

ON STATISTICAL SURVEILLANCE OF TWO-SAMPLED HIGH-DIMENSIONAL DATA

BY
FUHAD AHMED

A Thesis Presented to the
DEANSHIP OF GRADUATE STUDIES

KING FAHD UNIVERSITY OF PETROLEUM & MINERALS

DHAHRAN, SAUDI ARABIA

In Partial Fulfillment of the
Requirements for the Degree of

MASTER OF SCIENCE

In

APPLIED STATISTICS

DECEMBER 2023

KING FAHD UNIVERSITY OF PETROLEUM & MINERALS
DHAHRAN- 31261, SAUDI ARABIA
DEANSHIP OF GRADUATE STUDIES

This thesis, written by **Fuhad Ahmed** under the direction of his thesis advisor and approved by his thesis committee, has been presented and accepted by the Dean of Graduate Studies, in partial fulfillment of the requirements for the degree of **MASTER OF SCIENCE IN APPLIED STATISTICS**.



Prof. Monther Rashed Alfuraidan
Department Chairman

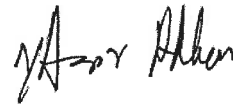


Prof. Suliman Saleh Al-Homidan
Dean of Graduate Studies

Date



Prof. Muhammad Riaz
(Advisor)



Dr. Nasir Abbas
(Member)



Dr. Mohammad Hafidz Omar
(Member)



© Fuhad Ahmed

2023

To my family

ACKNOWLEDGMENTS

At first, I express my gratefulness to almighty Allah for everything. Then I would like to thank my family for their enormous support throughout my entire journey.

I should not forget to thank King Fahd University of Petroleum and Minerals (KFUPM) for offering me the prestigious scholarship. I deeply acknowledge my beloved department of Mathematics at KFUPM for the excellent support from first to last. Special thanks to our respected Chairman Prof. Monther Rashed Alfuraidean for his continuous support.

As my thesis advisor, Prof. Muhammad Riaz deserves a great acknowledgement for guiding me since the first semester of my KFUPM life. He knows how to get the best out of a student. I would also like to acknowledge Dr. Nasir Abbas for helping me to run the big and lengthy simulations smoothly. He is perhaps the best teacher in my life; I have learned a lot from such an esteemed academician. I acknowledge Dr. Mohammad Hafidz Omar for guiding me regarding the direction of my thesis in the beginning. Working with him is a pleasant experience. I must not forget to express my gratitude to Dr. Tahir Mahmood (Lecturer, School of Computing, University of West of Scotland, UK) for his relentless efforts in my thesis work. I want to thank Dr. Jimoh Olawale Ajadi, a very good human being, for giving me nice suggestions. Thanks to Dr. Emmanuel Afuecheta for advising me regarding future possibilities. I thank Dr. Taleb Alkurdi for teaching probability. I also thank Prof. Adel S. Aldosary from the City and Regional Planning department, KFUPM, for his contribution to enriching my knowledge in the application of statistics. I have also been lucky to be supported by the Bangladeshi students' community at KFUPM. Mr. Omar Faruque, one of my seniors from the department of Chemical Engineering at KFUPM, helped me a lot. Finally, I thank everyone who supported me.

TABLE OF CONTENTS

ACKNOWLEDGMENTS	V
TABLE OF CONTENTS.....	VI
LIST OF TABLES.....	IX
LIST OF FIGURES.....	X
LIST OF ABBREVIATIONS.....	XIII
ABSTRACT	XVII
ملخص الرسالة	XVIII
CHAPTER 1 INTRODUCTION.....	1
1.1 Available High-Dimensional Control Chart Studies	3
1.2 General Structure of Multivariate Control Charts	8
1.3 Outlines of the Remaining Chapters	10
CHAPTER 2 LITERATURE REVIEW	12
2.1 High-Dimensional Control Charts using Dimension Reduction	14
2.1.1 The Hotelling's T ² Charts.....	15
2.1.2 The MEWMA Charts	18
2.1.3 The MCUSUM Charts.....	20
2.1.4 Other Charts.....	21
2.2 High-Dimensional Control Charts using Variable Selection Techniques.....	23

2.2.1	The Hotelling's T^2 Charts	24
2.2.2	The MEWMA Charts	24
2.2.3	The MCUSUM Charts	26
2.2.4	Other Charts	26
2.3	High-Dimensional Control Charts without the use of Dimension Reduction and Variable Selection Techniques	27
2.3.1	The Hotelling's T^2 Charts	28
2.3.2	The MEWMA Charts	29
2.3.3	The MCUSUM Charts	33
2.3.4	Other Charts	35
2.4	Research Questions	43
CHAPTER 3 HIGH-DIMENSIONAL CONTROL CHARTS UNDER SHEWHART ENVIRONMENT		44
3.1	Two-Sample Hotelling's T^2 Chart	45
3.2	Proposed Two-Sample High-dimensional Charts	46
3.3	Design of Simulation Study	48
3.4	Results and Discussions	50
3.4.1	Performance of the Proposed Charts under Lower Dimensions	50
3.4.2	Performance of the Proposed Charts under Higher Dimensions	56
3.5	A Case Study about Bearings of Wind Turbines	63
CHAPTER 4 HIGH-DIMENSIONAL CONTROL CHARTS UNDER EWMA ENVIRONMENT		68
4.1	Proposed Two-Sample High-dimensional Charts	69
4.2	Design of Simulation Study	72
4.3	Results and Discussions	73
4.3.1	Performance of the Proposed Charts	74

