STATISTICAL SELECTION AND WAVELET-BASED PROFILE MONITORING

A Thesis
Presented to
The Academic Faculty

by

Huizhu Wang

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
May 2015

STATISTICAL SELECTION AND WAVELET-BASED PROFILE MONITORING

Approved by:

Professor Seong-Hee Kim, Advisor
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology

Professor Xiaoming Huo
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology

Professor Jianjun Shi
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology

Professor James R. Wilson
Department of Industrial and Systems Engineering
North Carolina State University

Professor Youngmi Hur
Department of Mathematics
Yonsei University

Date Approved: 13 March 2015

ACKNOWLEDGEMENTS

First and foremost I would like to thank my advisor Dr. Seong-Hee Kim for her guidance during my years at Georgia Tech. She has been exceptionally generous with her time and support. With her warm encouragement, I keep pursuing my path in research. I always consider Dr. Kim as my role model, who shows me that persistence, patience and passionateness are essential key factors to being a successful researcher. What I have learned from Dr. Kim will be my invaluable treasures now and in the future.

Next, I would like to thank Dr. James R. Wilson and Dr. Youngmi Hur for their insight suggestions on quantile estimation and discrete wavelet transform. Their helpful comments make the work strong. I also highly appreciate discussions with Dr. Xiaoming Huo on the wavelet thresholding methods. Dr. Huo's input is very informative and inspiring, and also helps to improve the quality of the work. Besides, I would like to thank Dr. Jianjun Shi for his service in my doctoral committee and his consistent contact.

Additionally I would like to thank all of the excellent professors at Georgia Tech, from whom I have had excellent courses and influenced by their enthusiasm of research.

In the end, I would also like to thank Dan Steffy and my dear parents for their support and encouragement through my time at Georgia Tech, especially during difficult periods. I have been lucky to be at Georgia Tech and surrounded by many incredible people as my friends. To name a few: Naragain (Jan) Phumchusri, Jon Peterson, Haiyue Yu, Yaxian Li, Seonghye Jeon, Helder Inacio, Lulu Kang, Weijun Ding and Jia Yan.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

This thesis consists of two topics: statistical selection and profile monitoring. Statistical selection is related to ranking and selection in simulation and profile monitoring is related to statistical process control.

Ranking and selection (R&S) is to select a system with the largest or smallest performance measure among a finite number of simulated alternatives with some guarantee about correctness. Fully sequential procedures have been shown to be efficient, but their actual probabilities of correct selection tend to be higher than the nominal level, implying that they consume unnecessary observations. In the first part, we study three conservativeness sources in fully sequential indifference-zone (IZ) procedures and use experiments to quantify the impact of each source in terms of the number of observations, followed by an asymptotic analysis on the impact of the critical one. Then we propose new asymptotically valid procedures that lessen the critical conservativeness source, by mean update with or without variance update. Experimental results showed that new procedures achieved meaningful improvement on the efficiency.

The second part is developing a wavelet-based distribution-free tabular CUSUM chart based on adaptive thresholding. WDFTC$_a$ is designed for rapidly detecting shifts in the mean of a high-dimensional profile whose noise components have a continuous nonsingular multivariate distribution. First computing a discrete wavelet transform of the noise vectors for randomly sampled Phase I (in-control) profiles, WDFTC$_a$ uses a matrix-regularization method to estimate the covariance matrix of the wavelet-transformed noise vectors; then those vectors are aggregated (batched) so that the nonoverlapping batch means of the wavelet-transformed noise vectors have manageable covariances. Lower and upper in-control thresholds are computed for the resulting batch means of the wavelet-transformed noise vectors using the associated marginal Cornish-Fisher expansions that have been suitably adjusted for between-component correlations. From the thresholded batch means of the wavelet-transformed noise vectors, Hotelling's T^2-type statistics are computed to set the parameters of a CUSUM procedure. To monitor shifts in the mean profile during Phase II (regular) operation, WDFTC$_a$ computes a similar Hotelling's T^2-type statistic from successive thresholded batch means of the wavelet-transformed noise vectors using the in-control thresholds; then WDFTC$_a$ applies the CUSUM procedure to the resulting T^2-type statistics. Experimentation with several normal and nonnormal test processes revealed that WDFTC$_a$ outperformed existing nonadaptive profile-monitoring schemes.

CHAPTER I

INTRODUCTION

This thesis consists of two topics: statistical selection and profile monitoring. Statistical selection is related to ranking and selection in simulation and the other one is related to statistical process control. In this chapter, we provide an overview and brief literature review for each topic.

Ranking and selection (R&S) is to select a system with the largest or smallest performance measure among a finite number of simulated alternatives with some guarantee about a correct selection. Bechhofer [3] proposes the indifference-zone (IZ) approach, in which a decision maker is assumed to care about a practically meaningful difference only. Many statistical selection procedures are developed under the IZ approach. Kim and Nelson [30] provide an extensive review of existing R&S IZ procedures.

Among many IZ procedures, fully sequential IZ procedures take only one basic observation from each active system at each stage, possibly with some elimination steps for clearly inferior systems. For example, Procedures $\mathcal{KN}$ [28] and $\mathcal{KN}$++ [29] are fully sequential IZ procedures with elimination. In both [28] and [29], Kim and Nelson show that fully sequential IZ procedures are efficient compared to other existing IZ procedures in terms of the total number of observations required to make a selection. However, their probabilities of correct selection (PCS) are often higher than the nominal level, implying that unnecessary observations are consumed, especially when the number of systems is large. The conservativeness of fully sequential procedures reduces the efficiency of procedures, limits the number of competing alternatives, and further restricts their applications to optimization via simulation (OvS) in assisting a promising region/solution search [39].

In this thesis, we first identify conservativeness sources of fully sequential IZ procedures, by studying the proof of a simple fully sequential procedure, Procedure ($\mathcal{P}$), that is generalized from Paulson's procedure in [37], with known unequal variances across systems.

Then we use experiments to quantify the impact of each source in terms of the number of observations and determine the critical one. Additionally, asymptotic analysis reveals the critical conservativeness sources degrades the efficiency significantly for a large number of systems. Many fully sequential IZ procedures for R&S are extended from Procedure ($\mathcal{P}$), therefore our conclusions of conservativeness sources provide supportive information for other researchers to further improve the performance of procedures.

Then we develop two practically useful procedures that remove the critical source in their derivations and show their asymptotic validity [48]. Experimental results show that new procedures perform better than their competitor procedures on the efficiency.

The second topic of the thesis is on profile monitoring. Profile monitoring arises in statistical process control (SPC) for high-dimensional profiles. A profile describes a functional relationship between a response variable and one or more explanatory variables [51]. The functional relationship can be linear, nonlinear or too complicated to have an explicit form. In this thesis, we are interested in developing a distribution-free control chart for profiles with complicated functional relationships in manufacturing processes, where there is one explanatory variable, the same setting as [13, 23, 33]. As an example, consider a lumber manufacturing process. In order to control the quality of sawed boards, we monitor the thickness along a horizontal or vertical line from an edge of a sawed board. Since lumbering is a complicated process and affected by a combination of saw configurations, board types, and sensor locations, no explicit function can properly describe the relationship between the distance from the edge of the sawed board and its associated thickness. Therefore we describe the complicated relationship by taking a large number of pairs of points (distance from the edge and thickness), which results in a high-dimensional profile.

In a manufacturing process, SPC charts are often used to monitor profiles and determine whether the manufacturing process is in a state of statistical control. Montgomery [36] provides a comprehensive background of SPC charts. Some SPC charts have been developed with the aids from nonparametric methods, for example, smoothing spline [19], functional principal component analysis [41], Fourier transform [12], and discrete wavelet transform

(DWT) [13, 23, 24, 25, 33]. Among all nonparametric methods, DWT is a popular technique to incorporate with SPC charts, since it is a multi-resolution decomposition and has nice properties in dimension reduction and denoising. Moreover, DWT can preserve local properties, such as, sharps and jumps, which contain important local information. Most wavelet-based SPC charts first compute a profile's DWT or a noise vector's DWT, and next construct statistics from a subset of wavelet components, see [33]. The subset size is much smaller than the original profile's dimension, therefore these wavelet-based SPC charts avoid losing power in detecting shifts caused by high dimensionality, see [16].

Noises may exhibit non-normality, variance heterogeneity, or correlation between profile components, see [20, 33, 51]. Existing wavelet-based distribution-free control charts only monitor a static subset of wavelet components, whose locations are preselected and fixed, so they are not sensitive to some types of shifts, such as a shift whose DWT only affect unselected wavelet components. Therefore we are interested in developing a new wavelet-based distribution-free control chart with adaptive selection.

The new procedure [49], WDFTC$_a$, first applies DWT on a profile or its noise vector. Then it employs an adaptive thresholding methods to select a small number of wavelet components when in-control and more components when out-of-control. In several high-dimensional normal and non-normal test processes, WDFTC$_a$ outperforms its competitors in detecting various shifts.

The thesis is organized as follows. Chapter 2 studies conservativeness sources of fully sequential indifference-zone (IZ) procedures and proposes new procedures. Chapter 3 proposes a new wavelet-based distribution-free CUSUM chart with adaptive selection for profile monitoring, and show its experimental results. At the end, we summarize our contributions in Chapter 4.

CHAPTER II

REDUCING THE CONSERVATIVENESS OF FULLY SEQUENTIAL INDIFFERENCE-ZONE PROCEDURES

In order to find a system with the largest or smallest performance measure among a finite number of simulated systems, some R&S procedures take the IZ approach, see [3], where the decision maker is assumed to be indifferent to systems whose true means are within a practically meaningful difference, $\delta > 0$, from the best performance. The positive constant δ is called the IZ parameter. Fully sequential IZ procedures are one of efficient R&S procedures.

Among R&S IZ procedures, sequential procedures keep the number of stages small, say 1, 2 or 3, where a stage occurs whenever a simulation of a system is initiated to obtain observations. For example, see [42] for a sequential procedure with two stages. On the other hand, fully sequential procedures take only one basic observation from each active system at each stage and tend to have a large number of stages, but they are more efficient compared to other existing IZ procedures in terms of the total number of observations required to reach a decision, see [28, 29]. However, their PCS are often higher than the nominal level, implying that procedures are conservative and unnecessary observations are consumed.

Our contributions are in two aspects: (i) providing an insight on how each conservativeness source affects the performance of fully sequential procedures and which source is critical under various configurations, and (ii) proposing new improved fully sequential IZ procedures. Therefore, we first discuss three sources of conservativeness of fully sequential IZ procedures, use experiments to compare the number of additional observations overused due to each source, and identify which source is critical. Our experiments show that the slippage means configuration (SC) assumption, where all inferior systems' means are assumed to be exactly δ apart from the best, degrades the efficiency of fully sequential procedures the most. Then we present new asymptotically valid procedures that lessen conservativeness due to the SC assumption. The main idea of the new procedures is to replace the IZ parameter δ with a function of estimated mean differences between systems. A similar idea is used in sequential procedures [9, 10, 11]. Our procedures are different in that: (i) their function is related to a confidence interval for the best system's mean, while our function is related to a confidence interval of the mean difference between a pair of systems; (ii) their procedures are sequential while ours are fully sequential; and (iii) both procedures are asymptotically valid as δ goes to zero with a finite first-stage sample size, but the proofs are done under different assumptions. Their asymptotical validity is shown under the assumption that means are known while our procedures do not require this assumption.

This chapter is organized as follows: Section 2.1 defines a R&S problem with the IZ approach, sets up notation and assumptions, and discusses sources of conservativeness in fully sequential procedures. In Section 2.2, we quantify the impact of each source in terms of the number of additional observations by experiments and provide asymptotic analysis on the critical source, the SC assumption. Section 2.3 provides new asymptotically valid procedures that lessen the impact of the SC assumption and Section 2.4 presents experimental results of the new procedures. Concluding remarks are followed in Section 2.5.

2.1 Background

In this section, we first formulate a R&S problem with the IZ approach and set up notation and assumptions. We then explain why a statistically valid fully sequential procedure becomes conservative, using a simple procedure developed for known variances.

2.1.1 Problem, Notation and Assumptions

We assume that there are k simulated systems ($k \geq 2$). Let X_{ij} be the jth observation from system i for $i = 1, \ldots, k$ and $j = 1, 2, \ldots$. We let $\mu_i = \mathrm{E}[X_{ij}]$ and $\sigma_i^2 = \mathrm{Var}[X_{ij}]$ be the expected performance measure and marginal variance of system i, respectively. Observations X_{ij} and $X_{\ell j}$ from systems i and ℓ, $i \neq \ell$, can be correlated due to the use of common random numbers (CRN). Further, let $\boldsymbol{X}_j = (X_{1j}, X_{2j}, \ldots, X_{kj})^T$ denote a vector of jth observations for all k systems where $\boldsymbol{A}^T$ represents the transpose of a vector or matrix $\boldsymbol{A}$.

Further, let $\mathbb{R}^+$ and $\mathbb{Z}^+$ represent the sets of all positive real numbers and all positive integers, respectively. Our objective is to find a system with the largest performance measure μ_i with a pre-specified nominal level $1 - \alpha$.

The following assumptions are assumed to hold throughout this Chapter.

Assumption 1 $\boldsymbol{X}_j \overset{\text{IID}}{\sim} \mathbf{MN}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), j = 1, 2, \ldots,$ *where* $\overset{\text{IID}}{\sim}$ *represents 'are independent and identically distributed as',* $\mathbf{MN}$ *denotes multivariate normal,* $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_k)^T$*, and* $\boldsymbol{\Sigma}$ *is a positive definite* $k \times k$ *covariance matrix.*

Assumption 2 $\mu_1 - \delta \geq \mu_2 \geq \ldots \geq \mu_{k-1} \geq \mu_k$ *for* $\delta \in \mathbb{R}^+$*. Also, there exist* $c_i \geq 1$ *such that* $\mu_1 - \mu_i = c_i\delta$ *for* $c_i \in \mathbb{R}^+$*,* $i = 2, \ldots, k$*.*

Additional notation is defined in Table 1 as below. Note that $\mathcal{I}(\cdot)$ represents the indicator function, and $n_0 \geq 2$ and r denote the initial sample size and the current number of observations, respectively.

Table 1: Notation summary.

$\delta_{i\ell}$	$\equiv \max(\delta, \lvert\mu_i - \mu_\ell\rvert)$;
$\hat{\delta}_{i\ell}(r)$	$\equiv \max(\delta, \lvert\bar{X}_i(r) - \bar{X}_\ell(r)\rvert - A_{i\ell}(r))$ where $A_{i\ell}(r)$ is a non-negative random variable discussed in Assumption 3;
$\sigma^2_{i\ell}$	$\equiv$Var$(X_{ij} - X_{\ell j})$, the variance of differences between systems i and ℓ;
$\bar{X}_i(r)$	$\equiv \frac{1}{r}\sum_{j=1}^{r} X_{ij}$, the sample mean of system i based on r observations;
$S^2_{i\ell}(r)$	$\equiv \frac{1}{r-1}\sum_{j=1}^{r}\left[X_{ij} - X_{\ell j} - (\bar{X}_i(r) - \bar{X}_\ell(r))\right]^2$, the sample variance of differences between systems i and ℓ based on r observations;
$R(r; a, b, d)$	$\equiv \max\left\{0, \frac{bd}{a} - \frac{a}{2e}r\right\}$ for any $a, b, d \in \mathbb{R}^+$ and for $e \in \mathbb{Z}^+$;
$g(\eta)$	$\equiv \sum_{v=1}^{e}(-1)^{v+1}\left(1 - \frac{1}{2}\mathcal{I}(v = e)\right)\exp\left\{\frac{-\eta(2e - v)v}{e}\right\}$ for $e \in \mathbb{Z}^+$;
$g'(\eta, r)$	$\equiv \sum_{v=1}^{e}(-1)^{v+1}\left(1 - \frac{1}{2}\mathcal{I}(v = e)\right)\left(1 + \frac{2\eta(2e - v)v}{e}\right)^{-(r-1)/2}$ for $e \in \mathbb{Z}^+$;
T_i^d	$\equiv \min\Bigl\{r : r \in \mathbb{Z}^+ \text{ and } \sum_{j=1}^{r}(X_{1j} - X_{ij}) \geq R(r; \delta, \eta, \sigma^2_{1i})\Bigr\}$,
T_i^c	$\equiv \min\Bigl\{t : t \in \mathbb{R}^+ \text{ and } \sigma_{1i}\mathcal{W}\left(t, \frac{\mu_1 - \mu_i}{\sigma_{1i}}\right) \geq R(t; \delta, \eta, \sigma^2_{1i})\Bigr\}$.

2.1.2 Sources of Conservativeness

It is known that R&S procedures with the IZ approach tend to be conservative especially when the number of systems is large. We first present a simple fully sequential procedure denoted as Procedure ($\mathcal{P}$), a generalization of Paulson's procedure [37] from known equal variances to known unequal variances across systems, review the validity proof, and discuss why the procedure becomes conservative. Procedure ($\mathcal{P}$) is described in Figure 1. Many fully sequential IZ procedures for R&S are extended from this simple procedure and thus share similar steps for their validity proofs.

Procedure ($\mathcal{P}$)

Setup: Select the nominal level $1-\alpha$ and the IZ parameter δ. Calculate η from $g(\eta) = \alpha/(k-1)$. Set $I = \{1, \ldots, k\}$. Take one observation from each system. Set $r = 1$ and go to **Screening**.

Screening: Set $I^{\text{old}} = I$. Let

$$I = \left\{ i : i \in I^{\text{old}} \text{ and } \sum_{j=1}^{r} (X_{ij} - X_{\ell j}) \geq -R(r; \delta, \eta, \sigma_{i\ell}^2), \forall \ell \in I^{\text{old}} \text{ and } \ell \neq i \right\}.$$

Stopping Rule: If $|I| = 1$, then stop and select the system whose index is in I as the best one. Otherwise, take one additional observation $X_{i,r+1}$ from each active system $i \in I$, set $r = r+1$ and go to **Screening**.

Figure 1: Algorithmic statement of Procedure ($\mathcal{P}$).

Two lemmas are needed to prove the validity of a number of procedures presented.

Lemma 1 *[22] Suppose that a continuation region $H(t)$ is $(-h(t), h(t))$ given by a non-negative function $h(t), t \geq 0$. Consider two processes: a continuous process $\{\mathcal{W}(t,\Delta), t \geq 0\}$ with $\Delta > 0$, and a discrete process obtained by observing $\mathcal{W}(t,\Delta)$ at a random, increasing sequence of times $\{t_i : i = 1, 2, \ldots\}$ taking values in a given countable set. Let $T_C = \inf\{t : \mathcal{W}(t,\Delta) \notin H(t)\}$ and $T_D = \inf\{t_i : \mathcal{W}(t_i,\Delta) \notin H(t_i)\}$, and assume that $T_D < \infty$ almost surely. Note that $T_D \geq T_C$. The error probabilities are*

$$\Pr\{\mathcal{E}_C\} \equiv \Pr\{\mathcal{W}(T_C,\Delta) \leq -h(T_C)\} = \Pr\{\mathcal{W}(T_C,\Delta) < 0\},$$

$$\Pr\{\mathcal{E}_D\} \equiv \Pr\{\mathcal{W}(T_D,\Delta) \leq -h(T_D)\} = \Pr\{\mathcal{W}(T_D,\Delta) < 0\}.$$

Consider an outcome $\{(b(t); t \geq 0), \{t_i\}\}$, where $b(t)$ is the path of a Brownian motion. Assume that the conditional distribution of $\{t_i\}$ given $\mathcal{W}(t,\Delta) = b(t), t \geq 0$, is the same as *the conditional distribution of* $\{t_i\}$ *given* $\mathcal{W}(t,\Delta) = -b(t)$, $t \geq 0$. *Under these conditions,*

$$\Pr\{\mathcal{E}_D\} \leq \Pr\{\mathcal{E}_C\}.$$

Lemma 2 *[15] Consider a standard Brownian motion process with drift* $\mathcal{W}(t,\Delta)$ *with* $\Delta > 0$ *and* $t \geq 0$. *Let* $h(t) = a - \gamma t$ *for some* $a > 0$ *and* $\gamma \geq 0$. *Let* $H(t)$ *denote the interval* $(-h(t), h(t))$ *(so that* $H(t) = \emptyset$ *when* $-h(t) \geq h(t)$*), and let* $T = \inf\{t : \mathcal{W}(t,\Delta) \notin H(t)\}$ *be the first time* $\mathcal{W}(t,\Delta)$ *does not fall in the triangular continuation region defined by* $(H(t); t \geq 0)$. *Finally, let* $\mathcal{E}$ *be the event* $\{\mathcal{W}(T,\Delta) \leq -h(T)$ *and* $H(T) \neq \emptyset$, *or* $\mathcal{W}(T,\Delta) \leq 0$ *and* $H(T) = \emptyset\}$. *If* $\gamma = \Delta/(2e)$ *for any positive integer* e, *then*

$$\Pr\{\mathcal{E}\} = \sum_{v=1}^{e} (-1)^{v+1} \left(1 - \frac{1}{2}\mathcal{I}(v = e)\right) \exp\{-2a\gamma(2e - v)v\}.$$

The proof of statistical validity of Procedure $(\mathcal{P})$ is as follows. Let an incorrect selection (ICS) denote the event that the best system (system 1) is eliminated. This event implies that the partial sum $\sum_{j=1}^{r}(X_{1j} - X_{ij})$, which is the cumulated differences between system 1 and some alternative system i, exits through the lower boundary $-R(r;\delta,\eta,\sigma_{1i}^2)$.

Note that T_i^d is the first passage time that a discrete process $\sum_{j=1}^{r}(X_{1j} - X_{ij})$ exits the continuation region $\left(-R(t;\delta,\eta,\sigma_{1i}^2), R(t;\delta,\eta,\sigma_{1i}^2)\right)$ for $t \geq 0$; and T_i^c is the first passage time that a continuous process, $\sigma_{1i}\mathcal{W}(t, \mu_1 - \mu_i)$, exits the same continuation region. Then

$$\begin{aligned}
\Pr\{\text{ICS}\} &= \Pr\left\{\sum_{j=1}^{T_i^d}(X_{1j} - X_{ij}) \leq -R(T_i^d;\delta,\eta,\sigma_{1i}^2) \text{ for any } i \neq 1\right\} && (1)\\
&\overset{\text{BI}}{\leq} \sum_{i=2}^{k} \Pr\left\{\sum_{j=1}^{T_i^d} \frac{(X_{1j} - X_{ij})}{\sigma_{1i}} \leq -\max\left\{0, \left(\frac{\eta\sigma_{1i}}{\delta} - \frac{\delta}{2e\sigma_{1i}}T_i^d\right)\right\}\right\} && (2)\\
&\overset{\text{DTC}}{\leq} \sum_{i=2}^{k} \Pr\left\{\mathcal{W}\left(T_i^c; \frac{\mu_1 - \mu_i}{\sigma_{1i}}\right) \leq -\max\left\{0, \left(\frac{\eta\sigma_{1i}}{\delta} - \frac{\delta}{2e\sigma_{1i}}T_i^c\right)\right\}\right\} && (3)\\
&\overset{\text{SC}}{\leq} \sum_{i=2}^{k} \Pr\left\{\mathcal{W}\left(T_i^c; \frac{\delta}{\sigma_{1i}}\right) \leq -\max\left\{0, \left(\frac{\eta\sigma_{1i}}{\delta} - \frac{\delta}{2e\sigma_{1i}}T_i^c\right)\right\}\right\} && (4)\\
&= \sum_{i=2}^{k} \left[\sum_{\ell=1}^{e} (-1)^{\ell+1} \left(1 - \frac{1}{2}\mathcal{I}(\ell = e)\right) \exp\left\{-\frac{\eta(2e - \ell)\ell}{e}\right\}\right] && (5)\\
&= \sum_{i=2}^{k} g(\eta) = \alpha, && (6)
\end{aligned}$$

where the first inequality (2) is due to the BI; the second inequality (3) employs Lemma 1, denoted as DTC; the third inequality (4) holds because the smallest mean difference between the best system and any inferior system i is δ under the SC assumption; and Equation (5) is due to Lemma 2. Since we choose η such that $g(\eta) = \frac{\alpha}{k-1}$, the last equality holds.

