects modeling," *Technometrics*, vol. 52, no. 3, pp. 265–277, 2010.

[109] A. F. van Raan, "Sleeping beauties in science," *Scientometrics*, vol. 59, no. 3, pp. 467–472, 2004.

[110] C. Rago, P. Willett, and Y. Bar-Shalom, "Censoring sensors: A low-communication-rate scheme for distributed detection," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 32, no. 2, pp. 554–568, 1996.

[111] J. O. Ramsay, "Estimating smooth monotone functions," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 60, no. 2, pp. 365–375, 1998.

[112] J. O. Ramsay, G. Hooker, and S. Graves, *Functional data analysis with r and matlab*. Springer, 2009.

[113] J. O. Ramsay and B. W. Silverman, *Functional data analysis*. New York:Springer, 2005.

[114] J. Ramsay and B. Silverman, *Functional data analysis*, 1997.

[115] S Redner, "Citation statistics from 110 years of physical review," *Physics Today*, vol. 58, no. 6, pp. 49–54, 2005.

[116] W. J. Rey, *Introduction to robust and quasi-robust statistical methods*. Springer Science & Business Media, 2012.

[117] J. A. Rice and B. W. Silverman, "Estimating the mean and covariance structure nonparametrically when the data are curves," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 53, pp. 233–243, 1991.

[118] Y Ritov, "Decision theoretic optimality of the cusum procedure," *The Annals of Statistics*, vol. 18, no. 3, pp. 1464–1469, 1990.

[119] S. Roberts, "A comparison of some control chart procedures," *Technometrics*, vol. 8, no. 3, pp. 411–430, 1966.

[120] P. J. Rousseeuw, "Least median of squares regression," *Journal of the American statistical Association*, vol. 79, no. 388, pp. 871–880, 1984.

[121] N. Serban, A. M. Staicu, and R. J. Carroll, "Multilevel cross-dependent binary longitudinal data," *Biometrics*, vol. 69, no. 4, pp. 903–913, 2013.

[122] H. Shen, D. Wang, C. Song, and A.-L. Barabási, "Modeling and predicting popularity dynamics via reinforced poisson processes," in *Twenty-eighth AAAI conference on artificial intelligence*, 2014.

[123] A. N. Shiryaev, "On optimum methods in quickest detection problems," *Theory of Probability & Its Applications*, vol. 8, no. 1, pp. 22–46, 1963.

[124] G. Shmueli and H. Burkom, "Statistical challenges facing early outbreak detection in biosurveillance," *Technometrics*, vol. 52, no. 1, pp. 39–51, 2010.

[125] D. Siegmund, *Sequential analysis: Tests and confidence intervals*. Springer, New York, 1985.

[126] G. S. Stent, "Prematurity and uniqueness in scientific discovery," *Scientific American*, vol. 227, no. 6, pp. 84–93, 1972.

[127] A. Tartakovsky, I. Nikiforov, and M. Basseville, *Sequential analysis: Hypothesis testing and changepoint detection*. CRC Press, 2014.

[128] A. G. Tartakovsky, A. S. Polunchenko, and G. Sokolov, "Efficient computer network anomaly detection by changepoint detection methods," *IEEE Journal of Selected Topics in Signal Processing*, vol. 7, no. 1, pp. 4–11, 2013.

[129] A. G. Tartakovsky, B. L. Rozovskii, R. B. Blažek, and H. Kim, "Detection of intrusions in information systems by sequential change-point methods," *Statistical methodology*, vol. 3, no. 3, pp. 252–293, 2006.

[130] A. G. Tartakovsky and V. V. Veeravalli, "Change-point detection in multichannel and distributed systems," *Applied Sequential Methodologies: Real-World Examples with Data Analysis*, vol. 173, pp. 339–370, 2004.

[131] ——, "Asymptotically optimal quickest change detection in distributed sensor systems," *Sequential Analysis*, vol. 27, no. 4, pp. 441–475, 2008.

[132] W. P. Tay, J. N. Tsitsiklis, and M. Z. Win, "Asymptotic performance of a censoring sensor network," *IEEE Transactions on Information Theory*, vol. 53, no. 11, pp. 4191–4209, 2007.

[133] N. A. Thacker and P. A. Bromiley, "The effects of a square root transform on a poisson distributed quantity," *Tina memo*, vol. 10, p. 2001, 2001.

[134] J. Unnikrishnan, V. V. Veeravalli, and S. P. Meyn, "Minimax robust quickest change detection," *IEEE Transactions on Information Theory*, vol. 57, no. 3, pp. 1604–1614, 2011.