4.3.2	Performance Comparison of Memory Charts with Memoryless Charts	80
4.4	Application of proposed Charts on Wind Turbine Bearings.....	84
CHAPTER 5 HIGH-DIMENSIONAL CONTROL CHARTS UNDER CUSUM ENVIRONMENT		89
5.1	Proposed Two-Sample High-dimensional Charts	92
5.2	Design of Simulation Study.....	95
5.3	Results and Discussions	96
5.3.1	Performance of the Proposed Charts	97
5.3.2	Performance Comparison between DR, DREWMA, and DRCUSUM Charts	100
CHAPTER 6 SUMMARY, CONCLUSIONS AND RECOMMENDATIONS.....		105
6.1	Concluding Remarks Based on Shewahrt Setup	106
6.2	Concluding Remarks Based on EWMA Setup	107
6.3	Concluding Remarks Based on CUSUM Setup	109
6.4	Future Research Directions.....	110
REFERENCES.....		115
VITAE		124

LIST OF TABLES

Table 1: High-Dimensional datasets used in the selected studies.....	7
Table 2: Summary of the High-Dimensional control chart literature	13
Table 3: Control limits of the proposed charts under lower dimensions with $m = 30$ and $n = 30$	52
Table 4: ARL (SDRL) comparison of Hotelling's T^2 and proposed charts with $m = 30$, $n = 30$	53
Table 5: Control limits of the proposed charts under higher dimensions	58
Table 6: ARL (SDRL) comparison of DR and SD charts with $m = p \times 30\%$, $n =$ $p \times 30\%$ and $m = p \times 30\%$, $n = p \times 45\%$	60
Table 7: Mean, standard deviation, L1 and L2 for DREWMA and SDEWMA charts with $m = n = p \times 30\%$ and $m = p \times 30\%$, $n = p \times 45\%$	76
Table 8: ARL (SDRL) comparison of DREWMA and SDEWMA charts with $m = n =$ $p \times 30\%$ and $m = p \times 30\%$, $n = p \times 45\%$	77
Table 9: Comparison of DR with DREWMA charts, and SD with SDEWMA charts.....	81
Table 10: Mean, standard deviation, and h values for DRCUSUM chart with $m = n =$ $p \times 30\%$	98
Table 11: Run length profile of DRCUSUM based on ARL (SDRL) with $m = n =$ $p \times 30\%$	99
Table 12: Comparison between DR, DREWMA, and DRCUSUM charts for $m = n =$ $p \times 30\%$	101

LIST OF FIGURES

Figure 1: Number of high-dimensional control chart studies over the years.....	4
Figure 2: Count plot for high-dimensional control chart studies with respect to approaches and structures	4
Figure 3: Keyword co-occurrences of high-dimensional control chart studies	6
Figure 4: ARL comparison of Hotelling's T2 and proposed charts with $m = 30$, $n = 30$ for MVN0	54
Figure 5: ARL comparison of Hotelling's T2 and proposed charts with $m = 30$, $n = 30$ for MVN1	54
Figure 6: ARL comparison of Hotelling's T2 and proposed charts with $m = 30$, $n = 30$ for MVT	55
Figure 7: ARL comparison of Hotelling's T2 and proposed charts with $m = 30$, $n = 30$ for MVG	55
Figure 8: ARL comparison of DR and SD charts with 4 combinations of dimensions and sample sizes for MVN0	61
Figure 9: ARL comparison of DR and SD charts with 4 combinations of dimensions and sample sizes for MVN0	61
Figure 10: ARL comparison of DR and SD charts with 4 combinations of dimensions and sample sizes for MVT	62
Figure 11: ARL comparison of DR and SD charts with 4 combinations of dimensions and sample sizes for MVG.....	62
Figure 12: Wind turbine configuration (A) gearbox operated wind turbine [113] and (B) direct-drive wind turbine [114]	64

Figure 13: Change point analysis by fitting different models.....	65
Figure 14: Graphical displays of plotting statistics for DR, BS, and SD charts corresponding to their control limits	67
Figure 15: ARL comparison of DREWMA and SDEWMA charts with 4 combinations of dimensions and sample sizes for MVN0.....	78
Figure 16: ARL comparison of DREWMA and SDEWMA charts with 4 combinations of dimensions and sample sizes for MVN1.....	78
Figure 17: ARL comparison of DREWMA and SDEWMA charts with 4 combinations of dimensions and sample sizes for MVT	79
Figure 18: ARL comparison of DREWMA and SDEWMA charts with 4 combinations of dimensions and sample sizes for MVG	79
Figure 19: Comparison of (A) DR with DREWMA charts, and (B) SD with SDEWMA charts for MVN0	82
Figure 20: Comparison of (A) DR with DREWMA charts, and (B) SD with SDEWMA charts for MVN1	82
Figure 21: Comparison of (A) DR with DREWMA charts, and (B) SD with SDEWMA charts for MVT	83
Figure 22: Comparison of (A) DR with DREWMA charts, and (B) SD with SDEWMA charts for MVG	83
Figure 23: Wind turbine bearings [118].....	85
Figure 24: Graphical displays of plotting statistics for DREWMA, BSEWMA, and SDEWMA charts corresponding to their control limits.....	88
Figure 25: Comparison between DR, DREWMA, and DRCUSUM charts for MVN0..	102

Figure 26: Comparison between DR, DREWMA, and DRCUSUM charts for MVN1.. 102

Figure 27: Comparison between DR, DREWMA, and DRCUSUM charts for MVT.... 103

Figure 28: Comparison between DR, DREWMA, and DRCUSUM charts for MVG ... 103

LIST OF ABBREVIATIONS

1DRCAE	:	One-Dimension Residual Convolutional Auto-Encoder
ADD	:	Average Detection Delay
ADT	:	Area Delaunay Triangulation
ALEWMA	:	Adaptive LEWMA
ALT	:	Adaptive LASSO Thresholding
ARL	:	Average Run Length
ASD	:	Adaptive Step-Down
AVS-MEWMA	:	Adaptive VS-MEWMA
BSS	:	Blind Source Separation
BSSCD	:	Bayesian Spike-Slab Composite Decomposition
BSU	:	Bayesian Sequential Update
BSVDD	:	Bayesian Support Vector Data Description
CE	:	Conditional Entropy
CR	:	Correctness Rates
DFASD	:	Distribution Free Adaptive Step-Down
DL	:	Deep Learning
DNN	:	Deep Neural Network
DPCA	:	Dynamic PCA
DPEWMA	:	Data-Driven Pairwise EWMA
EER	:	Expected Error Rates
EQL	:	Extra Quadratic Loss
EWRL	:	Expectation Weighted Run Length
FAR	:	False Alarm Rate
Fast R-CNN	:	Fast Region based Convolutional Network
FBFP	:	Fed-Batch Fermentation Penicillin
FCR	:	Fault Correctness Ratio