Many fully sequential IZ procedures are developed similar to Procedure $(\mathcal{P})$. From the proof above, we can identify three sources of conservativeness:

2.1.2.1 Bonferroni inequality (BI)

Suppose that there are events E_i for $i = 1, \ldots, m$ and $m \in \mathbb{Z}^+$. The BI implies that $\Pr\{\cup_{i=1}^{m} E_i\} \leq \sum_{i=1}^{m} \Pr(E_i)$. The BI enables us to focus on the probability of incorrect selection for a pair of systems instead of that for all k systems, and and is useful to get a lower bound on PCS.

2.1.2.2 Discrete process to continuous process (DTC)

A lemma from [22] bounds the probability of making an error in a discrete stochastic process by the probability of making an error in the corresponding continuous stochastic process, usually a Brownian motion process. Many properties are known for a Brownian motion process, which is helpful in deriving statistically valid procedures.

2.1.2.3 Slippage means configuration (SC) assumption

This assumption makes the probability of incorrect selection free of unknown mean parameters but is often an unrealistic mean configuration for a large k.

2.2 Impact of Conservativeness Sources

In this section, we quantify the impact of the three conservativeness sources in terms of the number of observations by experiments under various mean and variance configurations, and identify a critical source that degrades the efficiency the most.

We first present three additional procedures: Procedures $(\mathcal{E})$, $(\mathcal{BI})$, and $(\mathcal{BI}+\mathcal{DTC})$ in Figure 2. These three procedures are statistically valid and designed to quantify the loss due to each conservativeness source. However, they are not implementable in practice because they require prior knowledge on the identity of the best system and the true mean difference between each pair of systems. They are only used in this chapter to study the impact of each conservativeness source on the efficiency of Procedure ($\mathcal{P}$). The difference between Procedures ($\mathcal{E}$) and ($\mathcal{BI}$), ($\mathcal{BI}$) and ($\mathcal{BI}+\mathcal{DTC}$), ($\mathcal{BI}+\mathcal{DTC}$) and ($\mathcal{P}$) indicates the impact of the BI, DTC, and the SC assumption, respectively.

Procedures ($\mathcal{E}$), ($\mathcal{BI}$), and ($\mathcal{BI}+\mathcal{DTC}$)

All steps in the three procedures are the same as those in Procedure ($\mathcal{P}$) except that δ is replaced with $\delta_{i\ell}$ in $R(r;.)$ for comparing systems i and ℓ, and η is the solution to the following equations:

Procedure ($\mathcal{E}$)

$$\Pr\left\{\sum_{j=1}^{T_i^d}(X_{1j}-X_{ij}) \le -R(T_i^d;\delta_{1i},\eta,\sigma_{1i}^2) \text{ for any } i \neq 1\right\} = \alpha,$$

Procedure ($\mathcal{BI}$)

$$\sum_{i=2}^{k}\Pr\left\{\sum_{j=1}^{T_i^d}(X_{1j}-X_{ij}) \le -R(T_i^d;\delta_{1i},\eta,\sigma_{1i}^2)\right\} = \alpha,$$

Procedure ($\mathcal{BI}+\mathcal{DTC}$)

$$\sum_{i=2}^{k}\Pr\left\{\mathcal{W}\left(T_i^c;\frac{\mu_1-\mu_i}{\sigma_{1i}}\right) \le -\max\left\{0,\left(\frac{\eta\sigma_{1i}}{\delta_{1i}}-\frac{\delta_{1i}}{2e\sigma_{1i}}T_i^c\right)\right\}\right\} = \alpha.$$

Figure 2: Algorithmic statement of Procedures ($\mathcal{E}$), ($\mathcal{BI}$), and ($\mathcal{BI}+\mathcal{DTC}$).

2.2.1 Experimental Setup

In this subsection, we present the experimental setup to quantify the impact of each conservativeness source. Two mean configurations are considered: SC and group decreasing means configuration (DC). Under the SC, $\mu_1 = \delta$ and $\mu_i = 0$ for $i = 2,\ldots,k$. The DC is designed in the way to avoid the situation where the mean difference between the best and an alternative is so large that comparison can be done with a few observations. The DC is as follows:

$$\mu_i = \begin{cases} \delta & \text{for } i = 1, \\ 0 & \text{for } i = 2,\ldots,\lceil (k-1)/5\rceil+1, \\ -0.5\delta & \text{for } i = \lceil (k-1)/5\rceil+2,\ldots,2\lceil (k-1)/5\rceil+1, \\ -\delta & \text{for } i = 2\lceil (k-1)/5\rceil+2,\ldots,3\lceil (k-1)/5\rceil+1, \\ -1.5\delta & \text{for } i = 3\lceil (k-1)/5\rceil+2,\ldots,4\lceil (k-1)/5\rceil+1, \\ -2\delta & \text{for } i = 4\lceil (k-1)/5\rceil+2,\ldots,k. \end{cases}$$

For the setting of the DC, three variance configurations are considered: equal variances

(EQUAL), $\sigma_i^2 = 1$; group increasing variances (INC), $\sigma_i^2 = 1+|\mu_i - \delta|$; and group decreasing variances (DEC), $\sigma_i^2 = 1/(1+|\mu_i - \delta|)$. We also run experiments with means in the SC, but with the same variances as in the DC. Therefore there are six mean-variance configurations in total: SC-EQUAL, SC-INC, SC-DEC, DC-EQUAL, DC-INC, and DC-DEC.

The number of systems k varies over $2, 5, 10, 25, 50, 75,$ and 100. The nominal level is $1 - \alpha = 0.95$. For correlation due to the use of CRN, we let ρ be a common correlation across each pair of systems and test $\rho = 0$, 0.3, and 0.6. We set the IZ parameter $\delta = 0.3$ and take the constant $e = 1$, which is recommended in [28].

For Procedures $(\mathcal{E})$ and $(\mathcal{BI})$, the value of η is searched empirically based on 40,000 macro replications for each configuration. On the other hand, η is determined analytically as the solution to $g(\eta) = \alpha/(k-1)$ for Procedures $(\mathcal{BI} + \mathcal{DTC})$ and $(\mathcal{P})$. We report average total number of observations (OBS) until a decision is made. The numbers are reported based on 40,000 macro replications for Procedures $(\mathcal{E})$ and $(\mathcal{BI})$ and 500 macro replications for Procedures $(\mathcal{BI} + \mathcal{DTC})$ and $(\mathcal{P})$. With 500 macro replications, the first three digits of average total number of observations are meaningful. See [43] for determining the number of meaningful digits in point estimates.

2.2.2 Results

We now report experimental results and discuss the impact of each source on the efficiency of Procedure $(\mathcal{P})$ to identify the most critical source. For reasons of brevity, we present numerical results for the SC and the DC with group increasing and decreasing variances and correlation $\rho = 0$ and 0.6. For a given mean configuration and ρ, average total number of observations with equal variances lie between those with group increasing and decreasing variances. Let OBS_E, OBS_{BI}, OBS_{BI+DTC}, and OBS_P represent average total number of observations for Procedures $(\mathcal{E})$, $(\mathcal{BI})$, $(\mathcal{BI} + \mathcal{DTC})$, and $(\mathcal{P})$, respectively. In Figures 3 and 6, lines represent percentages of average additional number of observations overconsumed due to conservative sources. More specifically, lines labeled BI, DTC, and SC show $\frac{\text{OBS}_{BI} - \text{OBS}_E}{\text{OBS}_E} \times 100\%$, $\frac{\text{OBS}_{BI+DTC} - \text{OBS}_{BI}}{\text{OBS}_E} \times 100\%$, and $\frac{\text{OBS}_P - \text{OBS}_{BI+DTC}}{\text{OBS}_E} \times 100\%$, respectively, as a function of k under the SC and the DC with group increasing and decreasing variances when $\rho = 0$.

Our main results are as follows.

2.2.2.1 Impact of the BI

From the lines labeled BI in Figures 3–6, we notice that the impact of the BI gets stronger as k increases. For example, from $k = 5$ to $k = 100$, the percentages of average additional observations due to the BI increase from 5% to 8% and 9% to 30% under SC-INC and SC-DEC with $\rho = 0$, respectively. This is expected because the Bonferroni bound becomes more conservative as the number of systems in consideration increases.

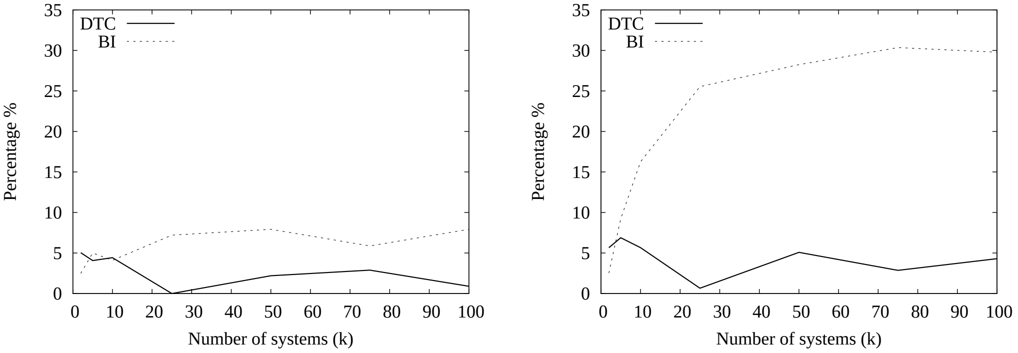

Figure 3: Percentages of average additional number of observations under the SC with INC (left) and DEC (right) variances when $\rho = 0$.

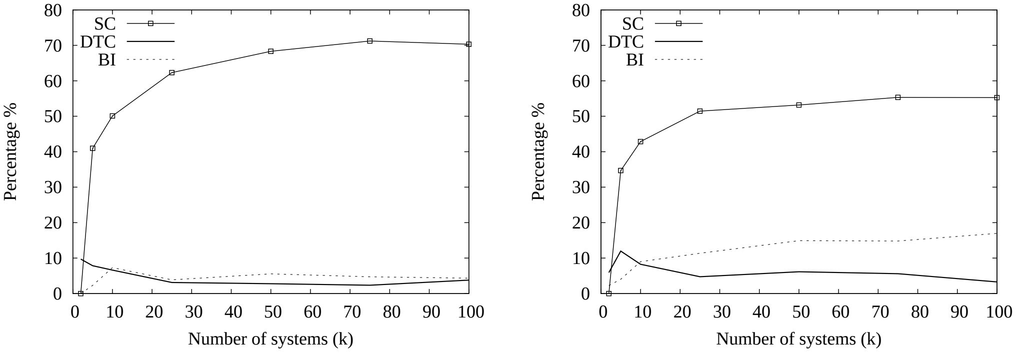

Figure 4: Percentages of average additional number of observations under the DC with INC (left) and DEC (right) variances when $\rho = 0$.

We also notice that the impact of the BI gets stronger under group decreasing than increasing variances. When $k = 100$, the percentage of additional observations is 17% under the DC-DEC, but only 4% under the DC-INC. We observe similar tendency for other

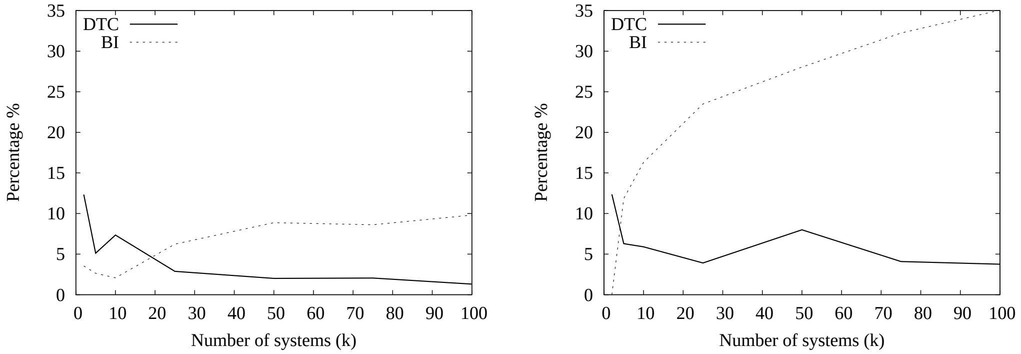

Figure 5: Percentages of average additional number of observations under the SC with INC (left) and DEC (right) variances when $\rho = 0.6$.

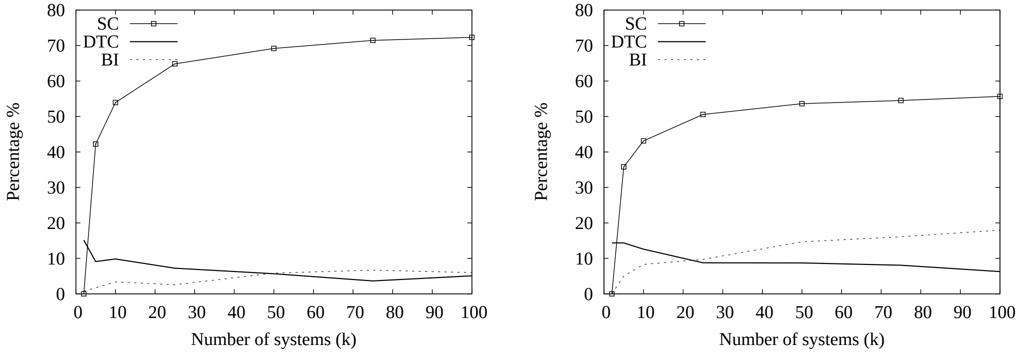

Figure 6: Percentages of average additional number of observations under the DC with INC (left) and DEC (right) variances when $\rho = 0.6$.

values of k. This is due to stronger positive correlation between $X_{1j} - X_{ij}$ and $X_{1j} - X_{\ell j}$ for $i \neq \ell \neq 1$ under group decreasing variances than under group increasing variances. It is well known that the BI becomes more conservative when events under consideration are positively correlated.

2.2.2.2 Impact of DTC

The lines labeled DTC in Figures 3–6 show the impact of DTC. The percentages of average additional observations due to DTC are small (around 5%) and no more than 10% in most cases. This implies that the partial sum of differences $\sum_{j=1}^{r}(X_{1j} - X_{ij})$ is well approximated by a continuous drifted Brownian motion process and the impact of DTC is not significant.

Table 2: Percentages of additional observations due to the SC assumption.

	DC-INC		DC-DEC	
	$k = 5$	$k = 100$	$k = 5$	$k = 100$
$\rho = 0$	41%	70%	35%	55%
$\rho = 0.6$	42%	72%	36%	56%

2.2.2.3 Impact of the SC Assumption

The lines labeled SC quantify the impact of the SC assumption in Figures 4 and 6. When the true mean configuration is the DC, Procedure ($\mathcal{P}$) uses δ to determine the continuation region $(-R(r;\cdot), R(r;\cdot))$ while Procedure ($\mathcal{BI} + \mathcal{DTC}$) uses $\delta_{i\ell}$. As a result, the continuation region of Procedure ($\mathcal{BI} + \mathcal{DTC}$) is narrower than that of Procedure ($\mathcal{P}$) and eliminations of inferior systems occur faster.

Table 2 shows the percentages of additional observations due to the SC assumption, and implies that the impact of the SC assumption overwhelms when k is large under the DC.

2.2.2.4 Impact of Correlation

By comparing Figures 3 and 5, 4 and 6, we observe that the impacts of the BI, DTC, and the SC assumption are similar for independent and correlated systems. The impact of the BI becomes slightly stronger for correlated systems than independent systems, because a positive correlation ρ makes incorrect selection events more strongly correlated. The impact of DTC is still very minor and the percentages of average additional observations due to the SC assumptions are similar for independent and correlated systems.

Overall, this study indicates that the impact of DTC is not significant but the BI and the SC assumption are more critical. Though the actual percentages of average additional observations due to the BI and the SC assumption depend on the true mean and variance configurations, our experimental results show that the BI and the SC are critical. Inefficiency of a fully sequential procedure is mainly due to the BI when the true mean configuration is indeed the SC. Unfortunately, developing a statistically valid fully sequential procedure without the use of the BI is very difficult because it requires studying the behaviors of a $(k-1)$-dimensional drifted Brownian motion process with component-wise correlation. On the other hand, when means are different from the SC, which is likely for large k, the SC assumption deteriorates efficiency significantly more than the BI. Thus, if one is to develop an efficient procedure that works well for large k, it seems more important to remove the SC assumption than the BI, which is the focus of this chapter. In next subsection, we analytically calculate asymptotic percentage of savings we can expect from Procedure $(\mathcal{BI}+\mathcal{DTC})$ compared to Procedure $(\mathcal{P})$ under general mean and variance configurations.

2.2.3 Asymptotic Analysis on the Impact of the SC Assumption

We employ a similar heuristic argument from [38] to approximate the number of observations for Procedures $(\mathcal{P})$ and $(\mathcal{BI}+\mathcal{DTC})$. The following happens for Procedures $(\mathcal{P})$ and $(\mathcal{BI}+\mathcal{DTC})$ as $\alpha \to 0$: (i) procedures stop making mistakes and system 1 is always found superior to other systems; and (ii) a larger number of observations is needed to make a decision, and hence sample mean $\bar{X}_i(r)$ behaves like the true mean μ_i for system i.

Let N_i^P and N_i^B be the number of observations that system i takes until a decision is reached in Procedures $(\mathcal{P})$ and $(\mathcal{BI}+\mathcal{DTC})$, respectively. For Procedure $(\mathcal{P})$, system i is eliminated when

$$\mu_1 - \mu_i \geq \frac{1}{r}R(r;\delta,\eta,\sigma_{1i}^2) = \frac{\eta\sigma_{1i}^2}{r\delta} - \frac{\delta}{2} \quad \text{and} \quad N_i^P \approx \frac{\eta\sigma_{1i}^2}{\delta(\mu_1-\mu_i+\delta/2)}, \text{ for } i = 2,\ldots,k.$$

System 1 keeps receiving additional observations until all inferior systems are eliminated and thus takes $N_1^P = \max(N_2^P,\ldots,N_k^P)$. Similarly, for Procedure $(\mathcal{BI}+\mathcal{DTC})$,

$$\mu_1 - \mu_i \geq \frac{1}{r}R(r;\delta_{1i},\eta,\sigma_{1i}^2) = \frac{\eta\sigma_{1i}^2}{r\delta_{1i}} - \frac{\delta_{1i}}{2} \quad \text{and} \quad N_i^B \approx \frac{\eta\sigma_{1i}^2}{\delta_{1i}(\mu_1-\mu_i+\delta_{1i}/2)}, \text{ for } i = 2,\ldots,k.$$

Also let $N_1^B = \max(N_2^B,\ldots,N_k^B)$.

Then, the asymptotic ratio of savings between the total number of observations for Procedures $(\mathcal{BI}+\mathcal{DTC})$ and $(\mathcal{P})$ is

$$\mathcal{S} \equiv 1 - \frac{\sum_{i=1}^k N_i^B}{\sum_{i=1}^k N_i^P} = 1 - \frac{\sum_{i=2}^k \frac{\eta\sigma_{1i}^2}{3(\mu_1-\mu_i)^2/2} + \max_{i\in\{2,\ldots,k\}} N_i^B}{\sum_{i=2}^k \frac{\eta\sigma_{1i}^2}{\delta(\mu_1-\mu_i+\delta/2)} + \max_{i\in\{2,\ldots,k\}} N_i^P}$$

The following Table 3 shows the asymptotic ratios $\mathcal{S}$ and empirical ratios based on 500 macro replications, when the mean configuration follows the DC with $\alpha = 0.01$ and $\rho = 0$.

Theoretical asymptotic ratios $\mathcal{S}$ match well empirical ratios. In addition, $\mathcal{S}$ range from 23% to 41% under the DC and this is the magnitude of savings one can expect when the SC assumption is removed under the DC. The savings can be greater if the means of systems are even more spaced out than the DC.

Table 3: Asymptotic ratios $\mathcal{S}$ and empirical ratios under the DC, with $\alpha = 0.01$ and $\rho = 0$.

	INC		EQUAL		DEC	
	$\mathcal{S}$	Empirical	$\mathcal{S}$	Empirical	$\mathcal{S}$	Empirical
$k = 5$	0.277	0.273	0.259	0.256	0.247	0.233
$k = 10$	0.345	0.334	0.324	0.311	0.310	0.276
$k = 25$	0.391	0.375	0.368	0.351	0.354	0.319
$k = 50$	0.406	0.393	0.384	0.365	0.370	0.323
$k = 75$	0.412	0.396	0.390	0.369	0.375	0.327
$k = 100$	0.415	0.406	0.392	0.375	0.384	0.378

Next we present practically implementable versions of Procedure $(\mathcal{BI} + \mathcal{DTC})$ that are asymptotically valid as δ goes to zero.

2.3 New Procedures

In this section, we propose two new procedures, namely, $\mathcal{WK}$ and $\mathcal{WK}$++ that lessen the impact of the SC assumption and are asymptotically valid as δ approaches to zero. Procedure $(\mathcal{BI} + \mathcal{DTC})$ replaces δ with the true mean differences $\delta_{i\ell}$ in calculating the continuation region $R(r; \cdot)$, which is impossible in practice. Instead, we replace δ with $\hat{\delta}_{i\ell}(r)$, a function of estimated mean differences. Procedure $\mathcal{WK}$ updates estimated mean differences between systems as more observations become available. We also consider a version with variance update, called $\mathcal{WK}$++, which incorporates the idea of updating sample variances as in [29, 35] to achieve further efficiency. A function of estimated mean differences $\hat{\delta}_{i\ell}(r)$ needs a random variable $A_{i\ell}(r)$ that satisfies the following assumption:

Assumption 3 *For $i \neq \ell$ and $i, \ell \in \{1, \ldots, k\}$, $A_{i\ell}(r) \geq 0$ for all $r \in \mathbb{Z}^+$ and $A_{i\ell}(r) \to 0$ with probability one, as $r \to \infty$.*

The purpose of $A_{i\ell}(r)$ is to prevent early termination of the new procedures. We use $A_{i\ell}(r) = \xi_q \frac{S_{i\ell}(n_0)}{\sqrt{r}}$ for $\mathcal{WK}$ and $A_{i\ell}(r) = \xi_q \frac{S_{i\ell}(r)}{\sqrt{r}}$ for $\mathcal{WK}$++ where ξ_q is the q quantile of

Procedures $\mathcal{WK}$ and $\mathcal{WK}$++

Setup: Select the nominal level $1-\alpha$, the IZ parameter δ and the first-stage sample size n_0. For $\mathcal{WK}$, $h^2 = \eta(n_0 - 1)$ where the constant η is the solution to the equation $g'(\eta, n_0) = \frac{\alpha}{k-1}$. For $\mathcal{WK}$++, $h^2 = \eta$ where η is the solution to $g(\eta) = \frac{\alpha}{k-1}$.