[135] V. V. Veeravalli, "Decentralized quickest change detection," *IEEE Transactions on Information theory*, vol. 47, no. 4, pp. 1657–1665, 2001.

[136] D. Wang, C. Song, and A. L. Barabási, "Quantifying long-term scientific impact," *Science*, vol. 342, no. 6154, pp. 127–132, 2013.

[137] J.-L. Wang, J.-M. Chiou, and H.-G. Mueller, "Review of functional data analysis," *ArXiv preprint arXiv:1507.05135*, 2015.

[138] J. Wang, "Unpacking the matthew effect in citations," *Journal of Informetrics*, vol. 8, no. 2, pp. 329–339, 2014.

[139] J. Wang, Y. Mei, and D. Hicks, "Comment on "quantifying long-term scientific impact"," *Science*, vol. 345, no. 6193, pp. 149–149, 2014.

[140] X. Wang, Y. Jiang, M. Huang, and H. Zhang, "Robust variable selection with exponential squared loss," *Journal of the American Statistical Association*, vol. 108, no. 502, pp. 632–643, 2013.

[141] Y. Wang and Y. Mei, "Large-scale multi-stream quickest change detection via shrinkage post-change estimation," *IEEE Transactions on Information Theory*, vol. 61, no. 12, pp. 6926–6938, 2015.

[142] Y. Wang, Y. Mei, and K. Paynabar, "Thresholded multivariate principal component analysis for phase i multichannel profile monitoring," *Technometrics*, vol. 60, no. 3, pp. 360–372, 2018.

[143] S. Wu, H. G. Müller, and Z. Zhang, "Functional data analysis for point processes with rare events," *Statistica Sinica*, vol. 23, pp. 1–23, 2013.

[144] Y. Xie and D. Siegmund, "Sequential multi-sensor change-point detection," *The Annals of Statistics*, vol. 41, no. 2, pp. 670–692, 2013.

[145] H. Yan, K. Paynabar, and J. Shi, "Image-based process monitoring using low-rank tensor decomposition," *IEEE Transactions on Automation Science and Engineering*, vol. 12, no. 1, pp. 216–227, 2015.

[146] F. Yao, H. G. Müller, and J. L. Wang, "Functional data analysis for sparse longitudinal data," *Journal of the American Statistical Association*, vol. 100, no. 470, pp. 577–590, 2005.

[147] V. J. Yohai, "High breakdown-point and high efficiency robust estimates for regression," *The Annals of Statistics*, vol. 15, pp. 642–656, 1987.

[148] R. Zhang, Y. Mei, and J. Shi, "Wavelet-based profile monitoring using order-thresholding recursive cusum schemes," in *New Frontiers of Biostatistics and Bioinformatics*, Springer, 2018, pp. 141–159.

[149] C. Zhou, K. Liu, X. Zhang, W. Zhang, and J. Shi, "An automatic process monitoring method using recurrence plot in progressive stamping processes," *IEEE Transactions on Automation Science and Engineering*, vol. 13, no. 2, pp. 1102–1111, 2016.

[150] S. Zhou, B. Sun, and J. Shi, "An spc monitoring system for cycle-based waveform signals using haar transform," *IEEE Transactions on Automation Science and Engineering*, vol. 3, no. 1, pp. 60–72, 2006.

[151] C. Zou and P. Qiu, "Multivariate statistical process control using lasso," *Journal of the American Statistical Association*, vol. 104, no. 488, pp. 1586–1596, 2009.

[152] C. Zou, P. Qiu, and D. Hawkins, "Nonparametric control chart for monitoring profiles using change point formulation and adaptive smoothing," *Statistica Sinica*, vol. 19, no. 3, pp. 1337–1357, 2009.

[153] C. Zou, F. Tsung, and Z. Wang, "Monitoring general linear profiles using multivariate exponentially weighted moving average schemes," *Technometrics*, vol. 49, no. 4, pp. 395–408, 2007.

[154] C. Zou, Z. Wang, X. Zi, and W. Jiang, "An efficient online monitoring method for high-dimensional data streams," *Technometrics*, vol. 57, no. 3, pp. 374–387, 2015.

[155] C. Zou, C. Zhou, Z. Wang, and F. Tsung, "A self-starting control chart for linear profiles," *Journal of Quality Technology*, vol. 39, no. 4, pp. 364–375, 2007.

[156] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the royal statistical society: Series B (statistical methodology)*, vol. 67, no. 2, pp. 301–320, 2005.