FDR	:	Fault Detection Rate
FPCA	:	Functional PCA
GLR	:	Generalized Likelihood Ratio
GMM	:	Gaussian Mixture Model
HBDS	:	High-Dimensional Binary Data Streams
HDEWMA	:	High-Dimensional EWMA
ICA	:	Independent Component Analysis
IHDEWMA	:	Individual HDEWMA
JD	:	Joint Decorrelation
KNN_MSE	:	K-Nearest Neighbor-Mean Squared Errors
KPCA	:	Kernel Principal Component Analysis
KPLSDA	:	Kernel PLSDA
KPLSDA_MAX_E	:	KPLSDA Maximum Error
KPLSDA_MAX_KE	:	KPLSDA top K Errors
KPLSDA_MSE	:	KPLSDA Mean Squared Errors
LAM_MSE	:	Linear Autoregressive Model-Mean Squared Errors
LEWMA	:	LASSO-based EWMA
LGPCA	:	Local and Global Principal Component Analysis
LPP	:	Locality Preserving Projections
LSMEWMA	:	Least Squares Support Vector Data Description
MEWMV	:	Multivariate Exponentially Weighted Moving Variance
MFPCA	:	Multichannel Functional Principal Component Analysis
MLPP	:	Multiway Locality Preserving Projections
MPCA	:	Multi-linear PCA
MPCA	:	Multiway Principal Component Analysis
MRSAE	:	Manifold Regularized Stacked Auto-Encoders
MEWMA	:	Multivariate Exponentially Weighted Moving Average
MCUSUM	:	Multivariate Cumulative Sum
NCSC	:	Nearest Correlation Spectral Clustering

NFL	:	National Football League
NKGM	:	Nonlinear Kernel Gaussian Mixture Model
OCP	:	One-Class Peeling
PCA	:	Principal Component Analysis
PHD-EWMA	:	Paired HDEWMA
PLR	:	Penalized Likelihood Ratio
PLSDA	:	Partial Least Squares Discriminant Analysis
PSA	:	Pressure Swing Adsorption
PVS – MPCA	:	Pareto Variable Selection-Multiway Principal
RARL	:	Relative Average Run Length
RMDP	:	Robust Minimum Diagonal Product
RMI	:	Relative Mean Index
RPLR	:	Ridge Penalized Likelihood Ratio
RTC	:	Real-Time Contrasts
SDAE	:	Stacked Denoising Auto-Encoders
SDRL	:	Standard Deviations of the Run-Length
SMEWMA	:	Support Vector Data Description MEWMA
SMFPCA	:	Sparse Multi-channel Functional Principal Component
SPC	:	Statistical Process Control
SPCRAE	:	Sparse Principal Component Regression with
SPE	:	Squared Prediction Error
SPE	:	Stochastic Proximity Embedding
SSMRL	:	Steady-State Median Run Length
STVDD	:	Support Tensor Vector Data Description
SVDD	:	Support Vector Data Description
TEP	:	Tennessee Eastman Process
TLPP	:	Tensor Locality Preserving Projections
TRAS	:	Top-r based Adaptive Sampling
UMPCA	:	Uncorrelated Multi-linear PCA

VAE	:	Variational Auto-Encoder
VDP	:	Vertical Density Profiles
VPCA	:	Vectorized PCA
VS-MSPC	:	Variable-Selection-Based Multivariate SPC
WVFD	:	Water Vapor Flux Divergence

ABSTRACT

Full Name : FUHAD AHMED

Thesis Title : ON STATISTICAL SURVEILLANCE OF TWO-SAMPLED HIGH-DIMENSIONAL DATA

Major Field : APPLIED STATISTICS

Date of Degree : DECEMBER 2023

Monitoring a process with high-dimensional data having more dimensions or features than observations is a significant endeavor in the field of Statistical Process Control (SPC). The shift detection ability of the conventional control charts in a process with low-dimensional data is satisfactory. However, the performance of these charts deteriorates as dimension increases. For high-dimensional data, almost all the previous studies focused on dealing with one-sample case, leaving a research gap to deal with two-sample cases. Hence, in this thesis, to monitor location shift in the two-sample high-dimensional process, both memoryless and memory-based new control charts are proposed without using traditional dimension reduction or variable selection methods. The performance of the proposed charts is found to be superior to the available charts based on the performance evaluation measures. The proposed memory-type charts outperform the proposed memoryless charts. Both types of constructed charts are also implemented on real-life high-dimensional data.

ملخص الرسالة

الاسم الكامل: فوهاد أحمد

عنوان الرسالة: الرصد الإحصائي لعينتين عالية الأبعاد.

التخصص: إحصاء تطبيقي

تاريخ الدرجة العلمية: ديسمبر 2023

تعد عملية المراقبة لبيانات عالية الأبعاد مسعىً مهماً في مجال التحكم الإحصائي للعمليات حيث تعرف البيانات عالية الأبعاد بأنها بيانات عدد السمات فيها يفوق عدد العينات. مخططات التحكم التقليدية لديها قدرة مرضية للكشف عن تحول ما في موقع عملية ذات بُعد منخفض، ولكن سرعان ما تتدهور هذه القدرة في حالة ازدياد الأبعاد. لبيانات عالية الأبعاد، معظم الدراسات السابقة ركزت على حالة التعامل مع عينة واحدة، تاركة فجوة بحثية لحالة التعامل مع عينتين. في هذا البحث، تم استخدام كلاً من مخططات التحكم الجديدة ذات الذاكرة وبدون الذاكرة لمراقبة التحول في موقع عملية مكونة من عينتين عالية الأبعاد من غير استخدام طرق التقليل البعدي التقليدية أو طرق اختيار المتغير. بناءً على مقاييس تقييم الأداء، وُجد أن أداء المخططات المقترحة متفوق مقارنة بأداء المخططات المتاحة، وأن أداء المخططات المقترحة ذات الذاكرة أفضل من المخططات بدون الذاكرة. كما تم تطبيق المخططات المقترحة على بيانات عالية الأبعاد من واقع الحياة.