Initialization: Set $I = \{1, \ldots, k\}$ and obtain n_0 observations $X_{ij}, j = 1, \ldots, n_0$ from all systems i, $i \in I$. Set $r = n_0$ and go to **Variance Update**.

Variance Update: For all $i \neq \ell$, $i, \ell \in I$, compute $S^2_{i\ell}(r)$, the sample variance of differences between systems i and ℓ based on r observations. Then go to **Mean Update**.

Mean Update: For all $i \neq \ell$, $i, \ell \in I$, compute $\bar{X}_i(r) - \bar{X}_\ell(r)$, the sample mean difference between systems i and ℓ based on r observations, and calculate $\hat{\delta}_{i\ell}(r)$ (the choice of $A_{i\ell}(r)$ is discussed in Assumption 3). Then go to **Screening**.

Screening: Set $I^{\text{old}} = I$. For $\mathcal{WK}$,

$$I = \left\{ i : i \in I^{\text{old}} \text{ and } \sum_{j=1}^{r} (X_{ij} - X_{\ell j}) \geq -R(r; \hat{\delta}_{i\ell}(r), h^2, S^2_{i\ell}(n_0)), \forall \ell \in I^{\text{old}}, \ell \neq i \right\}.$$

For $\mathcal{WK}$++,

$$I = \left\{ i : i \in I^{\text{old}} \text{ and } \sum_{j=1}^{r} (X_{ij} - X_{\ell j}) \geq -R(r; \hat{\delta}_{i\ell}(r), h^2, S^2_{i\ell}(r)), \forall \ell \in I^{\text{old}}, \ell \neq i \right\}.$$

Stopping Rule: If $|I| = 1$, then stop and select the system whose index is in I as the best. Otherwise, take one additional observation $X_{i,r+1}$ from each system $i \in I$ and set $r = r + 1$. For $\mathcal{WK}$, go to **Mean Update**. For $\mathcal{WK}$++, go to **Variance Update**.

Figure 7: Algorithmic statement of Procedures $\mathcal{WK}$ and $\mathcal{WK}$++.

the standard normal random variable for $q \geq 0.5$, but other random variables satisfying Assumption 3 can be used.

Procedures $\mathcal{WK}$ and $\mathcal{WK}$++ are described in Figure 7 and we now present our main theorems.

Theorem 1 *If Assumptions 1, 2, and 3 hold, then Procedure $\mathcal{WK}$ guarantees*

$$\liminf_{\delta \to 0} \Pr(\text{CS}) \geq 1 - \alpha,$$

where CS is the event that system 1 is selected.

Theorem 2 *If Assumptions 1, 2, and 3 hold, then Procedure $\mathcal{WK}$++ guarantees*

$$\liminf_{\delta \to 0} \Pr(\text{CS}) \geq 1 - \alpha.$$

The asymptotic validity of the proposed procedures is shown as $\delta \to 0$. Kim and Nelson [29] use the same asymptotic regime to prove the validity of their procedures and explain the importance of Assumption 2 which ensures the true mean differences also converges to zero as δ converges to zero. The proofs of Theorems 1 and 2 are shown in Appendix A.

Procedure $\mathcal{WK}$++ with $h^2(r)$

Same as Procedure $\mathcal{WK}$++ in Figure 7, except the following two steps:

Variance Update: For all $i \neq \ell$, $i, \ell \in I$, compute $S_{i\ell}^2(r)$, the sample variance of differences between systems i and ℓ based on r observations. Update η_r, the solution to $g'(\eta, r) = \frac{\alpha}{k-1}$ and $h^2(r) = \eta_r(r-1)$. Then go to **Mean Update**.

Screening: Set $I^{\text{old}} = I$. Let

$$I = \left\{ i : i \in I^{\text{old}} \text{ and } \sum_{j=1}^{r} (X_{ij} - X_{\ell j}) \geq -R(r; \hat{\delta}_{i\ell}(r), h^2(r), S_{i\ell}^2(r)), \forall \ell \in I^{\text{old}}, \ell \neq i \right\}.$$

Figure 8: Algorithmic statement of Procedure $\mathcal{WK}$++ with $h^2(r)$.

Note that $h^2 = \eta$ in $\mathcal{WK}$++ is the same constant used in Procedure ($\mathcal{P}$) which assumes known variances of all systems. Although $\mathcal{WK}$++ is asymptotically valid, the use of $h^2 = \eta$ may cause a severe under-coverage problem in finite samples, therefore $h^2(r) = \eta_r(r-1)$ is recommended where $g'(\eta, r) = \frac{\alpha}{k-1}$. The procedure is still asymptotically valid with $h^2(r)$ because $h^2(r)$ is larger than h^2. Then steps in $\mathcal{WK}$++ are modified as in Figure 8.

2.4 Experiments

In this section, we compare the performances of $\mathcal{WK}$ and $\mathcal{WK}$++ with two existing fully sequential procedures: $\mathcal{KN}$ [28] and $\mathcal{KN}$++ [29, 35]. Procedures $\mathcal{KN}$ and $\mathcal{KN}$++ use the BI, DTC, and the SC assumption for their validity. $\mathcal{WK}$ and $\mathcal{WK}$++ are counterparts of $\mathcal{KN}$ and $\mathcal{KN}$++ without the use of the SC assumption, respectively. For $\mathcal{WK}$++, we use the version with $h^2(r)$.

We employ the same mean and variance configurations described in Section 2.2.1 and report average total number of observations and estimated PCS based on 10,000 macro replications. The initial sample size is $n_0 = 10$. The new procedures are tested with five different values of ξ_q for $A_{i\ell}(r)$: $\xi_{0.5} = 0$, $\xi_{0.6} = 0.25$, $\xi_{0.8} = 0.84$, $\xi_{0.95} = 1.64$ and $\xi_{0.975} = 1.96$. We present selected results of our experiments.

1) Estimated PCS: Table 4 reports estimated PCS under various mean and variance configurations for $k = 10$ and $k = 100$ with $\rho = 0$. Estimated PCS are meaningful up to the thousandths place with 10,000 macro replications. We focus on the SC to discuss validity of the new procedures. When $k = 10$, Table 4 shows that estimated PCS are reasonably acceptable (all over 90%) except under the SC-INC configuration with q close to 0.5. For example, estimated PCS are 87.4% for $\mathcal{WK}$ and 83.3% for $\mathcal{WK}$++ when $q = 0.5$. When q

Table 4: Estimated PCS when $k = 10$ and 100 with $\rho = 0$.

$k = 10$	SC		DC	
	INC	DEC	INC	DEC
$\mathcal{KN}$	0.960	0.965	0.989	0.989
$\mathcal{WK}$ (0.50)	0.874	0.925	0.953	0.968
$\mathcal{WK}$ (0.60)	0.894	0.937	0.965	0.975
$\mathcal{WK}$ (0.80)	0.929	0.955	0.979	0.984
$\mathcal{WK}$ (0.95)	0.953	0.967	0.989	0.987
$\mathcal{WK}$ (0.975)	0.957	0.965	0.986	0.990
$\mathcal{KN}$++	0.955	0.962	0.989	0.988
$\mathcal{WK}$++ (0.50)	0.833	0.906	0.936	0.964
$\mathcal{WK}$++ (0.60)	0.869	0.922	0.955	0.968
$\mathcal{WK}$++ (0.80)	0.914	0.946	0.973	0.984
$\mathcal{WK}$++ (0.95)	0.943	0.958	0.983	0.986
$\mathcal{WK}$++ (0.975)	0.947	0.955	0.984	0.986
$k = 100$	SC		DC	
	INC	DEC	INC	DEC
$\mathcal{KN}$	0.964	0.971	0.990	0.992
$\mathcal{WK}$ (0.50)	0.835	0.925	0.944	0.963
$\mathcal{WK}$ (0.60)	0.858	0.931	0.951	0.978
$\mathcal{WK}$ (0.80)	0.908	0.957	0.971	0.984
$\mathcal{WK}$ (0.95)	0.942	0.965	0.985	0.988
$\mathcal{WK}$ (0.975)	0.943	0.966	0.985	0.990
$\mathcal{KN}$++	0.968	0.978	0.990	0.991
$\mathcal{WK}$++ (0.50)	0.748	0.898	0.904	0.949
$\mathcal{WK}$++ (0.60)	0.809	0.916	0.929	0.964
$\mathcal{WK}$++ (0.80)	0.879	0.942	0.964	0.979
$\mathcal{WK}$++ (0.95)	0.929	0.965	0.980	0.988
$\mathcal{WK}$++ (0.975)	0.944	0.967	0.982	0.986

is close to 0.5, $\hat{\delta}_{i\ell}(r)$ is essentially sample mean difference which can overestimate the true mean difference $\delta_{i\ell}$. This overestimation of the true mean differences can result in early termination of the procedures and degradation in PCS. On the other hand, q close to 1 is likely to result in $\hat{\delta}_{i\ell}(n_0) = \delta$ or $\hat{\delta}_{i\ell}(r) = \delta$. This helps achieve the nominal PCS but at the cost of observations because the new procedures will not be much different from Procedure $\mathcal{KN}$ or $\mathcal{KN}$++.

When $k = 100$, estimated PCS in Table 4 are lower than those for when $k = 10$ and quite small for $q = 0.5$ and 0.6 under the SC. However, low estimated PCS for $q = 0.5$ or 0.6 are under the very unrealistic mean configuration that ninety-nine systems are exactly δ away from the best. When $k = 100$, the DC is more realistic and estimated PCS under

Table 5: Average total number of observations (OBS) under the DC with $\rho = 0$.

	$k = 10$		$k = 100$	
	INC	DEC	INC	DEC
$\mathcal{KN}$	996	617	14434	7643
$\mathcal{WK}$ (0.50)	552	392	7596	4878
$\mathcal{WK}$ (0.60)	615	427	8327	5214
$\mathcal{WK}$ (0.80)	735	497	9846	5932
$\mathcal{WK}$ (0.95)	871	567	11574	6671
$\mathcal{WK}$ (0.975)	905	587	12155	6879
$\mathcal{KN}$++	630	406	8800	4919
$\mathcal{WK}$++ (0.50)	351	264	3664	2642
$\mathcal{WK}$++ (0.60)	385	287	4075	2883
$\mathcal{WK}$++ (0.80)	466	330	5141	3434
$\mathcal{WK}$++ (0.95)	551	375	6546	4091
$\mathcal{WK}$++ (0.975)	578	386	7041	4298

the DC are acceptable for all five values of q. When means are known to be different from the SC, a choice of $q = 0.6$ seems good in terms of achieving the nominal PCS. If the mean configuration is believed to be indeed the SC, the choice of $q = 0.95$ is safe.

2) Efficiency: Table 5 reports average total number of observations under the DC and various variance configurations for $k = 10$ and 100 with $\rho = 0$. We do not report average total number of observations under the SC because we already know that there is not much savings for $\mathcal{WK}$ and $\mathcal{WK}$++ compared to $\mathcal{KN}$ and $\mathcal{KN}$++ under the SC. Under the DC, we observe significant savings for $q = 0.5$, 0.6, and 0.8. For example, under the DC-INC, when $k = 10$ and $q = 0.6$, $\mathcal{WK}$ spends 38.3% fewer observations compared to $\mathcal{KN}$, and $\mathcal{WK}$++ achieves 38.9% savings compared to $\mathcal{KN}$++. When $k = 100$ and $q = 0.6$, the percentages of savings increase to 42.3% and 53.7% for $\mathcal{WK}$ and $\mathcal{WK}$++, respectively. With $q = 0.8$, the percentages of savings are 26.0% for both $\mathcal{WK}$ and $\mathcal{WK}$++ when $k = 10$ and increase to 31.8% and 41.6% when $k = 100$, for $\mathcal{WK}$ and $\mathcal{WK}$++, respectively. The choice of $q = 0.5$ or 0.6 brings large savings in observations but at the cost of PCS. The choice of $q = 0.8$ results in significant savings and good PCS. Considering both estimated PCS and average total number of observations, $q = 0.8$ seems a good compromise.

3) Impact of Correlation: From estimated PCS when $\rho = 0.3$ and 0.6, that we do not report here for reasons of brevity, we observe that estimated PCS increase and average total number of observations decrease as correlation due to CRN becomes stronger. This is well expected because variances of differences between systems get smaller and thus comparisons become easier. Estimated PCS with $\rho = 0.6$ are all over 90% even with $q = 0.5$ for all configurations we tested. From Figure 9, one can notice that percentage of savings in average total number of observations is still significant but smaller when $\rho = 0.6$ compared to $\rho = 0$. This is related to the fact that we have to take $n_0 = 10$ initial observations from all systems and savings in average total number of observations may not be fully recognized when comparison between a pair is so easy that ten observations are already more than enough to reach a decision. Figure 10 shows estimated PCS for $\mathcal{WK}$ and $\mathcal{WK}$++ when $\rho = 0$ and 0.6.

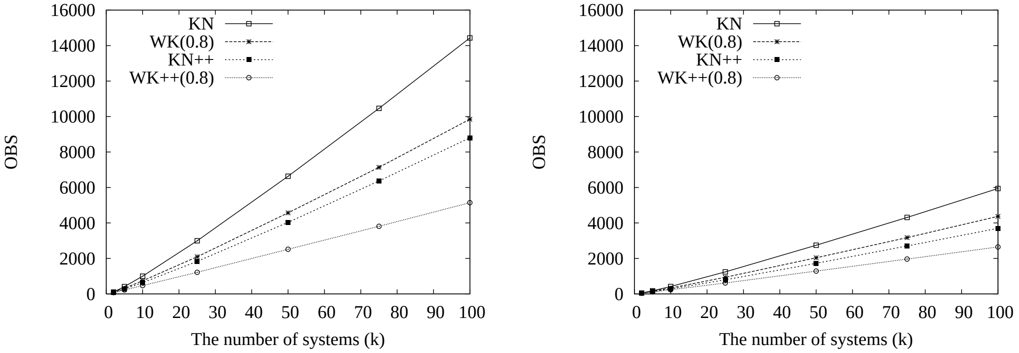

Figure 9: Average total number of observations (OBS) under the DC-INC with $\rho = 0$ (left) and $\rho = 0.6$ (right).

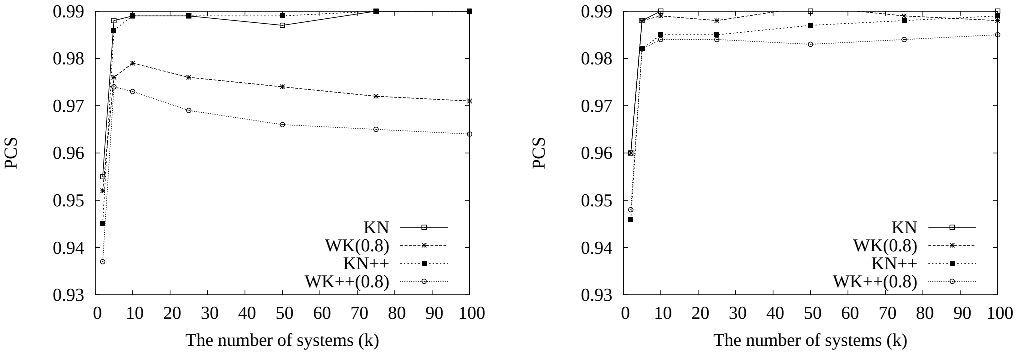

Figure 10: Estimated PCS under the DC-INC with $\rho = 0$ (left) and $\rho = 0.6$ (right).

2.5 Conclusion

We study conservativeness of fully sequential procedures. Both the BI and the SC assumption are critical sources that affect the efficiency of fully sequential procedures significantly. The impact of the SC assumption overwhelms when the number of system is large with the mean configuration different from the SC. Two new procedures, $\mathcal{WK}$ and $\mathcal{WK}$++, lessen the impact of the SC assumption and are shown to be efficient compared to existing procedures.

CHAPTER III

MONITORING NONLINEAR PROFILES ADAPTIVELY WITH A WAVELET-BASED DISTRIBUTION-FREE CUSUM CHART

With the rapid development of sensor technology, we can collect massive process and product data to monitor, control, and improve process performance and product quality. Massive data may lead to a data overabundance. For an example, Jin and Shi [24] study a tonnage signal from an automotive stamping process, where a press can typically perform 200 strokes per minute with more than 6000 data points in each stroke, i.e., 1.2 million data points per minute. Staudhammer et al. [45] use laser range sensors (LRSs) to measure the thickness of sawed boards from a lumber manufacturing process, and each LRS provides more than 2000 data points for each sawed board.

These series of data points for each stroke or sawed board are called profiles. A profile describes the functional relationship between a response variable and one or more explanatory variables [51]. In this chapter, we restrict attention to one explanatory variable, the same setting used in [13, 23, 33]. The functional relationship sometimes has a linear or nonlinear representation, but it also can be too complicated to have a well-known functional form. Examples, including tonnage signals or LRS data, often have jumps, cusps, oscillations, and other types of non-smooth behavior.

Monitoring linear profiles, such as some calibration processes, widely employ linear regression models. In order to detect a shift in the mean of successive observed profiles, control charts are applied to regression parameters, such as estimated intercepts and slopes. For simple linear profiles, Kang and Albin [27] perform a single control chart on a multivariate T^2-statistic constructed on estimated intercepts and slopes. For polynomial linear profiles and general linear profiles, Zou et al. [52] monitor all estimated regression parameters within a single chart using multivariate exponentially weighted moving average (EWMA) schemes.

Nonlinear profiles, such as product shapes and consecutive measurements of the same variable at different locations on individual products, are often modeled by nonlinear regression models; and a control chart is performed on estimated regression parameters from the observed profiles. Williams et al. [50] use a Hotelling's T^2-chart to monitor parameters of a nonlinear regression model for a particle board manufacturing process. To detect several complex sawing defects in lumber manufacturing, Staudhammer et al. [45] model LRS profiles with a combination of two regression models, multiple linear and nonlinear, to describe roughness and waviness of surfaces, respectively; and then they monitor regression parameters from different models with separate multiple profile charts.

Even though Gupta et al. [21] point out that the use of control charts based on estimated regression parameters is effective in terms of average run length performance, there are some issues in using regression models for profile monitoring. First, regression methods tend to use smooth models to estimate in-control mean profiles; thus we lose some important local information, such as jumps and cusps. Secondly, the performance of control charts highly depends on the choice of models. If an inappropriate model is selected, then we may get misleading results and not detect shifts in the mean profile adequately. For example, Chicken et al. [13] show that when a change occurs in the mean profile, the selected regression model possibly still yields the same parameter estimates so that the shift is undetectable. In addition, fitting a sufficiently accurate regression model to observed profiles can be time-consuming. Lastly, some process profiles, such as tonnage signals and LRS data, can be too complex to be modeled adequately by parametric regression models.

As a consequence, some profile monitoring charts have been based on nonparametric regression techniques, such as smoothing splines, functional principal components analysis (FPCA), Fourier analysis, and wavelet analysis. Gardner et al. [19] use a smoothing spline to model the thickness at selected locations of a wafer surface in a semiconductor manufacturing process. Ramsay and Silverman [41] propose methods based on FPCA that can be used in nonparametric profile monitoring. Chen and Nembhard [12] compute the Fast Fourier Transform (FFT) of profiles and construct an adaptive Neyman test statistic from the corresponding Fourier coefficients. However, splines, FPCA, and Fourier transforms cannot model locally sharp changes well—for example, see [13, 17, 23]. Instead, the discrete wavelet transform (DWT) has been suggested and shown to be effective for detecting and diagnosing process faults [16, 25]. Some advantages of using DWTs for profile monitoring are summarized as follows.

- Sparsity: the DWT can use substantially fewer wavelet basis functions to achieve a comparably accurate approximation for profiles, especially non-smooth profiles.
- Localization: the DWT is localized in both the frequency and time domains, that is, a frequency change causes changes only in certain time segments. By contrast, the Fourier transform is localized only in the frequency domain, which means a frequency change can cause changes everywhere in the time domain. Therefore the DWT can preserve locally sharp changes in profiles well and help diagnose process faults.
- Fast computation: For a profile of dimension n, the computational complexity of the profile's DWT is $O(n)$, see [34]; and this is smaller than the complexity $O(n \log_2 n)$ of the profile's FFT.

Therefore wavelet-based approaches have gained popularity, especially for monitoring high-dimensional profiles with non-smooth behaviors. The DWT is usually used as either a denoising method or a dimension-reduction method in nonparametric profile monitoring. Jeong et al. [23], Jin and Shi [25] use a DWT to denoise observed profiles via universal thresholds and then monitor the preserved components with control charts. The number and locations of preserved components change from profile to profile; therefore monitored wavelet components are selected adaptively. Their wavelet-based control charts with adaptive selection detect shifts effectively, but those control charts require assumptions of marginal normality and sometimes independence among components. Without the normality assumption, it is challenging to obtain distributions of monitored statistics and construct analytical control limits for control charts.

On the other hand, Jin and Shi [24], Lada et al. [32], and Lee et al. [33] employ the DWT as a dimension-reduction technique. Based on engineering knowledge of the automotive stamping operation that different process failures correspond to changes in different segments of the tonnage signal from the stamping process, Jin and Shi [24] construct different thresholds for different profile segments. Then they monitor a fixed selection of wavelet components according to the determined thresholds, and detect process faults in the tonnage signals. Lada et al. [32] reduce profile dimensions by minimizing the weighted relative reconstruction error (WRRE), in order to balance model parsimony against data reconstruction error. Lee et al. [33] apply WRRE to an in-control mean profile to determine locations of wavelet components whose values need to be monitored. A reduced-dimension DWT vector of a profile is formed by keeping wavelet components from the determined locations only; and then a distribution-free CUSUM chart is applied to Hotelling's T^2-type statistics computed from the reduced-dimension DWT vectors. Since both control charts in [24] and [33] select wavelet components from the same locations in each profile's DWT, the selection method for wavelet components is considered to be static or nonadaptive. Even though some wavelet-based control charts with static selection (specifically [33]) are distribution-free, they are insensitive to some shifts or even fail to detect some shifts—for instance, a shift only affecting the unselected wavelet components. Extending static selection to dynamic (adaptive) selection is not trivial. Because the selected components change from profile to profile and the number of selected components is affected by the correlations between profile components, monitoring statistics tend to have a large variance; and the distribution of monitored statistics can be quite complicated.

In this section we formulate and evaluate WDFTC$_a$, a wavelet-based distribution-free CUSUM chart with adaptive selection that dynamically adjusts the number and locations of preserved components for each monitored DWT vector. In brief WDFTC$_a$ consists of the following steps. First computing a DWT of the noise vectors for randomly sampled Phase I (in-control) profiles, WDFTC$_a$ uses a matrix-regularization method to estimate the covariance matrix of the wavelet-transformed noise vectors; then those vectors are aggregated (batched) so that the nonoverlapping batch means of the wavelet-transformed noise vectors have manageable covariances. Lower and upper in-control thresholds are computed for the resulting batch means of the wavelet-transformed noise vectors using the associated marginal Cornish-Fisher expansions that have been suitably adjusted for correlations between the components of those vectors. From the thresholded batch means of the wavelet-transformed noise vectors, Hotelling's T^2-type statistics are computed to set the parameters of a CUSUM procedure. To monitor shifts in the mean profile during Phase II (regular) operation, WDFTC$_a$ computes a similar Hotelling's T^2-type statistic from successive thresholded batch means of the wavelet-transformed noise vectors using the upper and lower in-control thresholds; then WDFTC$_a$ applies the CUSUM procedure to the resulting T^2-type statistics for rapid detection of an out-of-control condition.