CHAPTER 1

INTRODUCTION

In recent years, collecting and analyzing high-dimensional data has become integral to many fields, including bioinformatics, finance, physics, social science, and machine learning. High-dimensional data refers to data sets that contain many variables or features, typically with many more features (p) than observations (n) (i.e., $n < p$). The simultaneous collection of huge number of features on each person or object is a key element of modern data. These data are referred to as high-dimensional data. For instance, in high-throughput DNA sequencing, the number of features recorded for each patient can range in the tens of thousands, while the number of patients is modest. Similarly, in speech or image recognition, the data set may include many features but a few labeled examples. An image can be considered high-dimensional data, with each pixel representing a separate feature. High dimensional data is commonly used in the field of medical science, e-commerce, environmental science, and finance.

In the current era, analyzing and interpreting high-dimensional data is challenging because it can be difficult to visualize and comprehend data with many dimensions. It can also lead to issues with overfitting and spurious correlations if not adequately handled. High-dimensional data analysis thus poses unique challenges that traditional statistical methods cannot capture. The primary challenge associated with high-dimensional data relates to the well-known "curse of dimensionality" phenomenon. As the number of features increases,

the data become very sparse, and the distance between observations in the high-dimensional space becomes meaningless. Moreover, increasing the number of features increases the risk of overfitting, leading to the poor predictive accuracy of the model. Traditional statistical methods can only handle a small number of features and have become inefficient in handling high-dimensional data. Therefore, the need for methods that can handle high-dimensional data has given rise to a new area of research known as "high-dimensional data analysis." High-dimensional data analysis aims to develop statistical methods that can extract meaningful patterns or structures from high-dimensional data while overcoming the challenges posed by the "curse of dimensionality." These methods must be computationally efficient, statistically robust, and scalable.

In the era of industry 4.0, revolution in sensors technology provided a remarkable feature of storing high-dimensional data in temporal frames. Hence, most industries are interested in monitoring such high-dimensional data to enhance the quality of their service sector. Over the last few decades, statistical process control (SPC) methodologies have been utilized to help surveil and control a production process to ensure that it operates efficiently, effectively, and consistently. This methodology involves collecting and analyzing data in real-time to identify and respond to variations in the production process. Control charts are one of the seven magnificent SPC tools which aim to detect and prevent abnormalities or deviations from the expected process performance. However, with the advent of high-dimensional data, traditional control charts may not be sufficient because they fail to detect complex interactions between variables, leading to poor process monitoring. Therefore, in the last few decades, many studies have been designed to propose control charts based on high-dimensional data.

To monitor high-dimensional data, new SPC techniques [2-6] are required. Many studies also considered process monitoring as high-dimensional problems, even when the dimensions of data exceed three, but may not exceed the number of observations in the data [7-9]. Moreover, over the years, high-dimensional process monitoring has been carried out mainly for one-sampled datasets [6, 10-14]. Inspired from this research gap, the current study will focus on the application of SPC techniques on two-sampled high-dimensional data. The integration of new methods for two-sampled high-dimensional data into SPC can lead to more effective monitoring and control of high-dimensional industrial processes.

1.1 Available High-Dimensional Control Chart Studies

In this study, a total of 82 articles between 2004 and 2023 were reviewed. As shown in Figure 1, a few studies were initially conducted between 2004 and 2016, where only 25 articles were published on high-dimensional control charts during that time span. After that, there was an immense increase in literature, with 57 studies being published on high-dimensional monitoring schemes. This trend demonstrates the importance of designing review studies for the surveillance of high-dimensional data.