Ideally, in Phase I (in-control) operation WDFTC$_a$ will retain relatively few detail coefficients; on the other hand, when the observed profiles are out-of-control, WDFTC$_a$'s estimators of both the scaling and detail coefficients will exhibit significant shifts from their in-control counterparts. Thus we can rapidly detect such shifts by monitoring changes in T^2-type statistics computed from the selected components. In this chapter, we extend the wavelet-based control chart with static selection of [33] to formulate new control charts with adaptive selection. Experimental results show the new control charts can detect various shifts more effectively than the method of [33] under a general noise distribution.

Chapter 3 is organized as follows. Section 3.1 briefly defines the problem and notation, provides motivating examples, and introduces the DWT. In Section 3.2, we describe the new profile-monitoring procedure WDFTC$_a$, and we discuss distribution-free thresholding methods and covariance matrix estimation. Section 3.3 presents experimental results for WDFTC$_a$ and its competitors, followed by conclusions in Section 3.4.

3.1 Background

In this section, we first define our problem and provide notation; then we discuss a motivating example and present a brief overview of DWT.

3.1.1 Notation and Problem

For $j = 1, 2, \ldots$, we consider the jth observed profile $\boldsymbol{Y}_j = (Y_{j,1}, \ldots, Y_{j,n})^T$ as an n-dimensional random vector; and for $i = 1, \ldots, n$, the ith component $Y_{j,i}$ of that profile is the response associated with the ith prespecified level x_i of the designated explanatory variable so that we have the functional relationship

$$Y_{j,i} = f(x_i) + \varepsilon_{j,i} \text{ for } i = 1, \ldots, n \text{ and } j = 1, 2, \ldots\,, \tag{7}$$

where $\varepsilon_{j,i}$ is the error (noise) in the ith component of the jth profile. (Throughout this chapter, we let $\boldsymbol{A}^T$ denote the transpose of a vector or matrix $\boldsymbol{A}$.) If we let $\boldsymbol{x} = (x_1, \ldots, x_n)^T$ denote the $n \times 1$ vector of preselected levels of the explanatory variable that are used with each profile, and if we let $\boldsymbol{\epsilon}_j = (\epsilon_{j,1}, \ldots, \epsilon_{j,n})^T$ denote the corresponding $n \times 1$ vector of noise terms for the jth profile, then the functional relationship (7) may be compactly expressed as $\boldsymbol{Y}_j = \boldsymbol{f}(\boldsymbol{x}) + \boldsymbol{\epsilon}_j$ for $j = 1, 2, \ldots$, where $\boldsymbol{f}(\boldsymbol{x}) \equiv [f(x_1), \ldots, f(x_n)]^T$. We assume that $\mathrm{E}[\boldsymbol{\epsilon}_j] = 0$ and that the covariance matrix $\mathrm{Cov}[\boldsymbol{\epsilon}_\mathrm{j}] = \mathrm{E}[\boldsymbol{\epsilon}_\mathrm{j}\boldsymbol{\epsilon}_\mathrm{j}^\mathrm{T}] = \Sigma_0$. In general the components of $\boldsymbol{\epsilon}_j$ may be non-normal and correlated. With this setup, the jth profile $\boldsymbol{Y}_j$ has mean vector $\mathrm{E}[\boldsymbol{Y}_j] = \boldsymbol{f}(\boldsymbol{x})$ and covariance matrix $\boldsymbol{\Sigma}_0$. In addition, the marginal variance of $Y_{j,i}$ or $\epsilon_{j,i}$ is denoted by $\sigma_i^2 = [\boldsymbol{\Sigma}_0]_{ii}$.

We assume that when $\boldsymbol{f}(\cdot)$ is the in-control function $\boldsymbol{f}_0(\cdot)$, i.e., $\mathrm{E}[\boldsymbol{Y}_j] = \boldsymbol{f}_0(\boldsymbol{x})$, the monitored process is in control and $\boldsymbol{Y}_j$ is an in-control profile. When $\boldsymbol{f}(\cdot)$ is some out-of-control function $\boldsymbol{f}_1(\cdot)$, i.e., $\mathrm{E}[\boldsymbol{Y}_j] = \boldsymbol{f}_1(\boldsymbol{x}) \neq \boldsymbol{f}_0(\boldsymbol{x})$, the monitored process is out of control and $\boldsymbol{Y}_j$ is an out-of-control profile. Our problem is to detect rapidly the onset of an out-of-control condition.

To simplify notation, we let $\boldsymbol{f}_0 \equiv \boldsymbol{f}_0(\boldsymbol{x})$ and $\boldsymbol{f}_1 \equiv \boldsymbol{f}_1(\boldsymbol{x}) \neq \boldsymbol{f}_0$, where $\boldsymbol{f}_0 = [f_{0,1}, \ldots, f_{0,n}]^T$ and $\boldsymbol{f}_1 = [f_{1,1}, \ldots, f_{1,n}]^T$. Without loss of generality, we assume $\boldsymbol{f}_0$ is centered, i.e., $\sum_{i=1}^n f_{0,i} = 0$. The profile dimension is set to $n = 2^J$ for some positive integer J. Let $L \in \{0, \ldots, J-1\}$ denote the coarsest level of resolution in a given DWT system (see Section 3.1.3) and $\boldsymbol{W}$ denote the associated DWT matrix for the given DWT system with the selected L. We apply the DWT to $\boldsymbol{f}_0$, $\boldsymbol{Y}_j$ and $\boldsymbol{\epsilon}_j$ for $j = 1, 2, \ldots$, obtaining the corresponding DWTs $\boldsymbol{\theta}_0 = \boldsymbol{W}\boldsymbol{f}_0 = (\theta_{0,1}, \ldots, \theta_{0,n})^T$, $\boldsymbol{d}_j = \boldsymbol{W}\boldsymbol{Y}_j = (d_{j,1}, \ldots, d_{j,n})^T$, and $\boldsymbol{\omega}_j = \boldsymbol{W}\boldsymbol{\epsilon}_j = (\omega_{j,1}, \ldots, \omega_{j,n})^T$, respectively. The first 2^L components in $\boldsymbol{\theta}_0$, $\boldsymbol{d}_j$ and $\boldsymbol{\omega}_j$ are scaling coefficients and contain important features of the original vectors at the lower levels of resolution. The remaining $n - 2^L$ components are called detail coefficients. The covariance matrix of $\boldsymbol{d}_j$ or $\boldsymbol{\omega}_j$ is $\boldsymbol{\Lambda}_0 = \boldsymbol{W}\boldsymbol{\Sigma}_0\boldsymbol{W}^T$. All notation is summarized in Table 6.

Table 6: Notation summary.

$\boldsymbol{f}_0$	The $n \times 1$ in-control mean profile;
$\boldsymbol{Y}_j$	The jth $n \times 1$ observed profile, for $j = 1, \ldots$;
$\boldsymbol{\epsilon}_j$	$= \boldsymbol{Y}_j - \boldsymbol{f}_0,$ the $n \times 1$ noise vector of the jth observed profile $\boldsymbol{Y}_j$;
$\boldsymbol{\theta}_0$	$= \boldsymbol{W}\boldsymbol{f}_0,$ the $n \times 1$ DWT of the in-control mean profile;
$\boldsymbol{d}_j$	$= \boldsymbol{W}\boldsymbol{Y}_j,$ the $n \times 1$ DWT of the jth observed profile $\boldsymbol{Y}_j$;
$\boldsymbol{\omega}_j$	$= \boldsymbol{W}\boldsymbol{\epsilon}_j = \boldsymbol{d}_j - \boldsymbol{\theta}_0,$ the $n \times 1$ DWT of the noise vector of the jth observed profile $\boldsymbol{Y}_j$;
$\boldsymbol{\omega}_k(r)$	$= r^{-1}\sum_{u=1}^{r}\boldsymbol{\omega}_{(k-1)r+u},$ the DWT of the kth non-overlapping batch mean of noise vectors with batch size r;
$\boldsymbol{\lambda}^L$	$= \{\lambda_1^L, \ldots, \lambda_n^L\},$ an $n \times 1$ lower threshold vector;
$\boldsymbol{\lambda}^U$	$= \{\lambda_1^U, \ldots, \lambda_n^U\},$ an $n \times 1$ upper threshold vector;
$\boldsymbol{\omega}_k^*(r)$	$= \{\boldsymbol{\omega}_{k,1}^*(r), \ldots, \boldsymbol{\omega}_{k,n}^*(r)\},$ the thresholded DWT of the kth non-overlapping batch mean of noise vectors, i.e., applying lower and upper thresholds $\boldsymbol{\lambda}^L$ and $\boldsymbol{\lambda}^U$ on $\boldsymbol{\omega}_k(r)$, where $\boldsymbol{\omega}_{k,i}^*(r) = \boldsymbol{\omega}_{k,i}(r)(1 - \mathbf{1}_{\lambda_i^L < \boldsymbol{\omega}_{k,i}(r) < \lambda_i^U})$ and $\mathbf{1}$ is an indicator function;
$\overline{\boldsymbol{\varpi}}_N$	$= N^{-1}\sum_{j=1}^{N}\boldsymbol{\omega}_j,$ the sample mean of $\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_N,$ where N is the size of a data set;
$\widehat{\boldsymbol{\Lambda}}_0$	$= (N-1)^{-1}\sum_{j=1}^{N}(\boldsymbol{\omega}_j - \overline{\boldsymbol{\varpi}}_N)(\boldsymbol{\omega}_j - \overline{\boldsymbol{\varpi}}_N)^T, n \times n$ sample covariance matrix of noise vectors;
$\widetilde{\boldsymbol{\Lambda}}_0$	The regularizated version of $\widehat{\boldsymbol{\Lambda}}_0$, see discussion in Section 3.2.3;
$\widetilde{\boldsymbol{\Lambda}}_0(r)$	$= \widetilde{\boldsymbol{\Lambda}}_0 / r$.

In this chapter, we define the in-control average run length, ARL_0, as the average number of in-control profiles that are observed before a (false) out-of-control alarm is raised; similarly the out-of-control average run length, ARL_1, is the average number of out-of-control profiles that are observed before a (true) out-of-control alarm is raised. We compare and analyze different control charts in terms of ARL_0 and ARL_1.

3.1.2 Motivation

In this subsection, we show that a wavelet-based monitoring chart with static selection can miss certain shifts, which in turn demonstrates the need for new charts with adaptive selection. More specifically, we examine the performance of WDFTC_s from [33], which is a distribution-free tabular CUSUM chart on T^2-type statistics constructed from preselected locations of wavelet components. Recall that preselected locations are fixed for all profiles regardless of shift types. A brief description of Procedure WDFTC_s is provided in Figure 11 as below; also see Figure 2 in [33] for more details.

In our preliminary experiments, we use Mallat's piecewise smooth function as $\boldsymbol{f}_0$ with $n = 512$ equispaced data points, as depicted in Figure 12. This function exhibits some

Procedure WDFTC$_s$

Phase I—Obtain the $n \times 1$ DWT of the in-control mean profile $\boldsymbol{\theta}_0$ and assign weights q and $1-q$ ($0 \leq q \leq 1$), to the compression ratio and reconstructed error of the in-control mean profile $\boldsymbol{f}_0$ in WRRE, which is defined in Equation (6) from [32]. Then determine p ($p < n$) wavelet locations of $\boldsymbol{\theta}_0$ that minimize WRRE. WDFTC$_s$ monitors p components from selected locations instead of all n components in each DWT. For observed in-control profiles $\{\boldsymbol{Y}_j : j = 1, \ldots, N\}$, follow the following steps.

1. For $j = 1, \ldots, N$, obtain the reduced-dimension DWT $\boldsymbol{\omega}_j^{\#}$, by keeping components from the selected wavelet locations in the DWT $\boldsymbol{\omega}_j = \boldsymbol{W}(\boldsymbol{Y}_j - \boldsymbol{f}_0)$. Calculate sample covariance matrix, denoted as $\boldsymbol{\Lambda}_0^{\#}$, from $\{\boldsymbol{\omega}_j^{\#} : j = 1, \ldots, N\}$.
2. Apply Algorithm CMR (see Figure 3 in [33]) to regularize $\boldsymbol{\Lambda}_0^{\#}$ and obtain $\widetilde{\boldsymbol{\Lambda}}_0^{\#}$.
3. Determine the batch size r by Algorithm BSD (see Figure 4 in [33]) and calculate non-overlapping batch means, $\boldsymbol{\omega}_1^{\#}(r), \ldots, \boldsymbol{\omega}_{\lfloor N/r \rfloor}^{\#}(r)$, where $\boldsymbol{\omega}_k^{\#}(r) = r^{-1}\sum_{u=1}^{r}\boldsymbol{\omega}_{(k-1)r+u}^{\#}$. Let $\widetilde{\boldsymbol{\Lambda}}_0^{\#}(r) = \widetilde{\boldsymbol{\Lambda}}_0^{\#}/r$. The Hotelling's T^2-type statistic is constructed as
$$T_k^2(r) = \left(\boldsymbol{\omega}_k^{\#}(r)\right)^T \left(\widetilde{\boldsymbol{\Lambda}}_0^{\#}(r)\right)^{-1} \boldsymbol{\omega}_k^{\#}(r), \text{ for } k = 1, \ldots, \lfloor N/r \rfloor. \quad (8)$$
4. Calculate sample mean $\widehat{\mu}_{T^2(r)}$ and sample variance $\widehat{\sigma}_{T^2(r)}^2$ of $T_k^2(r)$. With $K = 0.1\widehat{\sigma}_{T^2(r)}$ and a prespecified in-control average run length ARL$_0$, solve the root H from Equation (12).

Phase II—Obtain new observed profiles $\{\boldsymbol{Y}_j : j = 1, 2, \ldots\}$ and compute non-overlapping batch means $\boldsymbol{Y}_k(r) = r^{-1}\sum_{u=1}^{r}\boldsymbol{Y}_{(k-1)r+u}$, for $k = 1, 2, \ldots$.

5. Compute the DWT of the kth batch mean $\boldsymbol{\omega}_k(r) = \boldsymbol{W}\boldsymbol{Y}_k(r)$ and its reduced-dimension version $\boldsymbol{\omega}_k^{\#}(r)$, for $k = 1, 2, \ldots$.
6. Raise an alarm when $S^+(k) \geq H$ or $S^-(k) \geq H$, where $S^{\pm}(k)$ is defined in (13).

Figure 11: Algorithmic description of WDFTC$_s$.

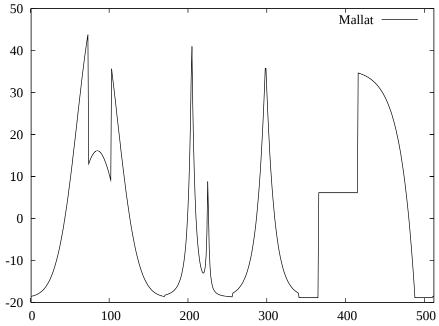

Figure 12: Mallat's piecewise smooth function.

problematic characteristics, such as cusps and jumps, which are often seen in manufacturing profiles. The noise components are assumed to follow independent and identically distributed (IID) standard normal distributions. We consider two types of shifts, namely the wavelet global (WG) shift and the wavelet local (WL) shift, with different values of the

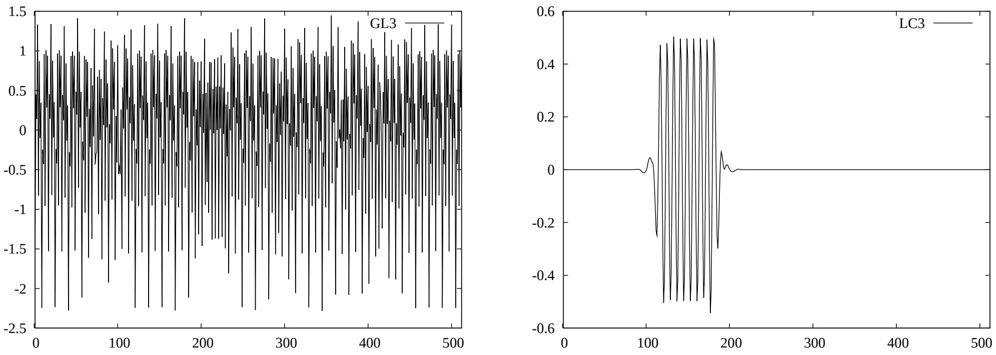

Figure 13: Wavelet global shift function (left) and wavelet local shift function (right).

shift size η; and thus the out-of-control mean profile is $\boldsymbol{f}_1 = \boldsymbol{f}_0 + \eta\boldsymbol{\zeta}$, where $\boldsymbol{\zeta}$ denotes the shift type and $\eta \in \{0.25, 0.5, 0.75, 1, 2\}$. These two types of shifts only affect unselected DWT components in WDFTC$_s$ and are shown in Figure 13. More details of WG and WL are in Section 3.3.1.1. The target ARL$_0$ is set to 200.

Table 7 shows ARL$_1$ of WDFTC$_s$ with different values of the shift size η. The performance of ARL$_1$ implies that WDFTC$_s$ fails to detect both types of shifts (WG and WL), even for a large shift size, such as $\eta = 2$. This is not surprising, because the WG and WL shifts do not affect the statistical properties of the preselected wavelet components; and thus constructed T^2-type statistics in WDFTC$_s$ have the same statistical properties (i.e., the mean and variance of the relevant T^2-type statistics do not change from in-control profiles to out-of-control profiles). Other existing wavelet-based control charts with a preselected subset of wavelet components also exhibit similar limitations. To overcome this problem, we consider DWTs of noise vectors of profiles and determine adaptively the locations of preserved wavelet components in the noise DWT vectors. More specifically, for each DWT of a noise vector, we preserve all scaling coefficients and a number of detail coefficients whose values turn out to be outliers. Therefore locations of preserved wavelet components can be different for each observed profile and can be affected by any type of shift.

To emphasize the need for a new chart, extreme shifts are considered in this section. Comparisons between WDFTC$_s$ and the proposed new chart on more realistic shifts are discussed in Section 3.3.

Table 7: ARL$_1$ delivered by WDFTC$_s$ for wavelet global shift and wavelet local shift.

	Wavelet global shift (WG)					Wavelet local shift (WL)				
η	0.25	0.5	0.75	1	2	0.25	0.5	0.75	1	2
ARL$_1$	202.72	195.54	198.19	203.19	189.25	212.47	206.64	202.42	201.03	202.22

3.1.3 Discrete Wavelet Transform

The wavelet transform is analogous to the Fourier transform, but it employs scaling and wavelet basis functions instead of the sine and cosine basis functions used in the Fourier transform. Let $\mathcal{L}^2[0,1]$ denote the space of real-valued square-integrable functions defined on the unit interval $[0,1]$. The wavelet transform of a function $g \in \mathcal{L}^2[0,1]$ represents g as an infinite series involving orthonormal scaling or wavelet basis functions. A scaling function $\phi \in \mathcal{L}^2[0,1]$ has an associated wavelet function $\psi \in \mathcal{L}^2[0,1]$; and from ϕ or ψ, we can derive an orthonormal set of basis functions for $\mathcal{L}^2[0,1]$ similar to the trigonometric functions used in the Fourier series representation. Assuming that ϕ and ψ are based on the Haar wavelet for simplicity, we explain the mechanism of the wavelet transform.

The wavelet representation of a function $g \in \mathcal{L}^2[0,1]$ is given by

$$g(z) = \lim_{B\to\infty} \sum_{\ell=-\infty}^{B-1} \sum_{m=0}^{\lceil 2^\ell \rceil -1} \langle g, \psi_{\ell,m}\rangle \psi_{\ell,m}(z) = \lim_{B\to\infty} \sum_{m=0}^{2^B-1} \langle g, \phi_{B,m}\rangle \phi_{B,m}(z) \tag{9}$$

for almost all $z \in [0,1]$, where: $h_{\ell,m}(z) = 2^{\ell/2}h(2^\ell z - m)$ for $h = \psi, \phi$; and $\langle g_1, g_2\rangle = \int_0^1 g_1(z)g_2(z)\,\mathrm{d}z$ is defined as the inner product operator for $g_1, g_2 \in \mathcal{L}^2[0,1]$.

Let $P_B(g)$ represent the Bth partial sum on the far right-hand side of Equation (9). It is clear that the quantity $P_B(g)$ becomes more accurate as B increases, and it provides an approximation to g for a finite B. The scaling coefficients of g are defined as the quantities $\{\mathsf{C}_{\ell,m} = \langle g, \phi_{\ell,m}\rangle\}$, which represent the low-frequency components or the smooth parts of $g(z)$. The detail coefficients of g are the quantities $\{\mathsf{D}_{\ell,m} = \langle g, \psi_{\ell,m}\rangle\}$; and they represent the high-frequency components or local behaviors of $g(z)$. Due to the discrete nature of measurements from a physical device, suppose that we take $g \approx P_J(g)$ for some finest (highest) level of resolution J; and we stop the successive function-approximation operations at some coarsest (lowest) level of resolution L, where $L < J$. Then we obtain an approximate representation of g based on its DWT,

$$g(z) \approx \sum_{m=0}^{2^J-1} \mathsf{C}_{J,m}\phi_{J,m}(z) = \sum_{m=0}^{2^L-1} \mathsf{C}_{L,m}\phi_{L,m}(z) + \sum_{\ell=L}^{J-1} \sum_{m=0}^{2^\ell-1} \mathsf{D}_{\ell,m}\psi_{\ell,m}(z) \tag{10}$$

for almost all $z \in [0, 1]$.

The DWT can also be presented in terms of matrices, and we use the matrix representation throughout the chapter. Let $\boldsymbol{W}$ denote an $n \times n$ orthogonal matrix associated with the selected functions $\phi(t)$ and $\psi(t)$, where n has the form $n = 2^J$ for some positive integer J. For any $n \times 1$ vector $\boldsymbol{Y}$, the matrix-vector multiplication $\boldsymbol{d} = \boldsymbol{W}\boldsymbol{Y}$ yields the corresponding DWT of $\boldsymbol{Y}$. Let $L \in \{0, 1, \ldots, J-1\}$ denote the coarsest level of resolution in the given wavelet system. Thus the first 2^L components of $\boldsymbol{d}$ are scaling coefficients, and all remaining components are detail coefficients.

3.2 New Procedure

In this section, we introduce a new chart that monitors statistics constructed from adaptively thresholded DWT vectors. We first present our proposed procedure WDFTC$_a$, and then discuss different adaptive thresholding methods and covariance matrix estimation.

3.2.1 Procedure WDFTC$_a$

WDFTC$_a$ first applies DWT to each observed profile, thresholds components of DWT vectors adaptively, and then constructs T^2-type statistics from thresholded DWT vectors. In order to maintain the power of control charts [16], a desirable thresholding method selects a small number of components (e.g., mainly scaling coefficients) when in-control, so that monitored statistics have manageable variability and can be approximated as a Brownian motion process properly. On the other hand, the thresholding method should select more components (e.g., including some extreme outliers of detail coefficients) when out-of-control, so that monitored statistics have significant changes and can quickly detect shifts in the mean profile. Additionally, scaling coefficients contain prominent information of a profile due to the nature of wavelet transform, while extreme outliers in DWT vectors vary for different observed profiles and thus should be selected adaptively.

Combining with an adaptive thresholding method, WDFTC$_a$ becomes more sensitive to detect a wide range of shift types than existing wavelet-based control charts with static selection. The detailed steps in WDFTC$_a$ are described in Figure 14 as below.

Procedure WDFTC$_a$

Phase I — Obtain vectors $\boldsymbol{d}_j = \boldsymbol{Y}_j - \boldsymbol{f}_0$, from observed profiles $\{\boldsymbol{Y}_j : j = 1, \ldots, N\}$ where N is the size of Phase I data set. Calculate DWTs $\boldsymbol{\omega}_j = \boldsymbol{W}\boldsymbol{d}_j$, and then perform the following steps:

1. Compute sample covariance matrix $\widehat{\boldsymbol{\Lambda}}_0$ from $\{\boldsymbol{\omega}_j : j = 1, \ldots, N\}$, see Notation Summary. Apply Algorithm CMR (Figure 15) as follow to regularize $\widehat{\boldsymbol{\Lambda}}_0$ and obtain $\widetilde{\boldsymbol{\Lambda}}_0$. Next determine the batch size r by Algorithm BSD (Figure 16) as follow and let $\widetilde{\boldsymbol{\Lambda}}_0(r) = \widetilde{\boldsymbol{\Lambda}}_0/r$. For $k = 1, 2, \ldots, \lfloor N/r \rfloor$, compute the kth non-overlapping batch mean vector $\boldsymbol{\omega}_k(r) = r^{-1}\sum_{u=1}^{r}\boldsymbol{\omega}_{(k-1)r+u}$.
2. Determine the lower and upper threshold vectors $\boldsymbol{\lambda}^L$ and $\boldsymbol{\lambda}^U$ as Section 3.2.2.
3. Threshold $\boldsymbol{\omega}_k(r)$ with $\boldsymbol{\lambda}^L$ and $\boldsymbol{\lambda}^U$; and obtain $\boldsymbol{\omega}^*_k(r)$, see Notation Summary. Compute the Hotelling's T^2-type statistics
$$T^2_k(r) = \left(\boldsymbol{\omega}^*_k(r)\right)^T \left(\widetilde{\boldsymbol{\Lambda}}_0(r)\right)^{-1} \left(\boldsymbol{\omega}^*_k(r)\right) \qquad (11)$$
4. Calculate sample mean $\widehat{\mu}_{T^2(r)}$ and sample variance $\widehat{\sigma}^2_{T^2(r)}$ from $\{T^2_k(r) : k = 1, 2, \ldots, \lfloor N/r \rfloor\}$. Solve a root H from Equation (12) with $K = 0.1\widehat{\sigma}_{T^2(r)}$ and prespecified ARL$_0$.
$$\frac{\widehat{\sigma}^2_{T^2(r)}}{2K^2}\left(\exp\left\{\frac{2K\left[H + 1.166\widehat{\sigma}_{T^2(r)}\right]}{\widehat{\sigma}^2_{T^2(r)}}\right\} - 1 - \left\{\frac{2K\left[H + 1.166\widehat{\sigma}_{T^2(r)}\right]}{\widehat{\sigma}^2_{T^2(r)}}\right\}\right) = \frac{2\mathrm{ARL}_0}{r}, \qquad (12)$$

Phase II — Obtain DWTs $\boldsymbol{\omega}_j = \boldsymbol{W}(\boldsymbol{Y}_j - \boldsymbol{f}_0)$ from new observed profiles $\boldsymbol{Y}_j$, $j = 1, 2, \ldots$, and non-overlapping batch mean vector $\boldsymbol{\omega}_k(r) = r^{-1}\sum_{u=1}^{r}\boldsymbol{\omega}_{(k-1)r+u}$, $k = 1, 2, \ldots$.

5. Threshold $\boldsymbol{\omega}_k(r)$ with $\boldsymbol{\lambda}^L$ and $\boldsymbol{\lambda}^U$; and obtain thresholded DWTs $\boldsymbol{\omega}^*_k(r)$. Calculate the associated statistic $T^2_k(r)$ from (11).
6. Raise an alarm for $\boldsymbol{\omega}_k(r)$ if $S^+(k) \geq H$ or $S^-(k) \geq H$, where
$$S^{\pm}(k) = \begin{cases} 0, & \text{for } k = 0, \\ \max\left\{0, S^{\pm}(k-1) \pm (T^2_k(r) - \widehat{\mu}_{T^2(r)}) - K\right\}, & \text{for } k = 1, 2, \ldots. \end{cases} \qquad (13)$$

Figure 14: Algorithmic description of WDFTC$_a$.

Remark 1 *In some applications, the in-control mean profile* $\boldsymbol{f}_0$ *may not be known. As an alternative, we use sample mean of Phase I data* $\widehat{\boldsymbol{f}}_0 = \sum_{j=1}^{N}\boldsymbol{Y}_j/N$ *as an estimator of* $\boldsymbol{f}_0$*. Correspondingly,* $\boldsymbol{\theta}_0$ *is replaced with* $\widehat{\boldsymbol{\theta}}_0 = \boldsymbol{W}\widehat{\boldsymbol{f}}_0$*.*

Remark 2 *Lee et al. [33] show batching reduces covariances and improves the ARL*$_1$ *performance of control charts.*

Algorithm CMR

1. Divide the Phase I data set into two disjoint subsets $\mathcal{S}_1$ and $\mathcal{S}_2$ with size $N_1 = \lfloor N(1 - \log N)/\log N \rfloor$ and $N_2 = N - N_1$, respectively.
2. Calculate the sample covariance matrices $\widehat{\mathbf{\Lambda}}^{\#}_{\mathcal{S}_1}$ and $\widehat{\mathbf{\Lambda}}^{\#}_{\mathcal{S}_2}$.
3. Compute the estimated threshold,

$$\widehat{\tau} = \arg\min_{\widehat{\tau}\geq 0} \sum_{\substack{(\mu\neq\nu)\text{ and}\\(2^L<\mu\text{ or }2^L<\nu)}} \left\{ \big[\widehat{\mathbf{\Lambda}}^{\#}_{\mathcal{S}_1}\big]_{\mu,\nu} 1\big[\big|\big[\widehat{\mathbf{\Lambda}}^{\#}_{\mathcal{S}_1}\big]_{\mu,\nu}\big| \geq \widehat{\tau}\big] - \big[\widehat{\mathbf{\Lambda}}^{\#}_{\mathcal{S}_2}\big]_{\mu,\nu}\right\}^2. \tag{14}$$

4. Calculate the sample covariance matrix $\widehat{\mathbf{\Lambda}}^{\#}_0$ using the entire Phase I data set of size N and apply the threshold $\widehat{\tau}$ from (14) to $\widehat{\mathbf{\Lambda}}^{\#}_0$ and obtain the regularized sample covariance matrix $\widetilde{\mathbf{\Lambda}}^{\#}_0$ as below.

$$\widetilde{\mathbf{\Lambda}}^{\#}_0 = \begin{cases} \big[\widehat{\mathbf{\Lambda}}^{\#}_0\big]_{\mu,\nu}, & \text{if (i) } \mu\neq\nu \text{ or (ii) } \mu\leq 2^L \text{ and } \nu\leq 2^L, \\ \big[\widehat{\mathbf{\Lambda}}^{\#}_0\big]_{\mu,\nu} 1\Big(\big|\big[\widehat{\mathbf{\Lambda}}^{\#}_0\big]_{\mu,\nu}\big| \geq \widehat{\tau}\Big), & \text{otherwise,} \end{cases}$$

where $1(\cdot)$ is the indicator function.

Figure 15: Algorithmic description of CMR.

WDFTC$_a$ and WDFTC$_s$ have quite different mechanisms. WDFTC$_s$ analyzes the DWT of the in-control mean profile and only monitors a subset of DWT components, while WDFTC$_a$ watches nonzero components from thresholded DWTs, whose locations vary for different observed profiles. More specifically, WDFTC$_s$ constructs T^2-type statistics by $p \times 1$ reduced-dimension DWTs of profiles and the corresponding $p \times p$ covariance matrix. On the other hand, WDFTC$_a$ constructs T^2-type statistics by $n \times 1$ thresholded DWTs of profiles and the $n \times n$ covariance matrix, but the number of nonzero components of thresholded DWTs is far smaller than n when in-control, so matrix multiplications to calculate monitored statistics can be accelerated by exploring sparsity of thresholded DWTs.

3.2.2 Thresholding Methods

Many researchers have developed wavelet thresholding methods with Gaussian noise as a canonical model. For example, minimax thresholding method and universal thresholding method from [14] are used for Gaussian white noise, while a level-dependent thresholding method from [26] works for cross-correlated Gaussian noise. Since WDFTC$_a$ is designed

Algorithm BSD

1. Obtain $\widehat{\tau}$ and the regularized sample covariance matrix $\widetilde{\mathbf{\Lambda}}_0^{\#}$ from Algorithm CMR. Set a subset $\mathcal{O} = \{(\mu,\nu) : \mu \neq \nu, 2^L < \mu \leq n, \text{ and } 2^L < \nu \leq n\}$ containing all off-diagonal elements of $\widetilde{\mathbf{\Lambda}}_0^{\#}$ but excluding the estimated covariances between pairs of scaling coefficients.
2. Let $\mathcal{Q}$ denote the number of nonzero elements in $\mathcal{O}$.

$$\mathcal{Q} = \sum_{(\mu,\nu)\in\mathcal{O}} 1\big(\big|[\widetilde{\mathbf{\Lambda}}_0^{\#}]_{\mu,\nu}\big| > 0\big)\,.$$

2a. If $\mathcal{Q} = 0$, then set $r = 1$ and stop. Otherwise, go to step [2b].

2b. Calculate the average magnitude ζ of nonzero elements in $\mathcal{O}$. Then set the batch size $r = \lceil \sqrt{2}\zeta/\widehat{\tau}\, \rceil$ and stop.

$$\zeta = \sum_{(\mu,\nu)\in\mathcal{O}} \big|[\widetilde{\mathbf{\Lambda}}_0^{\#}]_{\mu,\nu}\big|/\mathcal{Q}.$$

Figure 16: Algorithmic description of BSD.

to be distribution-free, we need a thresholding method that is both distribution-free and effective in the presence of cross correlations among noise components. In this subsection, we propose a distribution-free thresholding method to incorporate with WDFTC$_a$.

When the monitored process is in-control, the difference between an observed profile and the in-control mean profile has zero expected value, $\mathrm{E}[\boldsymbol{d}_j] = \mathrm{E}[\boldsymbol{Y}_j - \boldsymbol{f}_0] = 0$, so does its DWT vector $\mathrm{E}[\boldsymbol{\omega}_j] = \mathrm{E}[\boldsymbol{W}\boldsymbol{d}_j] = 0$. Thus it would be best to suppress all components of $\boldsymbol{\omega}_j$. Some existing thresholding methods can achieve this goal at least in the asymptotic sense for Gaussian white noise model, for example, universal thresholds that are mentioned above. On the other hand, when a shift occurs, the expected value of the difference between an out-of-control profile $\boldsymbol{Y}_j$ and an in-control mean profile $\boldsymbol{f}_0$ becomes $\boldsymbol{f}_1 - \boldsymbol{f}_0 \neq \mathbf{0}$, so it is desirable to keep more components after thresholding so that shifts can be detected faster. This implies that a desired thresholding method in WDFTC$_a$ should satisfy the following three critical properties: (a) it is distribution-free, (b) it keeps scaling coefficients but filters out most detail coefficients when in-control, and (c) it keeps some detail coefficients in addition to scaling coefficients when out-of-control.

Property (b) helps monitored statistics satisfy the underlying assumption of Equation (12) (i.e., cumulative sum statistics $S^{\pm}(k)$ behave like reflected Brownian motion processes when in-control). If we threshold both scaling coefficients and detail coefficients, it is possible that there is no component survived after thresholding for some profiles. Thus we may often obtain statistics with zero values, and behaviors of such statistics are hardly approximated well by a Brownian motion process. Additionally, scaling coefficients contain important information of profiles due to the nature of wavelet transform, such as the magnitude of noise. Therefore we prefer to keep all scaling coefficients without thresholding. On the other hand, if the number of preserved detail coefficients changes significantly from one profile to another when in-control, which is often caused by small thresholds or high correlations between components, the variability of T^2-type statistics becomes extremely large and it takes many observations until $S^{\pm}(k)$ exhibits appropriate asymptotic behaviors. As a result, actual ARL_0 of the control chart may be larger than the target value for finite target ARL_0. To ensure well-behaved T^2-type statistics when in-control, Property (b) is required.

Property (c) is critical to the effectiveness of control charts. Thresholds cannot be too large to filter out all of the important coefficients when out-of-control, since shift information can also be filtered out. This would cause a control chart inefficient in shift detection. It is recommended that one chooses relatively large but not extremely large thresholds and keep more detail coefficients when out-of-control, so that T^2-type statistics become larger and $S^{\pm}(k)$ exit control limits faster when a shift occurs. To ensure Properties (b) and (c), we design our thresholding method in a way that it keeps all scaling coefficients and relatively large outliers of detail coefficients only.

Next, we review the universal thresholding method and then propose our new thresholding method developed from the fourth moment Cornish-Fisher expansion from [4].

3.2.2.1 Universal Thresholds

Universal thresholds, developed by Donoho and Johnstone [14], are one of the most well-known wavelet thresholds. They provide some desirable asymptotic characteristics, which are stated below.

Lemma 3 *If z_1, z_2,...,z_n are independent and identically distributed as a standard normal distribution, $P\{\max_{1\leq i\leq n}|z_i| > \sqrt{2\log n}\} \to 0$ as $n \to \infty$. Therefore,*

$$\lim_{n\to\infty} \frac{\max_{1\le i\le n} |z_i|}{\sqrt{2\log n}} = 1, \textit{ almost surely.}$$

The universal threshold for the ith noise component is defined as $\lambda_i = \sigma_i\sqrt{2\log n}$, where σ_i is the marginal standard deviation of the ith component. Because the term $\sqrt{2\log n}$ is independent from distribution parameters, this thresholding method is called universal and known to be robust to different signals or profiles.

Supported by Lemma 3, universal thresholds are designed to filter out all detail coefficients asymptotically, so that zero components in the underlying mean profile are estimated as zero with high probability. This implies that universal thresholds satisfy Property (b) and (c). However, those properties do not hold any longer when noises are non-normal and/or have cross correlations. In [1] and [2], Averkamp and Houdré present thresholding methods for non-normal noises. But their thresholds vary from distributions and have no closed forms, making it hard to implement. Besides, their thresholds also tend to select more components than universal thresholds, violating Property (b). Instead, we present a distribution-free thresholding method that has similar properties as the universal thresholding method.

3.2.2.2 Cornish-Fisher Expansion Thresholds

We interpret the universal threshold by connecting it to a hypothesis testing:

$$H_0 : \mathrm{E}[\boldsymbol{\omega}_{j,i}] = 0 \text{ versus } H_1 : \mathrm{E}[\boldsymbol{\omega}_{j,i}] \neq 0, \text{ for } 2^L < i \le n \text{ and } j = 1, 2, \ldots.$$

When $|\boldsymbol{\omega}_{j,i}| > \widehat{\sigma}_i\sqrt{2\log n}$, we reject H_0 and keep $\boldsymbol{\omega}_{j,i}$. Otherwise, we filter out $\boldsymbol{\omega}_{j,i}$ and set $\boldsymbol{\omega}_{j,i} = 0$. Let $q = \Phi(\sqrt{2\log n})$, where Φ represents the standard normal cumulative distribution function. If noises are assumed to follow IID normal distributions, universal thresholds can be considered as q-quantile estimates. We extend this idea to a general noise distribution as follows.

- If $\boldsymbol{\omega}_{j,i} < z_i^{1-q}$ or $\boldsymbol{\omega}_{j,i} > z_i^q$, reject H_0 and keep $\boldsymbol{\omega}_{j,i}$, where z_i^{1-q} and z_i^q are the $(1-q)$- and q-quantiles of the underlying distribution of $\boldsymbol{\omega}_{j,i}$.

- Otherwise, filter out $\boldsymbol{\omega}_{j,i}$ and set $\boldsymbol{\omega}_{j,i} = 0$.

In other words, critical quantile values z_i^{1-q} and z_i^q serve as the lower and upper thresholds for $\boldsymbol{\omega}_{j,i}$. By using quantiles, we can determine whether a wavelet component is significant with an approximately similar confidence level as universal thresholds, no matter what underlying distributions are. Moreover, as $n \to 0$, $q \to 1$, Lemma 3 still holds.

For profiles with a moderate dimension size $n = 512$, the probability from universal thresholds is $q = \Phi(\sqrt{2\log 512}) = 0.9998$. The 0.9998-quantile is considered as an extreme quantile. It is known that accurate estimation of extreme quantiles is difficult. However, for our purpose, we only need relatively good thresholds that are large enough to guarantee Properties (a)–(c), instead of accurate quantile estimates. We consider indirect quantile estimation with the Cornish-Fisher (CF) expansion from [4], because it can provide good quantile estimates with a moderate data set and low data storage requirements. Additionally, experimental results in [4] show applaudable performance with the fourth moment expansion. A q-quantile estimator from the CF expansion is defined as follows.

$$z_i^q = \widehat{\mu}_i + \widehat{\sigma}_i x_i^q, \text{ where } x_i^p = z_q + \frac{1}{6}(z_q^2 - 1)\widehat{\gamma}_{1i} + \frac{1}{24}(z_q^3 - 3z_q)\widehat{\gamma}_{2i} - \frac{1}{36}(2z_q^3 - 5z_q)\widehat{\gamma}_{1i}^2, \quad (15)$$

where z_q is the q-quantile of the standard normal distribution, and $\widehat{\mu}_i$, $\widehat{\gamma}_{1i}$ and $\widehat{\gamma}_{2i}$ are sample mean, sample central standardized skewness, and sample central standardized excess kurtosis for the ith component from Phase I data, respectively.

Bekki et al. [4] also point out that quantile estimates based on the CF expansion may be inaccurate for extreme quantiles for skewed distributions, but perform reasonably well for symmetric distributions. Since wavelet coefficients are weighted averages and tend to be more symmetric than the raw data, so the CF expansion provides relatively good thresholds for WDFTC$_a$.

Consider profiles with IID exponential noise distributions. We set the profile dimension $n = 512$ and each noise component to follow centralized exponential distributions with parameter 1 (i.e., $\boldsymbol{\epsilon}_{ij} \sim \exp(1) - 1$, $i = 1, 2, \ldots, n$, and $j = 1, 2, \ldots$). If we apply Haar DWT, the highest level of wavelet components is difference of two exponential random variables, which follows the distribution, Laplace$(0, 1)$. We calculate theoretical values of the 0.0002- and 0.9998-quantiles from Laplace$(0, 1)$ and also estimate them with the CF expansion based on 500 randomly sampled noise vectors. Since wavelet components at the highest level ($n/2 < i \leq n$) follow the same distribution, we can obtain average and standard error (SE) of the 0.0002- and 0.9998-quantile estimates based on results of $n/2$ components. Table 8 provides average and SE for the CF expansion (CF), and also provides theoretical values for the Laplace distribution (Laplace) and universal thresholds (Universal). For IID exponential noises, the CF expansion provides more accurate quantile estimates than universal thresholds and the CF estimates are also close to the theoretical values. On the other hand, universal thresholds ±3.53 are clearly off for such a non-normal distribution.

Table 8: Average and Standard Error (SE) of the 0.0002- and 0.9998-quantile estimates for wavelet components at the highest level.

	The 0.0002-quantile			The 0.9998-quantile		
	CF	Universal	Laplace	CF	Universal	Laplace
Average	7.324	3.53	7.824	-7.537	-3.53	-7.824
SE	0.091	N/A	N/A	0.105	N/A	N/A

3.2.2.3 Inflation Factor

With the probability $q = \Phi(\sqrt{2\log n})$ from universal thresholds, the CF expansion provides good thresholds for independent normal and non-normal profiles. However, when positive cross correlations exist, more components tend to survive after thresholding and associated T^2-type statistics change significantly from one profile to another even when in-control. It becomes more difficult to achieve a good approximation of a Brownian motion process, and thus Equation (12) does not work well. We want to adjust q value when there are strong correlations. To this end, we employ an inflation factor γ and redefine $q = \Phi(\gamma\sqrt{2\log n})$ for estimating quantiles z_{1-q} and z_q.

Recall that $\mathbf{\Lambda}$ denotes the profile covariance matrix. When all components in a profile's DWT are independent, $\mathbf{\Lambda}_{ij} = 0$ for $i \neq j$. When there exist cross correlations among components in a profile's DWT, $\mathbf{\Lambda}_{ij}$ becomes nonzero for some $i \neq j$. Since positive correlations among detail coefficients or among detail coefficients and scaling coefficients, cause more survived components after thresholding, we determine an inflation factor by comparing the actual covariance matrix $\mathbf{\Lambda}$ and its "ideal" covariance matrix $\boldsymbol{S}$, where $\boldsymbol{S}$ has the same marginal variances for all coefficients as $\mathbf{\Lambda}$ (i.e., $\boldsymbol{S}_{ii} = \mathbf{\Lambda}_{ii}$), but zero correlation (i.e., $\boldsymbol{S}_{ij} = 0$ for $i \neq j$). Parameter t, set in Equation (16), can be interpreted as a measure of similarity between the actual covariance matrix $\boldsymbol{\Lambda}$ and its “ideal” covariance matrix $\boldsymbol{S}$. If $\boldsymbol{\Lambda} = \boldsymbol{S}$, i.e., all profile variables are independent, t is equal to 1; while t becomes small if $\boldsymbol{\Lambda}$ and $\boldsymbol{S}$ are different.

When $\boldsymbol{\Lambda}$ and $\boldsymbol{S}$ are quite different, $1/\sqrt{t}$ turns to be very large. However, if an inflation factor is too large, less components survive and Property (c) is violated. Therefore we set an upper bound for an inflation factor in Equation (17). We recommend to take $\gamma_{max} = 1.5$ to maintain Property (c). Note that without an inflation factor, estimated quantiles by CF expansion are already quite large, therefore an inflation factor should be relatively small and 1.5 is a good empirical choice. Algorithm CFI as follow states procedures for computing an inflation factor γ.

Algorithm CFI

The inflation factor γ is determined by the following steps.