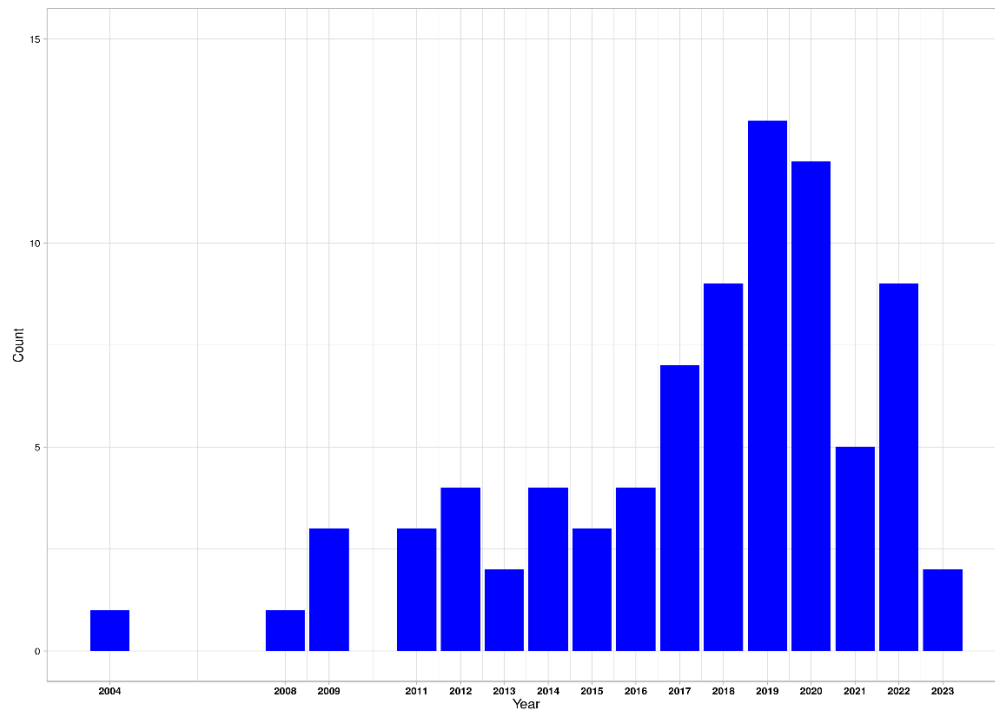


Figure 1: Number of high-dimensional control chart studies over the years

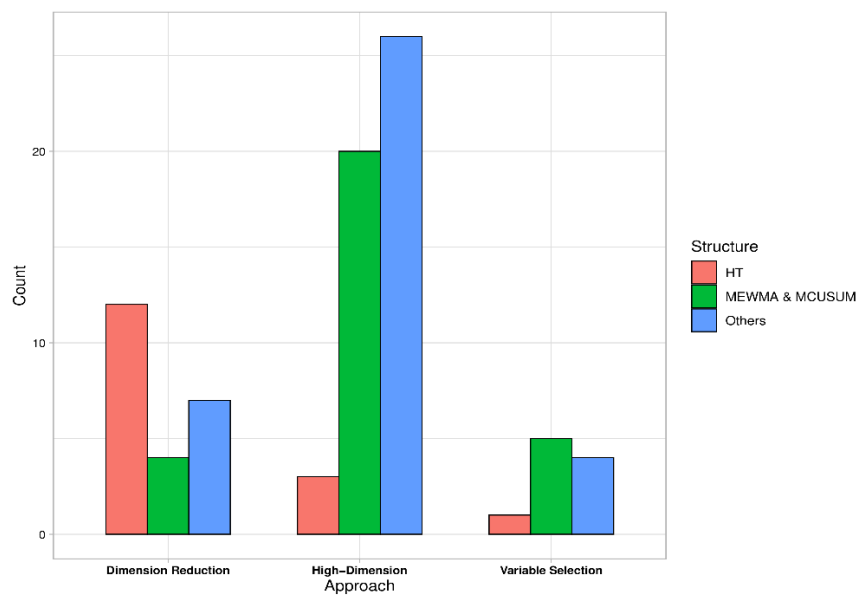


Figure 2: Count plot for high-dimensional control chart studies with respect to approaches and structures

As shown in Figure 2, among the 82 studies on high-dimensional process monitoring, 23 used dimension-reduction techniques to reduce the data dimensions before implementing traditional multivariate control charts. Of these, 12 studies applied Hotelling's T^2 (HT) control chart, 4 studies used Multivariate Exponentially Weighted Moving Average (MEWMA) or Multivariate Cumulative Sum (MCUSUM) charts, and the rest used other charts. Ten studies utilized the variable selection approach to reduce the data before implementing traditional multivariate control charts. Among them, only one article proposed Hotelling's T^2 (HT) scheme, five suggested MEWMA or MCUSUM control charts, and four applied charts that were different from Hotelling's T^2 (HT), MEWMA, and MCUSUM. A large number of studies (49 articles) considered neither dimension reduction nor variable selection methods but were purely based on high-dimensional data techniques. Among these studies, only three articles applied memory-less monitoring structures, 20 studies were based on the memory type, and 26 studies were based on other structures.

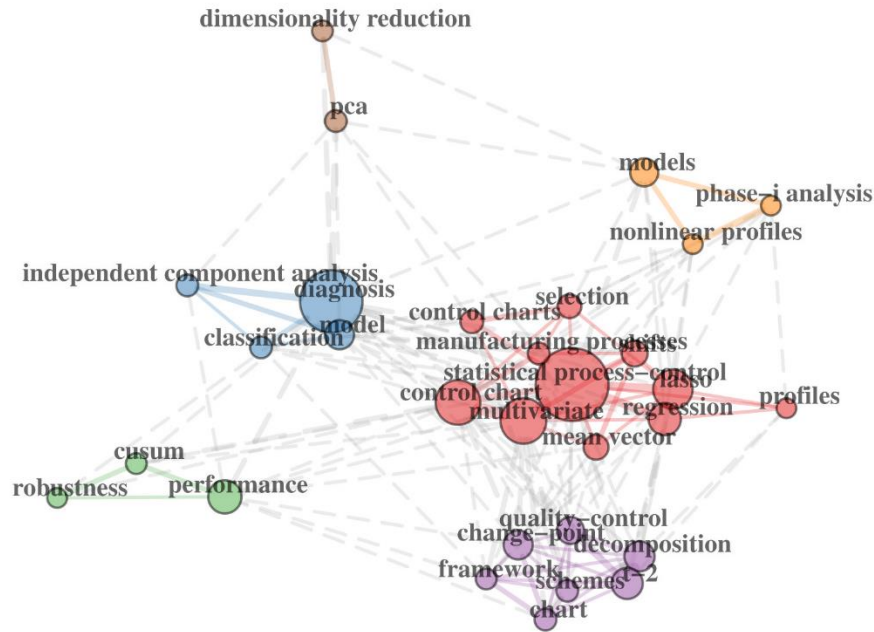


Figure 3: Keyword co-occurrences of high-dimensional control chart studies

Figure 3 shows the co-occurrences of keywords used in studies focusing on high-dimensional process monitoring. It is evident that the keywords “dimensionality reduction” and “pca” occurred together frequently, whereas “cusum,” “robustness,” and “performance” are the three keywords that appeared together in the high-dimensional monitoring studies. Similarly, other co-occurring keywords were “control charts,” “statistical process control,” “multivariate,” “regression,” “profiles,” and “mean vector.”

In the SPC literature, various high-dimensional data types are utilized, including those from the industrial, chemical, imaging, solar, and medical sciences. We found that the utilization of industrial data was most prevalent in this review. A total of 21 datasets based on the industry sector were found among the numerous high-dimensional dataset types. Whereas the numbers of high-dimensional datasets used from the chemical and medical areas are 7 and 5, respectively.

Table 1: High-Dimensional datasets used in the selected studies

Sector	Dataset	References	Dataset	References
Industrial Datasets	Industrial penicillin production plant	Hu and Yuan [15] Hu and Yuan [16] Wang, et al. [17]	PCB packaging process	Wang, et al. [18]
	Spur gear production	Abdella, et al. [8]	High-speed milling process	Kim, et al. [2]
	A hexagonal blot manufacturing	Abdella, et al. [19]	Chocolate conching batch process	Peres, et al. [20]
	Semiconductor manufacturing process	Zou, et al. [21] Zou, et al. [22], Shu and Fan [23], Abdella, et al. [24], Li [3], Li, et al. [25] Zhang, et al. [26], Mukherjee and Marozzi [27] Xiang, et al. [28] Ye, et al. [29]	Tennessee Eastman process (TEP)	Yu [30] Samuel and Cao [31] Yu and Zhang [13] Kim and Turkoz [32] Ye, et al. [29] Wang, et al. [17]
	Paper product manufacturing process	Alaeddini, et al. [33]	White wine production process	Zou, et al. [34], Li and Tsung [35], Li, et al. [9] Merlo, et al. [36]
	Bolt monitoring	Kim, et al. [37] Turkoz, et al. [38]	Forging process	Wang, et al. [39]
	Aluminium smelter company	Zhao, et al. [40]	Bearing	Yu [41]
	Aircraft maintenance	Qi, et al. [42]	Surface contact quality of the radio frequency MEMS switch	Huang, et al. [43]
	Steel rolling process, and a stamping process	Yan, et al. [44]	Soil temperature	Gómez, et al. [45]
	Continuous stirred tank heater	Kim, et al. [46]	Hard disk drives monitoring system	Du and Zou [47]
	Logistic company	Zou, et al. [48]		
Medical Science Data	Breast Cancer	Kim, et al. [49] Fan, et al. [50] Maboudou-Tchao, et al. [51]	Biomanufacturing Process & Ultra-precision Machining	Kan and Yang [11] Kan and Yang [52]
	Lung cancer	Maboudou-Tchao [53]	Metabolomics	Ullah, et al. [54]
	E. coli	Yan, et al. [55]		
Chemical Data	LNG process	Sangahn [56]		
	Wastewater treatment process	Yu [57] Odom, et al. [58]	Heavy oil-fired combustion test facility	Sun, et al. [10]
	Water quality monitoring	Chen, et al. [59]	Water vapor flux divergence (WVFD) dataset	Fang, et al. [60]
	Pressure swing adsorption (PSA)	Wang, et al. [17]	Thin-film-transistor liquid-crystal display process	Lee, et al. [61]
Image Data	Cherenkov gamma-ray telescope	Jang, et al. [62]	Camera performance measurement	He, et al. [63]
Solar Data	Solar data	Gómez, et al. [45]	Solar flare detection dataset	Guo, et al. [64]
	Solar data observatory	Yan, et al. [44]		
Miscellaneous	Monthly rates of vehicle robberies	Moala, et al. [65]	Traffic monitoring	Feng, et al. [66]
	National Football League (NFL)	Martinez, et al. [67]	Sonar	Maboudou-Tchao [53]
	KDDCUP99/NSL_KDD/UNSW-NB15/ISCX-URL2016/TON_IoT	Shaohui, et al. [68]	Vertical density profiles (VDP)	Wang and Jiang [69], Zhang and Albin [70] Chen and Nembhard [71]

1.2 General Structure of Multivariate Control Charts

Assume $X_1, X_2, \dots, X_p$ are random variables that follow a multivariate normal distribution with mean vector μ_0 and covariance matrix Σ_0 . Let the i^{th} observation vector of the process be denoted by X_{ij} , where $i = 1, 2, \dots, n$; $j = 1, 2, \dots, p$. The Hotelling's T^2 statistic is defined as:

$$T^2 = (X_{ij} - \bar{X}_j)' S^{-1} (X_{ij} - \bar{X}_j),$$

where $\bar{X}_j = \frac{1}{n} \sum_{i=1}^n X_{ij}$ and $S = \frac{1}{n-1} \sum_{i=1}^n (X_{ij} - \bar{X}_j)(X_{ij} - \bar{X}_j)'$.

The lower control limit of Hotelling's T^2 chart is set to be zero while the upper control limit under phase I is obtained by $UCL = \frac{(n-1)^2}{n} \beta_{\alpha, p/2, (n-p-1)/2}$, where $\beta_{\alpha, p/2, (n-p-1)/2}$ is upper α^{th} percentile of beta distribution with parameters $p/2$ and $(n-p-1)/2$ [72, 73]. Similarly, the upper control limit for phase II is obtained by $UCL = \frac{p(n+1)(n-1)}{n^2-np} F_{\alpha, p, (n-p)}$, where $F_{\alpha, p, (n-p)}$ is the upper α^{th} percentile of the F distribution with parameters p and $(n-p)$. The reasons behind using different distributions for the control limits under phase I and phase II are as follows: (a) In phase I, the chart is used retrospectively to check if the process is in-control state when the first subgroups of observations are drawn. Whereas, in Phase II, for future subgroups of observations, it is tested to check whether the process remains in-control or there is any shift from in-control situation. (b) In phase I, the i^{th} observation vector X_{ij} is not independent of the estimators

$\bar{X}_j$ and S , then the $\frac{T^2}{\frac{(n-1)^2}{n}}$ follows a β -distribution with $p/2$ and $(n - p - 1)/2$ degrees of freedom [72, 73]. On the other hand, in phase II, X_{ij} represents future observation vector which is independent of the estimators $\bar{X}_j$ and S , then the $\frac{T^2}{\frac{p(n+1)(n-1)}{n^2 - np}}$ follows a F -distribution with p and $(n - p)$ degrees of freedom [72, 73].