1. Let $\boldsymbol{\Lambda} = \widetilde{\boldsymbol{\Lambda}}_0(r)$, where $\widetilde{\boldsymbol{\Lambda}}_0(r)$ is a covariance matrix estimate in (11).
2. Set an $n \times n$ matrix $\boldsymbol{S}$ with $\boldsymbol{S}_{ii} = \boldsymbol{\Lambda}_{ii}$ and $\boldsymbol{S}_{ij} = 0$ for $i \neq j$ and $1 \leq i, j \leq n$.
3. Calculate a correlation-type statistic t as follows.
$$t = \frac{\sum_{1 \leq i,j \leq n} (\boldsymbol{\Lambda}_{ij} - \bar{\boldsymbol{\Lambda}})(\boldsymbol{S}_{ij} - \bar{\boldsymbol{S}})}{\sqrt{\left(\sum_{1 \leq i,j \leq n} (\boldsymbol{\Lambda}_{ij} - \bar{\boldsymbol{\Lambda}})^2\right)\left(\sum_{1 \leq i,j \leq n} (\boldsymbol{S}_{ij} - \bar{\boldsymbol{S}})^2\right)}}, \qquad (16)$$
where $\bar{\boldsymbol{A}} = \sum_{1 \leq i,j \leq n} \boldsymbol{A}_{ij}/n^2$ for any matrix $\boldsymbol{A}$.
4. Set an inflation factor
$$\gamma = \min(\frac{1}{\sqrt{t}}, \gamma_{max}) \qquad (17)$$
and let $q = \Phi(\gamma\sqrt{2 \log n})$.

Figure 17: Algorithmic description of Algorithm CFI.

We prove $0 < t \leq 1$ and $1 \leq \gamma \leq 1.5$ in the Appendix B. If profiles have independent noises, t equals to 1 and inflation factor γ also equals to 1, so the probability q is as same as the probability from universal thresholds. Additionally, a small t implies that a big discrepancy between the actual covariance matrix and the “ideal” case, and thus both γ and thresholds take larger values. On the other hand, when cross correlation is low, t takes large values and γ is small. Additionally, when negative correlations exist in $\boldsymbol{\Lambda}$, we get a larger t (or γ close to 1), compared to the case where all correlations are positive with the same level of correlation. WDFTC$_a$ with CF thresholds is given as below.

Step 2 of Procedure WDFTC$_a$

By employing CF thresholds, Step 2 of WDFTC$_a$ states as follows.

2. For $i = 1, \ldots, n$, calculate sample mean $\widehat{\mu}_i$, sample standard deviation $\widehat{\sigma}_i$, sample central standardized skewness $\widehat{\gamma}_{1i}$, and sample central standardized excess kurtosis $\widehat{\gamma}_{2i}$, from $\{\boldsymbol{\omega}_{k,i}(r) : k = 1, \ldots, \lfloor N/r \rfloor\}$, where $\boldsymbol{\omega}_{k,i}(r)$ is the ith component of the kth batch means vector $\boldsymbol{\omega}_k(r)$. Then determine the inflation factor γ as Equation (17) from Algorithm CFI. Obtain $q = \Phi(\gamma\sqrt{2\log n})$, and calculate corresponding estimated quantiles z_i^{1-q} and z_i^q from Equation (15). The lower and upper thresholds $\boldsymbol{\lambda}^L$ and $\boldsymbol{\lambda}^U$ are determined by

$$\boldsymbol{\lambda}^L = \{z_1^{1-q}, \ldots, z_n^{1-q}\} \text{ and } \boldsymbol{\lambda}^U = \{z_1^q, \ldots, z_n^q\}.$$

Figure 18: Algorithmic description of Step 2 of Procedure WDFTC$_a$ with CF thresholds.

3.2.3 Covariance Matrix Estimation

When WDFTC$_a$ constructs T^2-type statistics, a covariance matrix estimate of profile DWTs is required, if the true covariance matrix is unknown. Because the locations of survived DWT components vary from one profile to another, we need the information of the entire $n \times n$ covariance matrix.

Several researchers developed some covariance matrix estimation methods for high-dimensional vectors, such as [5] and [7]. We employ the regularization method from [5], because their method is robust to different forms of high-dimensional covariance matrices and easy to implement as shown in [33]. Bickel and Levina [5] divide the data set into two groups with the ratio $(\log N - 1) : 1$, where N is the size of the data set, and obtain sample covariance matrices for each group, $\widehat{\boldsymbol{\Sigma}}_1$ and $\widehat{\boldsymbol{\Sigma}}_2$, and search for a threshold s that minimizes the difference between the thresholded $\boldsymbol{\Sigma}_1$ by s and $\boldsymbol{\Sigma}_2$ in the Frobenius metric. Then the covariance matrix estimate is calculated by thresholding the sample covariance of the entire data set with s.

Algorithm CMR is the regularization method from [5] and also used in [33], but Lee et al. [33] use a splitting ratio $2 : 3$ that does not perform well for higher dimensional profiles. Note that we work with $n \times n$ matrix, while Lee et al. [33] work with $p \times p$ matrix, where $p \ll n$. We find that the original splitting ratio $(\log N - 1) : 1$ in [5] performs better. Thus

Algorithm CMR employs the splitting ratio $(\log N - 1) : 1$ for covariance matrix estimation.

3.3 Experiment

In this section, we test two different in-control mean profiles: Mallat's piecewise smooth function and the LRS data from a lumber manufacturing process. We present comprehensive experimental results of WDFTC$_a$ and compare with its competitor WDFTC$_s$ mentioned in Section 3.1.2. Recall that WDFTC$_s$ is an effective wavelet-based control chart with static selection, i.e., only monitoring a subset of wavelet components. On the other hand, WDFTC$_a$ monitors adaptively selected wavelet components that vary for different sample profiles.

3.3.1 Mallat's Piecewise Function

To investigate the effectiveness of WDFTC$_a$, we take Mallat's piecewise smooth function as the in-control mean profile and test processes on various noise distributions and shifts. We assume that the true covariance matrix of profiles is known for Mallat's piecewise smooth function, but this assumption is relaxed for the LRS data in the next subsection.

We take $n = 512$ equally spaced data points from each observed profile and apply Symmlet 8 wavelets with the coarsest level of resolution $L = 5$. The size of Phase I data set is $N = 20,000$. Set the target ARL$_0$ = 200 and actual ARL$_0$ and ARL$_1$ are calculated based on 1000 independent replications. Since we use the true covariance matrix, the batch size cannot be determined by a sample covariance matrix, we use the average batch size reported in [33] as our batch size for both WDFTC$_s$ and WDFTC$_a$. For example, the average batch size of WDFTC$_s$ for IID standard normal noises is 3 from Table 4 in [33], and thus we use 3 as batch size in both WDFTC$_s$ and WDFTC$_a$. The inflation factor is determined by comparing the true covariance matrix with its "ideal" case.

Additionally, the weight in WRRE for WDFTC$_s$ is set to 0.5 for Mallat's piecewise smooth function, and thus 62 fixed components are selected to construct their statistics.

3.3.1.1 Shift Configurations

The out-of-control mean profile is set to be $\boldsymbol{f}_1 = \boldsymbol{f}_0 + \eta\boldsymbol{\zeta}\boldsymbol{\sigma}$, where shift size $\eta \in \{0.25, 0.5, 0.75, 1, 2\}$ and $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_n)^T$ is the vector of marginal standard deviations of profiles. We define distinctive shifts by using different $\boldsymbol{\zeta}$. First, $\boldsymbol{\zeta}$ has a form of a diagonal matrix, i.e., $\boldsymbol{\zeta} = \boldsymbol{\Delta}$, where $\boldsymbol{\Delta} = \text{diag}(\delta_1, \ldots, \delta_n)$. We define four shifts as follows.

- Global straight line shift (G1): $\delta_i = 1$ for $i = 1, \ldots, n$.
- Global stepwise shift (G2): $\delta_i = 1$ for $i = 1, \ldots, n/2$ and $\delta_i = -1$ for $i = n/2+1, \ldots, n$.
- Local straight line shift (L1): $\delta_i = 1$ for $i \in \mathcal{A}_1 = \{3, 4, \ldots, 15\} \cup \{344, 345, \ldots, 347\}$ and $\delta_i = 0$ for $i \notin \mathcal{A}_1$.
- Local flare shift (L2): $\delta_i = (i - 480)/32$ for $i \in \mathcal{A}_2 = \{481, \ldots, 512\}$ and $\delta_i = 0$ for $i \notin \mathcal{A}_2$.

Next, $\boldsymbol{\zeta}$ changes its form to $\boldsymbol{\zeta} = \boldsymbol{W}^{-1}\boldsymbol{\Theta}$, where $\boldsymbol{\Theta} = \text{diag}(\theta_1, \ldots, \theta_n)$, a diagonal matrix defined as the same as Section 3.1.2, only affecting non-selected components from the static selection in WDFTC$_s$.

- Wavelet global shift (WG): $\theta_i = 1$ in $\boldsymbol{\zeta}_1$, for $i = 1, \ldots, n$, see Figure 13 (Left).
- Wavelet local shift (WL): $\theta_i = 1$ in $\boldsymbol{\zeta}_2$, for $i \in \mathcal{B}_1 = \{80, 81, \ldots, 88\}$ and $\theta_i = 0$ for $i \notin \mathcal{B}_1$, see Figure 13 (Right).

3.3.1.2 Noise Distributions

We consider three multivariate normal distributions, SMN, CMN and GMN, with mean zero and covariance matrices Σ_S, Σ_{CN}, and Σ_G, respectively, and two multivariate shifted exponential distributions, EXP and CEXP, with mean zero and covariance matrices Σ_E and Σ_{CE}, respectively. All five noise distributions SMN, CMN GMN, EXP and CEXP, with their covariance matrices are defined as below.

- SMN: $\Sigma_S = \boldsymbol{I}_n$, where $\boldsymbol{I}_n$ is an $n \times n$ identity matrix.

- CMN: All diagonal elements in Σ_{CN} are equal to 1 and all off-diagonal elements in Σ_{CN} have value 0.5.

- GMN: Diagonal elements $[\Sigma_G]_{ii}$ are taken from Example 2 of [18], defined as Equation (18) with values from 9.5 to 14.8, and off-diagonal elements are $[\Sigma_G]_{i_1,i_2} = [\Sigma_G]_{i_1 i_1}^{1/2}$ $[\Sigma_G]_{i_2 i_2}^{1/2}\, \rho(i_1 - i_2)$ for $i_1 \neq i_2$, where $\rho(\cdot)$ are determined by the following Equation (19), originally taken from [47].

$$[\Sigma_G]_{ii} = \sigma_0^2 \left(1 + \left\{0.5 - 2.5\big[(i-1)/n - 0.515\big]^2\right\}^2\right)^2, \text{ where } \sigma_0^2 = 9.50. \tag{18}$$

$$\begin{aligned} \rho(\ell) = \mathrm{Corr}[\epsilon_{\mathrm{i,j}}, \epsilon_{\mathrm{i}+\ell,\mathrm{j}}] = (-\alpha_2)^{|\ell/2|} \left[\frac{\sin\left(|\ell|\boldsymbol{\omega} + \xi\right)}{\sin\left(\xi\right)}\right], \text{ where } \ell = 0, \pm 1, \ldots, \pm(\mathrm{n}-1), \alpha_1 = \frac{4}{3}, \\ \alpha_2 = -\frac{8}{9}, \boldsymbol{\omega} = \cos^{-1}\left[\frac{\alpha_1}{2\sqrt{-\alpha_2}}\right] \cong 0.785, \xi = \tan^{-1}\left[\frac{1-\alpha_2}{1+\alpha_2}\tan(\boldsymbol{\omega})\right] \cong 1.51. \end{aligned} \tag{19}$$

- EXP: $\Sigma_E = \boldsymbol{I}_n$, the same as Σ_S.

- CEXP: We employ the NORTA method in [8] to generate CEXP, by transforming a CMN vector into a shifted exponential (CEXP) vector. Thus the covariance matrix Σ_{CE} has diagonal elements $[\Sigma_{CE}]_{ii} = 1$ and off-diagonal elements $[\Sigma_{CE}]_{i_1,i_2}$ close to 0.5 but slightly less than 0.5 on the average.

3.3.1.3 Results

We first present the actual values of ARL_0 delivered by WDFTC_s and WDFTC_a for various distributions. Table 9 shows that WDFTC_s delivers an actual value of ARL_0 close to the target, while WDFTC_a delivers an actual value of ARL_0 that slightly deviates from the target. This is mainly due to the larger variance of the T^2-type statistics used in WDFTC_a compared with WDFTC_s. Thus we get a less accurate estimate of the relevant covariance matrix with $20,000$ Phase I data. However, we confirm that this problem goes away when we increase Phase I data. Nevertheless, the deviation from target 200 in Table 9 is acceptable.

Tables 10–11 compare ARL_1 between WDFTC_s and WDFTC_a. For all tested noise distributions, both control charts are able to detect the first four shifts (G1, G2, L1, and L2), but WDFTC_a outperforms WDFTC_s. Moreover, WDFTC_a performs significantly better than WDFTC_s on small shift size. For example, when shift size $\eta = 0.25$, WDFTC_a detects local shifts L1 and L2 by 9.4% to 32.2% faster than WDFTC_s, where local shifts

Table 9: Actual ARL_0 from WDFTC_s and WDFTC_a for Mallat's piecewise smooth function under various noise distributions.

Noise type	$\bar{r}$	γ	WDFTC_s	WDFTC_a
SMN	3	1.00	202.61	190.62
CMN	3	1.50	204.22	199.58
GMN	8	1.29	196.42	197.26
EXP	3	1.00	196.57	195.91
CEXP	3	1.50	203.99	214.33

are more difficult to detect than global ones. Furthermore, WDFTC_a can effectively detect wavelet shifts $\boldsymbol{\zeta}_1$ and $\boldsymbol{\zeta}_2$, where WDFTC_s fails. Even though correlations in the CEXP noise distribution lead to a large inflation factor and make it difficult for WDFTC_a to detect wavelet shifts until the shift size η reaches 2, WDFTC_a still performs better than WDFTC_s because WDFTC_s cannot detect wavelet shifts of any size. Results with the theoretical covariance matrix for Mallat's piecewise smooth function show that WDFTC_a performs better than WDFTC_s in most scenarios and also detects shifts missed by WDFTC_s.

3.3.2 Lumber Manufacturing Process

We compare experimental results between WDFTC_s and WDFTC_a for LRS data from a lumber manufacturing process. LRS data are generated by a statistical model developed in [46] and implemented in [33, 44, 45]. An LRS profile is the thickness of a large number of tested points on a sawed board, measured along a vertical line from an edge of the board, by an LRS at a certain location of the sawed board.

Given a saw configuration, a board type and a LRS location, a sampled LRS profile is $\boldsymbol{Y}_j = \{Y_{j,1}, \ldots, Y_{j,n}\}$, where $Y_{j,i} = f_{0,i} + \epsilon_{j,i}$ is the measured thickness of the jth board at the ith horizontal distance x_i cm along the length of the board, $i = 1, \ldots, n$, and $f_{0,i}$ denotes the true mean of thickness at the same tested point. $\boldsymbol{f}_0 = \{f_{0,1}, \ldots, f_{0,n}\}$ is the in-control mean of LRS data, taken over the population. However, $\boldsymbol{f}_0$ shown in Figure 19, cannot be presented by an explicit form.

We take $n = 2048$ tested points on each sawed board and again use Symmlet 8 wavelets with the coarsest level of resolution $L = 5$. Set the target ARL_0 to 370 and the size of Phase I data to $N = 30,000$. Delivered ARL's are calculated based on 1000 independent

Table 10: ARL_1 for Mallat's piecewise smooth function with SMN, CMN and GMN noise vectors.

		SMN		CMN		GMN	
Shift type	η	WDFTC_s	WDFTC_a	WDFTC_s	WDFTC_a	WDFTC_s	WDFTC_a
G1	0.25	3.81	3.44	195.45	191.12	8.00	8.00
	0.5	3.00	3.00	150.01	123.36	8.00	8.00
	0.75	3.00	3.00	86.27	62.03	8.00	8.00
	1	3.00	3.00	49.86	34.14	8.00	8.00
	2	3.00	3.00	12.69	9.38	8.00	8.00
G2	0.25	3.88	3.51	3.00	3.00	8.00	8.00
	0.5	3.00	3.00	3.00	3.00	8.00	8.00
	0.75	3.00	3.00	3.00	3.00	8.00	8.00
	1	3.00	3.00	3.00	3.00	8.00	8.00
	2	3.00	3.00	3.00	3.00	8.00	8.00
L1	0.25	123.58	103.38	73.12	51.56	40.42	28.34
	0.5	33.63	30.83	18.15	13.16	11.81	9.08
	0.75	15.58	13.78	8.65	6.69	8.02	8.00
	1	9.09	8.08	5.55	4.12	8.00	8.00
	2	3.03	3.00	3.00	3.00	8.00	8.00
L2	0.25	141.21	116.22	84.31	60.26	44.88	30.43
	0.5	40.26	35.06	21.31	15.20	13.58	9.76
	0.75	17.84	15.89	9.80	7.38	8.07	8.00
	1	10.47	9.47	6.18	4.77	8.00	8.00
	2	3.16	3.07	3.00	3.00	8.00	8.00
WG	0.25	200.64	58.88	191.77	189.33	203.66	8.00
	0.5	197.93	11.96	198.71	139.10	203.87	8.00
	0.75	198.33	4.19	213.91	31.00	201.18	8.00
	1	195.84	3.01	202.34	6.14	202.12	8.00
	2	201.35	3.00	198.42	3.00	190.07	8.00
WL	0.25	205.53	161.06	199.45	196.02	203.02	190.21
	0.5	207.34	49.86	210.75	125.44	188.08	145.39
	0.75	202.78	11.03	202.32	13.24	204.72	58.94
	1	200.93	4.39	200.85	3.43	200.82	18.97
	2	200.46	3.00	193.36	3.00	202.68	8.00

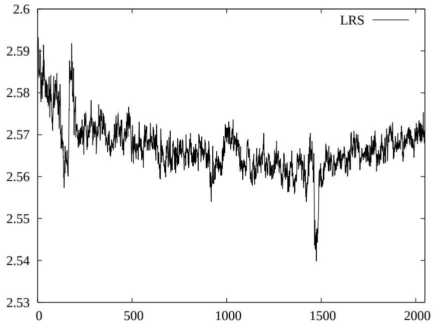

Figure 19: LRS data from a lumber manufacturing process (cm).

Table 11: ARL$_1$ for Mallat's piecewise smooth function with EXP and CEXP noise vectors.

		EXP		CEXP	
Shift type	η	WDFTC$_s$	WDFTC$_a$	WDFTC$_s$	WDFTC$_a$
G1	0.25	4.23	4.32	199.33	172.44
	0.5	3.00	3.00	171.77	153.04
	0.75	3.00	3.00	157.02	111.57
	1	3.00	3.00	131.06	75.63
	2	3.00	3.00	43.50	19.82
G2	0.25	4.35	4.39	6.05	3.34
	0.5	3.00	3.00	3.00	3.00
	0.75	3.00	3.00	3.00	3.00
	1	3.00	3.00	3.00	3.00
	2	3.00	3.00	3.00	3.00
L1	0.25	139.06	124.92	185.56	138.96
	0.5	38.49	36.97	66.38	34.26
	0.75	17.20	16.54	27.54	15.12
	1	9.93	9.67	15.37	8.84
	2	3.09	3.02	5.06	3.00
L2	0.25	145.67	131.99	191.12	148.23
	0.5	44.50	43.58	76.19	41.38
	0.75	19.94	19.45	33.38	17.61
	1	11.22	11.08	17.87	10.23
	2	3.43	3.33	5.57	3.01
WG	0.25	206.36	117.92	202.43	218.82
	0.5	199.38	38.63	209.50	215.50
	0.75	201.11	13.64	208.58	211.14
	1	197.65	6.39	210.53	207.17
	2	208.70	3.00	203.63	160.85
WL	0.25	203.22	178.99	207.62	211.78
	0.5	203.96	99.86	210.83	206.97
	0.75	202.04	22.81	201.89	211.72
	1	211.41	6.96	209.04	203.71
	2	197.41	3.00	211.53	5.31

replications. For LRS data, we estimate covariance matrix by using Algorithm CMR with ratio $(\log N - 1) : 1$.

Additionally, the weight in WRRE for WDFTC$_s$ is set to 0.7 for LRS data, and thus 92 fixed components are selected to construct their statistics.

3.3.2.1 Noise Distributions

As mentioned early, Staudhammer et al. [44] points that the thickness of a sawed board is affected by a saw configuration, a board type and an LRS location, and thus noise components have the form

$$\epsilon_{j,i} = f(x_i) + \mathcal{B}_j + \mathcal{L}_v + \mathcal{B}L_{jv} + \ell_{jvi} \text{ for } i = 1, \ldots, n, \tag{20}$$

where: (a) $\mathcal{B}_j \sim N(0, \sigma_{\mathcal{B}}^2)$ with $\sigma_{\mathcal{B}} = 0.0204$ cm, is the random effect of the jth sample board; (b) $\mathcal{L}_v \sim N(0, \sigma_{\mathcal{L}}^2)$ with $\sigma_{\mathcal{L}} = 0.0052$ cm, is the random effect of the vth laser location; (c) $\mathcal{B}L_{uv} \sim N(0, \sigma_{\mathcal{BL}}^2)$ with $\sigma_{\mathcal{B}L} = 0.0238$ cm, is the random effect caused by the interaction of the board and laser-location effects; (d) ℓ_{jvi} is the random affect arising from the interaction of the board, laser-location and the distance x_i along the board, so that the process $\{\ell_{jvi} : i = 1, \ldots, n\}$ is assumed to be stationary and represented by an ARIMA(1,1,1) time series model, where

$$(1 - \alpha\mathcal{B})(\ell_{jvi} - \ell_{jv(i-1)}) = (1 - \beta\mathcal{B})\varepsilon_i \text{ for } i = 1, 2, \ldots, \tag{21}$$

where: (a) $\mathcal{B}$ is the backshift operator so that $(1 - \alpha\mathcal{B})\ell_{jvi} = \ell_{jvi} - \alpha\ell_{jv(i-1)}$; and (b) $\{\varepsilon_i : i = 1, 2, \ldots\}$ is a white noise process, where $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$ with $\widehat{\sigma}_\varepsilon = 0.00967$ cm. Staudhammer et al. [44] use the autoregressive parameter $\widehat{\alpha} = 0.00053$ cm and the moving-average parameter $\widehat{\beta} = 0.00178$ cm.

3.3.2.2 Shift Configurations

We take the same four types of process faults as [33]: machine positioning problem (MPP), taper, flare and snake, see Figure 20, and briefly explain their causes and associated shifts as follows. Set $\boldsymbol{\delta} = \{\delta_1, \ldots, \delta_n\}$ as an associated shift vector, and thus the out-of-control mean profile has the form $\boldsymbol{f}_1 = \boldsymbol{f}_0 + \boldsymbol{\delta}$. More details can be found in [44] and [45].

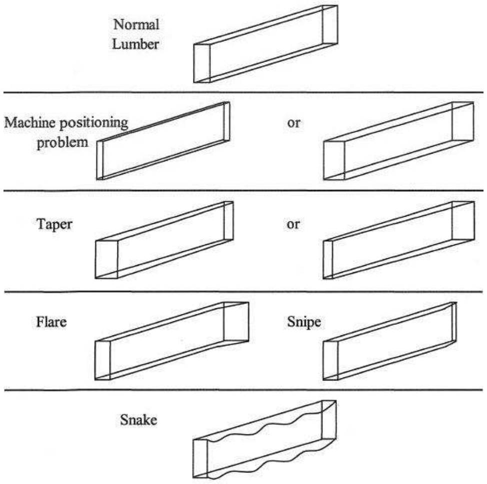

Figure 20: Four major types of process faults.