Note that Hotelling's T^2 chart is effective only for detecting large shifts in the parameter of interest. However, a Multivariate Exponentially Weighted Moving Average (MEWMA) chart is preferred for detecting small to moderate shifts in the parameters of interest. Lowry, et al. [74] introduced the MEWMA chart, where the MEWMA statistic is defined as:

$$z_{ij} = R(X_{ij} - \mu_0) + (I - R)z_{ij-1},$$

Where $R = \text{diag}(r_1, r_2, \dots, r_p)$ is user-chosen smoothing parameters $0 \leq r_j \leq 1$ for $j = 1, 2, \dots, p$, $z_{0j} = 0$, and I is the identity matrix. If the charted statistic $E_i^2 = z^T \Sigma_z^{-1} z > UCL$, the MEWMA chart gives an out-of-control signal. Here, Σ_z is the variance-covariance matrix of z_{ij} , and UCL is the upper control limit of the MEWMA chart.

The Multivariate Cumulative Sum (MCUSUM) chart is an alternative for monitoring small to moderate shifts in the parameter of interest. Crosier [75] presented two procedures by extending the univariate CUSUM chart to a multivariate case, and named it the MCUSUM control chart. The MCUSUM statistic can be expressed as:

$$C_i = \left[(S_{i-1} + \bar{X}_i - \mu_0)' \left(\frac{\Sigma}{n} \right)^{-1} (S_{i-1} + \bar{X}_i - \mu_0) \right]^{1/2}$$

where the covariance matrix is Σ , and the cumulative sums S_i are defined as

$$S_i = \begin{cases} 0, & \text{if } C_i \leq k \\ (S_{i-1} + \bar{X}_i - \mu_0) \left(1 - \frac{k}{C_i}\right), & \text{if } C_i > k \end{cases}$$

If the charting statistic $T_i^2 = [S_i' \left(\frac{\Sigma}{n}\right)^{-1} S_i]^{1/2} > h$, the MCUSUM chart gives an out-of-control signal. It is noteworthy that h is the upper control limit (UCL).

1.3 Outlines of the Remaining Chapters

The outlines of the remaining chapters are as follows:

- The aim of Chapter 2 is to formulate a comprehensive review of high-dimensional control chart studies to find the possible gaps with detailed insights so that further contribution can be offered.
- Chapter 3 will focus on the construction of the memory-less structure (Shewhart-type) to capture a large amount of shifts in two-sampled high-dimensional data. The constructed charts will be evaluated based on run length profile. A case study will also be conducted to determine the effectiveness of the charts.
- The target of Chapter 4 is to design the memory-type monitoring schemes under EWMA environment to detect small or moderate amounts of shifts in two-sampled high-dimensional data. Performance evaluation measures will be used to evaluate the efficiency of the proposed scheme. The application of the proposed scheme on real-life data will also be justified.

- In Chapter 5, the aim is to propose better memory charts than EWMA setup. With this goal, the chart will be constructed based on CUSUM setup. Then further comparison will be made between the performance of these two memory charts along with the comparison of memoryless chart.
- Finally, Chapter 6 will discuss the summary, concluding remarks, and the future research directions based on the complete study conducted in this thesis.

CHAPTER 2

LITERATURE REVIEW

In this review, we found that the literature on high-dimensional monitoring methods can be classified into three approaches:

- (i) Dimension-reduction-based control charts: First, the dimension-reduction technique is used to reduce the dimensions of the data, and then traditional multivariate control charts are implemented.
- (ii) Variable-selection-based control charts: First, the variable selection method is used to reduce the data, and then traditional multivariate control charts are implemented.
- (iii) High-dimensional-technique-based control charts, which adopt high-dimensional control charts without dimension reduction or variable selection.

Table 2: Summary of the High-Dimensional control chart literature

Technique	Control Charts		
	Hotelling's T^2	MEWMA & MCUSUM	Others
Dimension Reduction	Yu [30], Lee, et al. [76], Hu and Yuan [15], Skubalska-Rafajłowicz [77], Sun, et al. [10], Samuel and Cao [31], Fujiwara and Kano [78], Odom, et al. [58], Lee, et al. [61], Yu and Zhang [13], Shaohui, et al. [68], Yu and Liu [79]	Yu [41], Zhang, et al. [26], Chen, et al. [59] Gómez, et al. [45]	Ning and Tsung [80], Bae, et al. [81], Kan and Yang [52], Zhang, et al. [82], Zhang, et al. [83], Wang, et al. [39], Kim, et al. [49]
Variable Selection	Kim, et al. [37]	Zou, et al. [34], Zou, et al. [48], Abdella, et al. [19], Sangahn [56], Zhao, et al. [40]	Wang and Jiang [69] Peres, et al. [20], Yan, et al. [55], Guo, et al. [64]
High-Dimension	Ullah, et al. [54], Ahmad and Ahmed [6], Ahmadi-Javid and Ebadi [14]	Huang, et al. [43], Liu, et al. [84], Zou, et al. [22], Zou, et al. [22] Qi, et al. [42], Wang, et al. [18], Alaeddini, et al. [33], Du and Zou [47], Li [3], Li and Tsung [35], Moala, et al. [65], Ahmad, et al. [85], Feng, et al. [66], Kim, et al. [2], Li [86], Li, et al. [9], Fang, et al. [60], Kim, et al. [46], Maboudou-Tchao, et al. [51], Ye, et al. [29]	Hu and Yuan [16], Zhang and Albin [70], Chen and Nemhard [71], Yu [57], Li, et al. [87], Zou, et al. [21], Kan and Yang [11] Turkoz, et al. [88], Jang, et al. [62], Wang, et al. [17], He, et al. [63], Shu and Fan [23], Yan, et al. [44], Abdella, et al. [24], Guo, et al. [89], Kim, et al. [7] Turkoz, et al. [38] Abdella, et al. [8], Li, et al. [25], Martinez, et al. [67], Fan, et al. [50], Maboudou-Tchao [53], Merlo, et al. [36], Mukherjee and Marozzi [27], Kim and Turkoz [32], Xiang, et al. [28]

Furthermore, the literature on each approach is categorized with respect to different structures such as memory-less (Hotelling's T^2), memory type (i.e., MEWMA and

MCUSUM), and other control charts. The divisions of all studies with respect to approaches and structures are listed in Table 2. The details of each approach with respect to the structures are provided in the preceding subsections.