- MPP is caused by incorrect positions of saw guides and results in a uniform change in thickness along the length of the board, with $\delta_i = \tau$ for $i = 1, \ldots, n$
- Taper is caused by machine misalignment and results in a gradual increment or decrement in the board thickness along the length of the board, with $\delta_i = x_i\tau/x_n$ for $i = 1, \ldots, n$.
- Flare is caused when feed roll does not engage at a proper time, and results in an unexpected increment in the board thickness at the end of the board, with $\delta_i = 0$ for $i < i_0$ and $\delta_i = (x_i - x_{i_0})\tau/(x_n - x_{i_0})$ for $i \geq i_0$, where $i_0 = \max\{i : x_i < x_n - 15\}$.
- Snake is usually caused by several saw problems and results in a waveform with the period along the length of the board, with $\delta_i = \tau \sin(2\pi x_i/\mathsf{P})$ for $i = 1, \ldots, n$, where τ is the amplitude and the period $\mathsf{P} = 182.88$ cm.

Parameter τ takes values from $\{0.0254, 0.0508, 0.0762, 0.1016\}$ cm for MPP, while $\tau \in \{0.0508, 0.1016, 0.1524, 0.2032\}$ cm for others.

3.3.2.3 Results

From Algorithm CFI, we obtain an inflation factor for WDFTC$_a$ from each replication and the average of 1000 replications is $\bar{\gamma} = 1.35$. Table 12 show ARL$_0$ and ARL$_1$ for both WDFTC$_s$ and WDFTC$_a$. Both charts deliver actual ARL$_0$ close to the target value, but WDFTC$_a$ performs significantly better than WDFTC$_s$ for all shifts and detects some shifts with almost half observed profiles compared to WDFTC$_s$. Similar to Mallat case, WDFTC$_a$ achieves more ARL$_1$ saving on small shifts. For some shifts that are difficult to detect, such as flare, WDFTC$_a$ uses less than 28.9% to 42.9% observed profiles than WDFTC$_s$, and this saving is larger than some global shifts, such as MPP.

WDFTC$_a$ keeps only 32.000022 components on the average when in-control, including all $2^L = 32$ scaling coefficients. On the other hand, WDFTC$_s$ selects 92 components for each observed profile. WDFTC$_a$ does not only select significantly fewer wavelet components than WDFTC$_s$, but also chooses components adaptively with as much information as possible. Moreover, fewer monitored components speed up shift detecting and especially benefits small shifts, see [23]. Hence WDFTC$_a$ outperforms WDFTC$_s$ on LRS data and achieves practical improvement.

Table 12: ARL's for LRS data.

	τ	WDFTC$_s$	WDFTC$_a$
ARL$_0$	0	361.72	367.23
MPP	0.0254	156.48	138.74
	0.0508	43.18	38.87
	0.0762	20.94	19.86
	0.1016	13.65	12.75
Taper	0.0508	168.92	89.06
	0.1016	46.47	27.11
	0.1524	22.24	13.40
	0.2032	14.07	9.71
Flare	0.0508	269.26	153.87
	0.1016	91.75	65.27
	0.1524	42.30	27.81
	0.2032	24.45	14.08
Snake	0.0508	88.29	57.07
	0.1016	23.85	16.73
	0.1524	12.54	9.49
	0.2032	10.17	8.19

3.4 Conclusion

We develop a wavelet-based control chart with adaptive thresholds for high-dimensional profiles with general noise distributions and some correlations. The main idea of WDFTC$_a$ is to employ an adaptive thresholding method to select a small number of wavelet components when in-control and more components when out-of-control. CF thresholds are the adaptive thresholds implemented in WDFTC$_a$ and can be considered as the lower and upper quantiles associated with the probability of universal thresholds in the standard normal distribution. WDFTC$_a$ performs well for high-dimensional profiles and is especially sensitive to small shifts and/or local shifts. It can also detect shifts that are missed by existing wavelet-based control charts with static selection. Experimental results show WDFTC$_a$ outperforms WDFTC$_s$ for both Mallat's piecewise smooth function and LRS data. WDFTC$_a$ achieves significant improvement with LRS data and reveals its potential in real applications.

CHAPTER IV

DISCUSSION AND CONCLUSIONS

In Chapter 2, we studied the conservativeness of fully sequential procedures. There are three conservativeness sources: (a) the BI, which enables us to focus on making selections between a pair of systems instead of among all k systems; (b) the DTC, which bounds the probability of making an incorrect selection in a discrete stochastic process by that in a corresponding continuous stochastic process; and (c) the SC assumption, which sets us free of unknown mean parameters.

Both the BI and the SC assumption are critical sources that affect the efficiency of fully sequential procedures significantly. The other source, the DTC, affects the procedures' efficiency insignificantly. When the number of system is large, the impact of the SC assumption overwhelms. To develop an efficient procedure that works well for large k, we proposed two new procedures, $\mathcal{WK}$ and $\mathcal{WK}$++, which lessen the impact of the SC assumption and achieve more than 30% savings under the DC, compared to existing procedures. The savings can be even greater if the means of the systems are even more spaced out. Additionally, the same technique can help other researchers to improve the efficiency of some other fully sequential procedures, such as [39, 40].

In Chapter 3, we developed a wavelet-based distribution-free tabular CUSUM chart based on adaptive thresholding, WDFTC$_a$, for quickly detecting shifts in the mean of high-dimensional profiles with general noise distributions and some correlations. The idea of adaptive selection is to select a small number of wavelet components (e.g., mainly scaling coefficients) when in-control and more components (e.g., including some extreme outliers of detail coefficients) when out-of-control, so that when the process is in-control, monitored statistics have manageable variability and can be well approximated as a Brownian motion process; while when the process is out-of-control, monitored statistics change significantly and can quickly detect shifts in the mean profile.

On the other hand, it is not trivial to extend some existing wavelet-based distribution-free control charts [33] from static selection to adaptive selection. Since (a) with adaptive selection, the selected components change from profile to profile and the number of selected components is also affected by the correlations between profile components, both result in a large variance for monitoring statistics; and (b) the distribution of monitored statistics can be very complicated. WDFTC$_a$ was designed to overcome these difficulties. In order to maintain manageable covariances, WDFTC$_a$ computes monitored statistics from nonoverlapping batch means of the wavelet-transformed noise vectors and uses a matrix-regularization method to estimate their covariance matrix. WDFTC$_a$ makes adaptive selection with CF thresholds, which uses the associated marginal Cornish-Fisher expansions that have been suitably adjusted for between-component correlations.

Moreover, experimentation with several normal and nonnormal test processes reveals that WDFTC$_a$ outperforms existing nonadaptive profile-monitoring schemes. WDFTC$_a$ performs well on small shifts and/or local shifts and can detect shifts missed by its competitors. Additionally, WDFTC$_a$ achieves significant improvement in a lumber manufacturing case and shows its potential in practical applications.

APPENDIX A

PROOFS OF THEOREMS 1 AND 2

A.1 Proof of Theorem 1

We need the following lemmas for the proof of Theorem 1.

Lemma 4 *Define the standardized partial sum for system i as*

$$C_i(t,r) \equiv \frac{\sum_{j=1}^{\lfloor rt \rfloor} X_{ij} - rt\mu_i}{\sigma_i\sqrt{r}}, \qquad 0 \leq t \leq 1 \textit{ and } r \in \mathbb{Z}^+, \tag{22}$$

where $\lfloor \cdot \rfloor$ indicates truncation of any fractional part. The probability distribution of $C_i(t,r)$ over $D[0,1]$, the space of functions that are right-continuous and have left-hand limits, converges to that of a standard Brownian motion process, $\mathcal{W}(t)$, as r increases; i.e., $C_i(\cdot,r) \Rightarrow \mathcal{W}(\cdot)$ as $r \to \infty$, where $\Rightarrow$ denotes convergence in distribution. Further, we assume that for any $t \in [0,1]$, the family of random variables $\{C_i^2(t,r) : r = 1,2,\ldots\}$ is uniformly integrable.

Lemma 4 is basically Functional Central Limit Theorem [6]. It is straightforward that Lemma 4 holds under Assumption 1.

Lemma 5 *If Assumptions 1, 2 and 3 hold, then there exists $N \in \mathbb{Z}^+$ such that $\delta \leq \hat{\delta}_{1i}(r) < (c_i + \kappa)\delta$ with probability 1 (w.p.1), for any $r > N$ and some $\kappa \in \mathbb{R}^+$.*

Proof: Under Assumptions 1, 2 and 3, $|\bar{X}_1(r) - \bar{X}_i(r)| - A_{1i}(r) \to c_i\delta$ w.p.1 as $r \to \infty$ for $i \in \{2,\ldots,k\}$. This follows by the strong law of large numbers and continuous mapping theorem (CMT) [6]. Set $\epsilon = \kappa\delta$ for some $\kappa \in \mathbb{R}^+$. Then there exists some $N \in \mathbb{Z}^+$, $(c_i - \kappa)\delta < |\bar{X}_1(r) - \bar{X}_i(r)| - A_{1i}(r) < (c_i + \kappa)\delta$ and therefore $\delta \leq \hat{\delta}_{1i}(r) < (c_i + \kappa)\delta$ for $r > N$ w.p.1. □

Proof of Theorem 1: Let $r \in \{n_0, n_0 + 1, \ldots\}$,

$$T_{1i} \equiv T_{1i}(\delta) = \min\left\{ r : \left| \sum_{j=1}^{r} (X_{1j} - X_{ij}) \right| \geq R(r; \hat{\delta}_{1i}(r), h^2, S_{1i}^2(n_0)) \right\}$$

and $N_{1i}(r) = \left\lfloor \frac{2eh^2 S_{1i}^2(n_0)}{\hat{\delta}_{1i}^2(r)} \right\rfloor$.

First of all, we will show that $T_{1i} \to \infty$ as $\delta \to 0$ w.p.1, which is a fundamental requirement for the convergence of estimators. We first represent the output process $\{X_{1j} - X_{ij}\}$ as $\{Z_j + c_i\delta\}$, where $Z_j \overset{\text{IID}}{\sim} N(0, \sigma_{1i}^2)$, $i = 2, \ldots, k$ and $j = 1, 2, \ldots$. Consider a sample path going up for which $\limsup_{\delta \to 0} T_{1i} = T^* \leq \infty$. (When the path goes down, the proof is analogous.) There must exist $\delta^* > 0$ such that for all $\delta \leq \delta^*$,

$$\sum_{j=1}^{T^*} Z_j + T^* c_i \delta \geq \frac{h^2 S_{1i}^2(n_0)}{\hat{\delta}_{1i}(T^*)} - \frac{\hat{\delta}_{1i}(T^*)}{2e} T^*. \tag{23}$$

By Lemma 5, there must exist some $\kappa \in \mathbb{R}^+$ and $n < T^*$ that ensure $\delta \leq \hat{\delta}_{1i}(r) < (c_i + \kappa)\delta$ for any $r > n$ w.p.1. This implies that $\delta \leq \hat{\delta}_{1i}(T^*) < (c_i + \kappa)\delta$ and we obtain

$$\sum_{j=1}^{T^*} Z_j + T^* c_i \delta \geq \frac{h^2 S_{1i}^2(n_0)}{(c_i + \kappa)\delta} - \frac{(c_i + \kappa)\delta}{2e} T^*. \tag{24}$$

Equation (24) can be rewritten as follows:

$$\sum_{j=1}^{T^*} Z_j + T^* \delta (c_i + \frac{c_i + \kappa}{2e}) \geq \frac{h^2 S_{1i}^2(n_0)}{(c_i + \kappa)\delta}.$$

Note that when T^* is finite, $\sum_{j=1}^{T^*} Z_j$ is finite w.p.1, but $T^*\delta$ goes to zero while $\frac{h^2 S_{1i}^2(n_0)}{(c_i+\kappa)\delta}$ goes to infinity as $\delta \to 0$. Thus (23) occurs with probability zero and then $T_{1i} \to \infty$ w.p.1 as $\delta \to 0$. This also ensures that as $\delta \to 0$,

$$\hat{\delta}_{1i}(T_{1i})/\delta_{1i} \to 1 \quad w.p.1. \tag{25}$$

Due to (25) and CMT, it is also true that $N_{1i}(T_{1i}) \to \infty$ as $\delta \to 0$. Let ICS$_i$ denote the event that the best system gets eliminated by an inferior system i, $i = 2, \ldots, k$. Then

$$\begin{aligned}
\Pr\{\text{ICS}_i\} &= \Pr\left\{ \sum_{j=1}^{T_{1i}} (X_{1j} - X_{ij}) \leq -\max\left\{0, \frac{h^2 S_{1i}^2(n_0)}{\hat{\delta}_{1i}(T_{1i})} - \frac{\hat{\delta}_{1i}(T_{1i})}{2e} T_{1i} \right\} \right\} \\
&= \Pr\left\{ \frac{\sum_{j=1}^{T_{1i}} (X_{1j} - X_{ij}) - \delta_{1i} T_{1i}}{\sigma_{1i}\sqrt{N_{1i}(T_{1i}) + 1}} + \frac{\delta_{1i} T_{1i}}{\sigma_{1i}\sqrt{N_{1i}(T_{1i}) + 1}} \leq \right. \\
&\qquad \left. -\max\left\{0, \frac{h^2 S_{1i}^2(n_0)}{\hat{\delta}_{1i}(T_{1i})\sigma_{1i}\sqrt{N_{1i}(T_{1i}) + 1}} - \frac{\hat{\delta}_{1i}(T_{1i})}{2e\sigma_{1i}\sqrt{N_{1i}(T_{1i}) + 1}} T_{1i} \right\} \right\}
\end{aligned} \tag{26}$$

For $0 \leq t \leq 1$, let

$$C_{1i}(t, \delta) = \frac{\sum_{j=1}^{\lfloor (N_{1i}(T_{1i})+1)t \rfloor} (X_{1j} - X_{ij}) - (N_{1i}(T_{1i}) + 1)\delta_{1i} t}{\sigma_{1i}\sqrt{N_{1i}(T_{1i}) + 1}}$$

Further, define

$$\widehat{T}_{1i} = \min\left\{ t \in \left\{ \frac{n_0}{N_{1i}(T_{1i})+1}, \frac{n_0+1}{N_{1i}(T_{1i})+1}, \ldots, 1 \right\} : \left| C_{1i}(t,\delta) + \frac{(N_{1i}(T_{1i})+1)\delta_{1i}t}{\sigma_{1i}\sqrt{N_{1i}(T_{1i})+1}} \right| \right.$$
$$\left. \geq \frac{h^2 S_{1i}^2(n_0)}{\hat{\delta}_{1i}(T_{1i})\sigma_{1i}\sqrt{N_{1i}(T_{1i})+1}} - \frac{\hat{\delta}_{1i}(T_{1i})(N_{1i}(T_{1i})+1)}{2e\sigma_{1i}\sqrt{N_{1i}(T_{1i})+1}}t \right\}.$$

Clearly, $\widehat{T}_{1i} = T_{1i}/(N_{1i}(T_{1i})+1)$. Also define the stopping time of the corresponding continuous-time process as

$$\tilde{T}_{1i} = \min\left\{ t \geq \frac{n_0}{N_{1i}(T_{1i})+1} : \left| C_{1i}(t,\delta) + \frac{(N_{1i}(T_{1i})+1)\delta_{1i}t}{\sigma_{1i}\sqrt{N_{1i}(T_{1i})+1}} \right| \right.$$
$$\left. \geq \frac{h^2 S_{1i}^2(n_0)}{\hat{\delta}_{1i}(T_{1i})\sigma_{1i}\sqrt{N_{1i}(T_{1i})+1}} - \frac{\hat{\delta}_{1i}(T_{1i})(N_{1i}(T_{1i})+1)}{2e\sigma_{1i}\sqrt{N_{1i}(T_{1i})+1}}t \right\}.$$

Notice that for fixed δ, $C_{1i}(\widehat{T}_{1i},\delta)$ corresponds to the right-hand limit of a point of discontinuity of $C_{1i}(\cdot,\delta)$. We can show that $\widehat{T}_{1i} \to \tilde{T}_{1i}$ w.p.1 as $\delta \to 0$, making use of the fact that $1/(N_{1i}(T_{1i})+1) \to 0$ w.p.1. Thus, in the limit, we can focus on $C_{1i}(\tilde{T}_{1i},\delta)$.

Now condition on $S_{1i}^2(n_0)$. Then Lemma 4, (25) and CMT imply that

$$C_{1i}(t,\delta) + \frac{(N_{1i}(T_{1i})+1)\delta_{1i}t}{\sigma_{1i}\sqrt{N_{1i}(T_{1i})+1}} \Rightarrow \mathcal{W}(t,\Delta)$$

as $\delta \to 0$, where

$$\Delta = \lim_{\delta\to 0} \frac{(N_{1i}(T_{1i})+1)\delta_{1i}}{\sigma_{1i}\sqrt{N_{1i}(T_{1i})+1}} = \frac{\sqrt{2eh^2 S_{1i}^2(n_0)}}{\sigma_{1i}}.$$

Still conditional on $S_{1i}^2(n_0)$, let

$$\mathcal{A}(\delta) = \frac{h^2 S_{1i}^2(n_0)}{\hat{\delta}_{1i}(T_{1i})\sigma_{1i}\sqrt{N_{1i}(T_{1i})+1}} \xrightarrow{\delta\to 0} \frac{\sqrt{2eh^2\, S_{1i}^2(n_0)}}{2e\sigma_{1i}} \equiv \mathcal{A}$$

$$\mathcal{B}(\delta) = \frac{\hat{\delta}_{1i}(T_{1i})(N_{1i}(T_{1i})+1)}{2e\sigma_{1i}\sqrt{N_{1i}(T_{1i})+1}} \xrightarrow{\delta\to 0} \frac{\sqrt{2eh^2\, S_{1i}^2(n_0)}}{2e\sigma_{1i}} \equiv \mathcal{B}.$$

Notice that the stopping time $\tilde{T}_{1i}$ is the first time t at which the event

$$\left\{ \left| C_{1i}(t,\delta) + \frac{(N_{1i}(T_{1i})+1)\delta_{1i}t}{\sigma_{1i}\sqrt{N_{1i}(T_{1i})+1}} \right| - \mathcal{A}(\delta) + \mathcal{B}(\delta)t \geq 0 \right\}$$

occurs. Define the mapping $s_\delta : D[0,1] \to \mathbb{R}$ such that $s_\delta(Y) = Y(T_{Y,\delta})$ where

$$T_{Y,\delta} = \inf\{t : |Y(t)| - \mathcal{A}(\delta) + \mathcal{B}(\delta)t \geq 0\}$$

for every $Y \in D[0,1]$ and $\delta > 0$. Similarly, define $s(Y) = Y(T_Y)$ where

$$T_Y = \inf\left\{t : |Y(t)| - \mathcal{A} + \mathcal{B}t \geq 0\right\}$$

for every $Y \in D[0,1]$ and $\delta > 0$. Notice that

$$s_\delta\left(C_{1i}(t,\delta) + \frac{(N_{1i}(T_{1i})+1)\delta_{1i}t}{\sigma_{1i}\sqrt{N_{1i}(T_{1i})+1}}\right) = C_{1i}(\tilde{T}_{1i},\delta) + \frac{(N_{1i}(T_{1i})+1)\delta_{1i}\tilde{T}_{1i}}{\sigma_{1i}\sqrt{N_{1i}(T_{1i})+1}}$$

$$s\left(\mathcal{W}(\cdot,\Delta)\right) = \mathcal{W}(T_{\mathcal{W}(\cdot,\Delta)},\Delta).$$

We need to show that

$$s_\delta\left(\mathcal{G}_{1i}(\cdot,\delta)\right) \Rightarrow s\left(\mathcal{W}(\cdot,\Delta)\right) \tag{27}$$

as $\delta \to 0$, where

$$\mathcal{G}_{1i}(t,\delta) \equiv C_{1i}(t,\delta) + \frac{(N_{1i}(T_{1i})+1)\delta_{1i}t}{\sigma_{1i}\sqrt{N_{1i}(T_{1i})+1}} \tag{28}$$

for $t \in [0,1]$ and $\delta > 0$. This follows Proposition 2 from [31] which is established on the extended CMT [6].

Unconditioning on S_{1i}^2, (27) gives

$$\begin{aligned}
&\limsup_{\delta\to 0}\Pr\{\mathrm{ICS}_i\}\\
&= \mathrm{E}\left[\Pr\left\{\mathcal{W}(t,\Delta) \text{ exits continuation region through the lower boundary}|S_{1i}^2(n_0)\right\}\right]\\
&= \mathrm{E}\left[\sum_{v=1}^{e}(-1)^{v+1}\left(1-\frac{1}{2}\mathcal{I}(v=e)\right)\exp\left\{-\frac{h^2S_{1i}^2(n_0)}{e\sigma_{1i}^2}(2e-v)v\right\}\middle| S_{1i}^2(n_0)\right] && (29)\\
&= \mathrm{E}\left[\sum_{v=1}^{e}(-1)^{v+1}\left(1-\frac{1}{2}\mathcal{I}(v=e)\right)\exp\left\{-\frac{\eta(n_0-1)S_{1i}^2(n_0)}{e\sigma_{1i}^2}(2e-v)v\right\}\middle| S_{1i}^2(n_0)\right] && (30)\\
&= \sum_{v=1}^{e}(-1)^{v+1}\left(1-\frac{1}{2}\mathcal{I}(v=e)\right)\mathrm{E}\left[\exp\left\{-\frac{\eta\chi_{n_0-1}^2}{e}(2e-v)v\right\}\right]\\
&= \sum_{v=1}^{e}(-1)^{v+1}\left(1-\frac{1}{2}\mathcal{I}(v=e)\right)\left(1+\frac{2\eta(2e-v)v}{e}\right)^{-(n_0-1)/2} && (31)\\
&= \frac{\alpha}{k-1},
\end{aligned}$$

where (30) holds due to Lemma 2, χ_f^2 is a chi-squared random variable with degrees of freedom f, and (31) is from the moment generating function of χ_f^2. Finally the probability of incorrect selection is less than the nominal level as below.

$$\limsup_{\delta\to 0} \Pr\{\mathrm{ICS}\} \le \limsup_{\delta\to 0} \sum_{i=2}^{k} \Pr\{\mathrm{ICS}_i\} \le \alpha.$$

□

A.2 Proof of Theorem 2

For the proof of $\mathcal{WK}$++ we need the following lemma:

Lemma 6 *[6]*

If $Y_n \Rightarrow Y$ *and* $Z_n \xrightarrow{P} a$ *where* a *is a constant vector, then* $(Y_n,\, Z_n) \Rightarrow (Y,\, a)$*.*

Proof of Theorem 2: First set T_{1i} and $N_{1i}(r)$ in the same way as the proof of Theorem 1, except replacing $S_{1i}^2(n_0)$ with $S_{1i}^2(r)$. Sample variance $S_{1i}^2(n_0)$ in (23) and (24) is also replaced with $S_{1i}^2(T^*)$. Then using a similar argument we can show $T_{1i} \to \infty$ as $\delta \to 0$ w.p.1 to guarantee the convergence of estimators. This ensures that as $\delta \to 0$,

$$\hat{\delta}_{1i}(T_{1i})/\delta_{1i} \to 1 \quad \text{and} \quad S_{1i}^2(T_{1i}) \to \sigma_{1i}^2, \text{ w.p.1.} \tag{32}$$

Also, $N_{1i}(r) \to \infty$ as $\delta \to 0$ w.p.1 due to (32) and CMT. The definition of $C_{1i}(t,\delta)$ is the same as in the proof of Theorem 1. Redefine $\hat{T}_{1i}$ and $\tilde{T}_{1i}$ as follows:

$$\begin{aligned}\widehat{T}_{1i} = \min\Bigg\{ t \in \left\{\frac{n_0}{N_{1i}(T_{1i})+1}, \frac{n_0+1}{N_{1i}(T_{1i})+1}, \ldots, 1\right\} : \left| C_{1i}(t,\delta) + \frac{(N_{1i}(T_{1i})+1)\delta_{1i}t}{\sigma_{1i}\sqrt{N_{1i}(T_{1i})+1}} \right| \\ \ge \frac{h^2 S_{1i}^2(T_{1i})}{\hat{\delta}_{1i}(T_{1i})\sigma_{1i}\sqrt{N_{1i}(T_{1i})+1}} - \frac{\hat{\delta}_{1i}(T_{1i})(N_{1i}(T_{1i})+1)}{2e\sigma_{1i}\sqrt{N_{1i}(T_{1i})+1}} t \Bigg\}.\end{aligned}$$

$$\begin{aligned}\tilde{T}_{1i} = \min\Bigg\{ t \ge \frac{n_0}{N_{1i}(T_{1i})+1} : \left| C_{1i}(t,\delta) + \frac{(N_{1i}(T_{1i})+1)\delta_{1i}t}{\sigma_{1i}\sqrt{N_{1i}(T_{1i})+1}} \right| \\ \ge \frac{h^2 S_{1i}^2(T_{1i})}{\hat{\delta}_{1i}(T_{1i})\sigma_{1i}\sqrt{N_{1i}(T_{1i})+1}} - \frac{\hat{\delta}_{1i}(T_{1i})(N_{1i}(T_{1i})+1)}{2e\sigma_{1i}\sqrt{N_{1i}(T_{1i})+1}} t \Bigg\}.\end{aligned}$$

We can show that $\widehat{T}_{1i} \to \tilde{T}_{1i}$ w.p.1 as $\delta \to 0$ making use of the fact that $1/(N_{1i}(T_{1i})+1) \to 0$ w.p.1. Thus, in the limit, we can focus on $C_{1i}(\tilde{T}_{1i},\delta)$.

Since strong consistency implies convergence in probability, (32) and Lemma 6 can be applied to $\left(C_{1i}(T_{1i},\delta), \begin{bmatrix} \hat{\delta}_{1i}(T_{1i})/\delta_{1i} \\ S_{1i}^2(T_{1i}) \end{bmatrix} \right)$. By the random-change-of-time theorem [6], the followings hold.

$$C_{1i}(t,\delta) + \frac{(N_{1i}(T_{1i})+1)\delta_{1i}t}{\sigma_{1i}\sqrt{N_{1i}(T_{1i})+1}} \Rightarrow \mathcal{W}(t,\Delta)$$

as $\delta \to 0$, where

$$\Delta = \lim_{\delta\to 0} \frac{(N_{1i}(T_{1i})+1)\delta_{1i}}{\sigma_{1i}\sqrt{N_{1i}(T_{1i})+1}} = \sqrt{2eh^2}.$$

$$\mathcal{A}(\delta) = \frac{h^2 S_{1i}^2(T_{1i})}{\hat{\delta}_{1i}(T_{1i})\sigma_{1i}\sqrt{N_{1i}(T_{1i})+1}} \overset{\delta\to 0}{\longrightarrow} \sqrt{\frac{h^2}{2e}} \equiv \mathcal{A}$$

$$\mathcal{B}(\delta) = \frac{\hat{\delta}_{1i}(T_{1i})(N_{1i}(T_{1i})+1)}{2e\sigma_{1i}\sqrt{N_{1i}(T_{1i})+1}} \overset{\delta\to 0}{\longrightarrow} \sqrt{\frac{h^2}{2e}} \equiv \mathcal{B}$$

The definitions of mappings s_δ, s, $T_{Y,\delta}$, and T_Y follow those in the proof of Theorem 1 . Then (27) holds again. Therefore,

$$\begin{aligned}\limsup_{\delta\to 0} \Pr\{\mathrm{ICS_i}\} &= \Pr\{\mathcal{W}(t,\Delta) \text{ exits continuation region through the lower boundary}\} \\ &= \sum_{v=1}^{e} (-1)^{v+1}\left(1-\frac{1}{2}\mathcal{I}(v=e)\right)\exp\left\{-2\frac{h^2}{2e}(2e-v)v\right\} \text{ (by Lemma 2)} \\ &= \sum_{v=1}^{e} (-1)^{v+1}\left(1-\frac{1}{2}\mathcal{I}(v=e)\right)\exp\left\{-\frac{\eta}{e}(2e-v)v\right\} = \alpha/(k-1) \qquad (33)\end{aligned}$$

where the equality follows from the way we choose η. Finally,

$$\limsup_{\delta\to 0} \Pr\{\mathrm{ICS}\} \le \limsup_{\delta\to 0} \sum_{i=2}^{k} \Pr\{\mathrm{ICS}_i\} \le \alpha.$$

□

APPENDIX B

PROOFS OF $0 < T \leq 1$ IN SECTION 3.2.2.3

We prove $0 < t \leq 1$ for Equation (16) and thus $\gamma = 1/\sqrt{t} \geq 1$. First set the difference between two matrix averages to be $\nabla = \bar{\mathbf{\Lambda}} - \bar{\boldsymbol{S}}$. Next we show

$$\bar{\mathbf{\Lambda}} < n\bar{\boldsymbol{S}}. \tag{34}$$

Note that $1 \leq i, j \leq n$, and

$$\begin{aligned} n^2\bar{\mathbf{\Lambda}} &= \sum_i \sum_j \mathbf{\Lambda}_{ij} = \sum_i \mathbf{\Lambda}_{ii} + \sum_i \sum_{j \neq i} \mathbf{\Lambda}_{ij} \\ &\leq \sum_i \mathbf{\Lambda}_{ii} + \sum_i \sum_{j \neq i} \sqrt{\mathbf{\Lambda}_{ii}\mathbf{\Lambda}_{jj}} \\ &\leq \sum_i \mathbf{\Lambda}_{ii} + \sum_i \sum_{j \neq i} \frac{\mathbf{\Lambda}_{ii} + \mathbf{\Lambda}_{jj}}{2} \\ &= n \sum_i \mathbf{\Lambda}_{ii} = n \sum_i \boldsymbol{S}_{ii} = n^3 \bar{\boldsymbol{S}}, \end{aligned} \tag{35}$$

where the first inequality holds due to the property of correlation; the second inequality holds due to the inequality of arithmetic and geometric means and $\mathbf{\Lambda}_{ii} \geq 0$, $1 \leq i \leq n$. On the other side, the equality holds only when all elements in $\mathbf{\Lambda}$ are the same, i.e., $\mathbf{\Lambda}_{i,j} = c$, $1 \leq i, j \leq n$ and c is a fixed constant. Such $\mathbf{\Lambda}$ implies that all variables are perfectly positive correlated, which is impossible in reality. Therefore strict inequality (34) holds. Since the denominator of t is always positive, we first check the sign of the numerator.

$$\begin{aligned} &\sum_i \sum_j (\mathbf{\Lambda}_{ij} - \bar{\mathbf{\Lambda}})(\boldsymbol{S}_{ij} - \bar{\boldsymbol{S}}) \\ =\ &\sum_i \sum_{j \neq i} (\mathbf{\Lambda}_{ij} - \bar{\mathbf{\Lambda}})(0 - \bar{\boldsymbol{S}}) + \sum_i (\mathbf{\Lambda}_{ii} - \bar{\mathbf{\Lambda}})(\boldsymbol{S}_{ii} - \bar{\boldsymbol{S}}) \\ =\ &\sum_i \sum_{j \neq i} (\mathbf{\Lambda}_{ij} - \bar{\mathbf{\Lambda}})(-\bar{\boldsymbol{S}}) + \sum_i (\boldsymbol{S}_{ii} - \bar{\boldsymbol{S}})^2 - \nabla \sum_i (\boldsymbol{S}_{ii} - \bar{\boldsymbol{S}}). \end{aligned} \tag{36}$$

Equation (36) holds due to the definition of $\boldsymbol{S}$, $\mathbf{\Lambda}_{ii} = \boldsymbol{S}_{ii}$ for $1 \leq i \leq n$, and $\bar{\mathbf{\Lambda}} = \bar{\boldsymbol{S}} + \nabla$. Further,

$$\begin{aligned} \sum_i \sum_{j \neq i} (\mathbf{\Lambda}_{ij} - \bar{\mathbf{\Lambda}}) &= \sum_i \sum_{j \neq i} \mathbf{\Lambda}_{ij} - n(n-1)\bar{\mathbf{\Lambda}} = \sum_i \sum_j \mathbf{\Lambda}_{ij} - \sum_i \mathbf{\Lambda}_{ii} - n(n-1)\bar{\mathbf{\Lambda}} \\ &= n^2\bar{\mathbf{\Lambda}} - n^2\bar{\boldsymbol{S}} - n(n-1)\bar{\mathbf{\Lambda}} = n\bar{\mathbf{\Lambda}} - n^2\bar{\boldsymbol{S}} = n(\bar{\boldsymbol{S}} + \nabla) - n^2\bar{\boldsymbol{S}} \qquad (37) \\ &= n(1-n)\bar{\boldsymbol{S}} + n\nabla \end{aligned}$$

$$\nabla\sum_i(\boldsymbol{S}_{ii}-\bar{\boldsymbol{S}}) \;=\; \nabla(\sum_i\boldsymbol{S}_{ii}-n\bar{\boldsymbol{S}})=\nabla(n^2\bar{\boldsymbol{S}}-n\bar{\boldsymbol{S}}) \tag{38}$$

where (37) holds due to $\sum_i\sum_j\boldsymbol{\Lambda}_{ij}=n^2\bar{\boldsymbol{\Lambda}}$ and (38) holds due to (35). Then

$$\begin{aligned}
(36) &= n(n-1)\bar{\boldsymbol{S}}^2-n\nabla\bar{\boldsymbol{S}}+\sum_i(\boldsymbol{S}_{ii}-\bar{\boldsymbol{S}})^2-n^2\nabla\bar{\boldsymbol{S}}+n\nabla\bar{\boldsymbol{S}}\\
&= n^2\bar{\boldsymbol{S}}^2-n\bar{\boldsymbol{S}}^2+\sum_i\boldsymbol{S}_{ii}^2-2\sum_i\boldsymbol{S}_{ii}\bar{\boldsymbol{S}}+n\bar{\boldsymbol{S}}^2-n^2\bar{\boldsymbol{S}}\nabla\\
&= \sum_i\boldsymbol{S}_{ii}^2-n^2\bar{\boldsymbol{S}}^2-n^2\bar{\boldsymbol{S}}\nabla \qquad (39)\\
&= \sum_i\boldsymbol{S}_{ii}^2-n^2\bar{\boldsymbol{S}}\bar{\boldsymbol{\Lambda}}\\
&> \sum_i\boldsymbol{S}_{ii}^2-n^2\bar{\boldsymbol{S}}n\bar{\boldsymbol{S}}=\sum_i\boldsymbol{S}_{ii}^2-\frac{(\sum_i\boldsymbol{S}_{ii})^2}{n} \qquad (40)\\
&\geq \sum_i\boldsymbol{S}_{ii}^2-\frac{n\sum_i\boldsymbol{S}_{ii}^2}{n}=0 \qquad (41)
\end{aligned}$$

Inequality (40) holds due to Inequality (34). The last inequality (41) employs Jensen's inequality. Therefore $t>0$ holds for any $\boldsymbol{\Lambda}$. According to Cauchy–Schwarz inequality,

$$\left[\sum_i\sum_j(\boldsymbol{\Lambda}_{ij}-\bar{\boldsymbol{\Lambda}})(\boldsymbol{S}_{ij}-\bar{\boldsymbol{S}})\right]^2\leq\left(\sum_i\sum_j(\boldsymbol{\Lambda}_{ij}-\bar{\boldsymbol{\Lambda}})^2\right)\left(\sum_i\sum_j(\boldsymbol{S}_{ij}-\bar{\boldsymbol{S}})^2\right),$$

thus it is guaranteed that $t\leq 1$ and thus $\gamma\geq 1$. By definition of γ in (17), $1\leq\gamma\leq 1.5$.

Additionally, when negative correlations exist in $\boldsymbol{\Lambda}$, ∇, the difference between $\bar{\boldsymbol{\Lambda}}$ and $\bar{\boldsymbol{S}}$ gets smaller and the numerator of t gets larger due to (39), compared to the covariance matrix with the same level of correlation but all positively correlated, i.e., a matrix with the absolutely value of $\boldsymbol{\Lambda}$.

REFERENCES

[1] Averkamp, R. and Houdré, C., "Wavelet thresholding for non necessarily gaussian noise: Idealism," *The Annals of Statistics*, vol. 31, no. 1, pp. 110–151, 2003.

[2] Averkamp, R. and Houdré, C., "Wavelet thresholding for non necessarily gaussian noise: Functionality," *The Annals of Statistics*, vol. 33, no. 5, pp. 2164–2193, 2005.

[3] Bechhofer, R. E., "A single-sample multiple decision procedure for ranking means of normal populations with known variances," *Annals of Mathematical Statistics*, vol. 25, pp. 16–39, 1954.

[4] Bekki, J. M., Fowler, J. W., Mackulak, G. T., and Nelson, B. L., "Indirect cycle time quantile estimation using the Cornish–Fisher expansion," *IIE Transactions*, vol. 42, no. 1, pp. 31–44, 2009.

[5] Bickel, P. J. and Levina, E., "Covariance regularization by thresholding," *The Annals of Statistics*, vol. 36, no. 6, pp. 2577–2604, 2008.

[6] Billingsley, P., *Convergence of Probability Measures.* John Wiley & Sons, Inc., first ed., 1968.

[7] Cai, T. T. and Yuan, M., "Adaptive covariance matrix estimation through block thresholding," *Working Paper*, 2012.

[8] Cario, M. C. and Nelson, B. L., "Autoregressive to anything: Time-series input processes for simulation," *Operations Research Letters*, vol. 19, pp. 51–58, 1996.

[9] Chen, E. J., "Using ordinal optimization approach to improve efficiency of selection procedures," *Discrete Event Dynamic Systems: Theory and Applications*, vol. 14, pp. 153–170, 2004.

[10] Chen, E. J. and Kelton, W. D., "An enhanced two-stage selection procedure," in *Proceedings of the 2000 Winter Simulation Conference* (Joines, J. A., Barton, R. R., Kang, K., and Fishwick, P. A., eds.), pp. 727–735, IEEE, 2000.

[11] Chen, E. J. and Kelton, W. D., "Sequential selection procedures: Using sample means to improve efficiency," *European Journal of Operational Research*, vol. 166, pp. 133–153, 2005.

[12] Chen, S. and Nembhard, H. B., "A high-dimensional control chart for profile monitoring," *Quality and Reliability Engineering International*, vol. 27, no. 4, pp. 451–464, 2011.

[13] Chicken, E., Pignatiello Jr., J. J., and Simpson, J. R., "Statistical process monitoring of nonlinear profiles using wavelets," *Journal of Quality Technology*, vol. 41, no. 2, pp. 198–212, 2009.

[14] Donoho, D. L. and Johnstone, I. M., "Ideal spatial adaptation by wavelet shrinkage," *Biometrika*, vol. 81, no. 3, pp. 425–455, 1994.

[15] Fabian, V., "Note on anderson's sequential procedures with triangular boundary," *Annals of Statistics*, vol. 2, pp. 170–176, 1974.

[16] Fan, J., "Test of significance based on wavelet thresholding and Neyman's truncation," *Journal of American Statistical Association*, vol. 91, no. 434, pp. 674–688, 1996.

[17] Ganesan, R., Das, T. K., Sikder, A. K., and Kumar, A., "Wavelet based indentification of delamination emission signal," *IEEE Transactions on Semiconductor Manufacturing*, vol. 16, no. 4, pp. 677–685, 2003.

[18] Gao, H.-Y., "Wavelet shrinkage estimates for heteroscedastic regression models," tech. rep., Seattle: MathSoft, Inc., 1997.

[19] Gardner, M. M., Lu, J.-C., Gyurcsik, R. S., Wortman, J. J., Hornung, B. E., Heinisch, H. H., Rying, E. A., Rao, S., Davis, J. C., and Mozumder, P. K., "Equipment fault detection using spatial signatures," *IEEE TRANSACTIONS ON COMPONENTS, PACKAGING, AND MANUFACTURING TECHNOLOGY, PART C*, vol. 20, no. 4, pp. 295–304, 1997.

[20] Guo, H., Paynabar, K., and Jin, J., "Multiscale monitoring of autocorrelated processes using wavelets analysis," *IIE Transactions*, vol. 44, no. 4, pp. 312–326, 2012.

[21] Gupta, S., Montgomery, D. C., and Woodall, W. H., "Performance evaluation of two methods for online monitoring of linear calibration profiles," *International Journal of Production Research*, vol. 44, no. 10, pp. 1927–1942, 2006.

[22] Jennison, C., Johnstone, I. M., and Turnbull, B. W., "Asymptotically optimal procedures for sequential adaptive selection of the best of several normal means," in *Statistical Decision Theory and Related Topics, iii* (Gupta, S. S. and Berger, J. O., eds.), vol. 2, pp. 55–86, Academic Press, 1982.

[23] Jeong, M. K., Lu, J. C., and Wang, N., "Wavelet-based SPC procedure for complicated functional data," *International Journal of Production Research*, vol. 44, no. 4, pp. 729–744, 2006.

[24] Jin, J. and Shi, J., "Feature-preserving data compression of stamping tonnage information using wavelets," *Technometrics*, vol. 41, no. 4, pp. 327–339, 1999.

[25] Jin, J. and Shi, J., "Automatic feature extraction of waveform signals for in-process diagnostic performance improvement," *Journal of Intelligent Manufacturing*, vol. 12, no. 3, pp. 257–268, 2001.

[26] Johnstone, I. M., "Wavelet shrinkage for correlated data and inverse problems adaptivity results," *Statistica Sinica*, vol. 9, no. 1, pp. 51–83, 1999.

[27] Kang, L. and Albin, S. L., "On-line monitoring when the process yields a linear profile," *Journal of Quality Technology*, vol. 32, pp. 418–426, 2000.

[28] Kim, S.-H. and Nelson, B. L., "A fully sequential procedure for indifference-zone selection in simulation," *ACM Transactions on Modeling and Computer Simulation*, vol. 11, pp. 251–273, 2001.

[29] Kim, S.-H. and Nelson, B. L., "On the asymptotic validity of fully sequential selection procedures for steady-state simulation," *Operations Research*, vol. 54, pp. 475–488, 2006.

[30] Kim, S.-H. and Nelson, B. L., "Selecting the best system," in *Handbooks in Operations Research and Management Science: Simulation* (Henderson, S. G. and Nelson, B. L., eds.), pp. 501–534, Elsevier, 2006.

[31] Kim, S.-H., Nelson, B. L., and Wilson, J. R., "Some almost-sure convergence properties useful in sequential analysis," *Sequential Analysis*, vol. 24, pp. 411–419, 2005.

[32] Lada, E. K., Lu, J. C., and Wilson, J. R., "A wavelet based procedure for process fault detection," *IEEE Transactions on Semiconductor Manufacturing*, vol. 15, pp. 79–90, 2002.

[33] Lee, J., Hur, Y., Kim, S.-H., and Wilson, J. R., "Monitoring nonlinear profiles using a wavelet-based distribution-free CUSUM chart," *International Journal of Production Research*, 2012.

[34] Mallat, S. G., "Multifrequency channel decompositions of images and wavelet models," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 37, no. 12, pp. 2091–2110, 1989.

[35] Malone, G. J., Kim, S.-H., Goldsman, D., and Batur, D., "Performance of variance updating procedures on various data," in *Proceedings of the 2005 Winter Simulation Conference* (Kuhl, M. E., Steiger, N. M., Armstrong, F. B., and Joines, J. A., eds.), pp. 825–832, IEEE, 2005.

[36] Montgomery, D. C., *Statistical Quality Control.* Wiley, 7 ed., 2012.

[37] Paulson, E., "A sequential procedure for selecting the population with the largest mean from k normal populations," *Annals of Mathematical Statistics*, vol. 35, pp. 174–180, 1964.

[38] Perng, S. K., "A comparison of the asymptotic expected sample sizes of two sequential procedures for ranking problem," *Annals of Mathematical Statistics*, vol. 40, pp. 2198–2202, 1969.

[39] Pichitlamken, J. and Nelson, B. L., "A combined procedure for optimization via simulation," *ACM Transactions on Modeling and Computer Simulation*, vol. 13, pp. 155–179, 2003.

[40] Pichitlamken, J., Nelson, B. L., and Hong, L. J., "A sequential procedure for neighborhood selection-of-the-best in optimization via simulation," *European Journal of Operational Research*, vol. 173, pp. 283–298, 2006.

[41] Ramsay, J. O. and Silverman, B. W., *Functional Data Analysis.* New York, NY: Springer, 1997.

[42] Rinott, Y., "On two-stage selection procedures and related probability inequalities," *Communications in Statistics*, vol. A7, pp. 799–811, 1978.

[43] Song, W.-M. T. and Schmeiser, B., "Omitting meaningless digits in point estimates: The probability guarantee of leading-digit rules," *Operations Research*, vol. 57, pp. 109–117, 2009.

[44] Staudhammer, C., Kozak, R. A., and Maness, T. C., "Spc methods for detecting simple sawing defects using real-time laser range sensor data," *Wood and Fiber Science*, vol. 38, no. 4, pp. 696–716, 2006.

[45] Staudhammer, C., Maness, T. C., and Kozak, R. A., "Profile charts for monitoring lumber manufacturing using laser range sensor data," *Journal of Quality Technology*, vol. 39, no. 3, pp. 224–240, 2007.

[46] Staudhammer, C. L., LeMay, V. M., Maness, T. C., and Kozak, R. A., "Mixed-model development of real-time statistical process control data in wood products manufacturing," *Forestry Biometry, Modelling and Information Sciences*, vol. 1, pp. 19–35, 2005.

[47] von Sachs, R. and MacGibbon, B., "Non-parametric curve estimation by wavelet thresholding with locally stationary errors," *Scandinavian Journal of Statistics*, vol. 27, no. 3, pp. 475–499, 2000.

[48] Wang, H. and Kim, S.-H., "Reducing the conservativeness of fully sequential indifference-zone procedures," *IEEE Transactions on Automatic Control*, vol. 58, no. 6, pp. 1613–1619, 2013.

[49] Wang, H., Kim, S.-H., Huo, X., Hur, Y., and Wilson, J. R., "Monitoring nonlinear profiles adaptively with a wavelet-based distribution-free CUSUM chart," *International Journal of Production Research*, forthcoming.

[50] Williams, J. D., Woodall, W. H., and Birch, J. B., "Statistical monitoring of nonlinear product and process quality profiles," *Quality and Reliability Engineering International*, vol. 23, pp. 925–941, 2007.

[51] Woodall, W. H., "Current research on profile monitoring," *Produção*, vol. 17, pp. 420–425, Dec. 2007.

[52] Zou, C., Tsung, F., and Wang, Z., "Monitoring general linear profiles using multivariate exponentially weighted moving average schemes," *Technometrics*, vol. 49, no. 4, pp. 395–408, 2007.