2.1 High-Dimensional Control Charts using Dimension Reduction

Dimension reduction is the process of reducing a dataset's features, dimensions, or variables to a lower-dimensional space while preserving as much essential information as possible. Dimension reduction aims to reduce the complexity of data for better understanding, visualization, and analysis, while avoiding issues that arise with high-dimensional data, such as the dimensionality curse. In this review, dimension reduction methods are classified into three types: component analysis, projection-based, and other. Currently, component analysis is widely used for dimension reduction in SPCs. Commonly used component analysis methods are Principal Component Analysis (PCA), Local and Global PCA (LGPCA), Multiway PCA (MPCA), Kernel PCA (KPCA), Multistate Adaptive-Dynamic PCA (MSAD-PCA), Adaptive-Dynamic PCA (AD-PCA), Functional PCA (FPCA), Multichannel Functional PCA (MFPCA), Uncorrelated Multi-linear PCA (UMPCA), Multi-linear PCA (MPCA), Vectorized PCA (VPCA), and Sparse Multichannel Functional PCA (SMFPCA). Some projection-based methods used in SPC include Locality Preserving Projections (LPP), Multiway Locality Preserving Projections (MLPP), Normal Random Projections (NRP), Ensemble Random Projections (ERP), and Tensor Locality Preserving Projections (TLPP). Whereas Joint Decorrelation (JD) and Stochastic Proximity Embedding (JD-SPE), One-Dimension Residual Convolutional Auto-Encoder (1DRCAE), Recursive Tensor Recovery (RTR), Spatial Scanning (SS), Local Outlier Factor (LOF), and Simplicial Depth Statistic (SDS) are some other methods that are also

used for reducing dimensionality in SPC. The details on the SPC methods using the above-stated dimension reduction methods are presented in the following subsections.

2.1.1 The Hotelling's T^2 Charts

In the last few decades, several studies have been designed in which dimension reduction techniques were used to reduce the dimensionality of high-dimensional process data, and Hotelling's T^2 chart was implemented to monitor the parameter(s) of the process. For example, Lee, et al. [76] used the KPCA to reduce the variables of high-dimensional data by capturing nonlinear relationships in the process variables. Furthermore, Hotelling's T^2 and squared prediction error (SPE) charts were proposed to monitor variations in the process mean, and the implications of the proposed charts were demonstrated using the biological wastewater treatment process. Hu and Yuan [15] implemented MLPP for the dimension reduction of high-dimensional data and proposed Hotelling's T^2 and SPE charts for process monitoring. A case study was carried out with data from industrial penicillin fed-batch cultivations, and the findings revealed that the MLPP-based approach could preserve the neighborhood structure of the data, be more robust, and capture intrinsic information more effectively than the multiway principal component analysis (MPCA).

Yu [30] used the LGPCA to extract feature information from observational data by projecting them onto a low-dimensional space. The proposed LGPCA-based method combining Hotelling's T^2 and SPE statistic-based control chart was compared with charts based on PCA and LPP. The proposed chart outperformed the LPP- and PCA-based monitoring methods in terms of better data separation capability of normal and abnormal observations, higher fault detection rate, and higher sensitivity to faults. Skubalska-

Rafajłowicz [77] proposed a multivariate Hotelling's T^2 chart based on normal random projections for the surveillance of changes (or anomalies) in high-dimensional discrete-time processes. Sun, et al. [10] proposed the KPCA method for dimension reduction and used Hotelling's T^2 and Q statistics to detect abnormal conditions in the combustion process. Based on the false warning rate, the proposed KPCA approach was compared with the conventional PCA approach. The KPCA method performed better than the PCA method in detecting abnormal conditions even in situations with few training samples. Samuel and Cao [31] proposed extending the KPCA method to kernel density estimation (KDE) to account for nonlinearities in process monitoring. The initial step was to calculate Hotelling's T^2 and Q (SPE) statistics and establish their KDE-based control limits at a 99 percent confidence level. Subsequently, KPCA with KDE was applied to the Tennessee Eastman process (TEP) for fault detection, and the results were compared with those of KPCA with a normality assumption and the corresponding PCA methods. The results showed that the KPCA-based methods performed better overall than the PCA-based methods in terms of the Fault Detection Rate (FDR). Furthermore, the KDE-based methods performed better than those without KDE, whereas KPCA with the KDE approach yielded the highest FDRs. The KPCA-based False Alarm Rates (FARs) were slightly higher than those of the PCA methods; however, all the FARs were within the acceptable range. Thus, associating KPCA with KDE improves the nonlinear process monitoring performance.

Fujiwara and Kano [78] proposed a JD-SPE- T^2 chart based on a new correlation-based data visualization method, which integrates the JD and SPE methods. The JD method is a Blind Source Separation (BSS) method that can separate samples based on correlation, while the SPE method is a self-organizing algorithm that can map high-dimensional data to a two-

dimensional plane. A comparative study was designed using the existing Hotelling's T^2 chart based on MSPC, and the findings revealed that the JD-SPE- T^2 chart outperformed the existing method under consideration. Furthermore, Odom, et al. [58] used the multistate adaptive-dynamic principal component analysis (MSAD-PCA) to reduce the dimensions of nonlinear, nonstationary, and serially autocorrelated multistate multivariate process data. After dimension reduction, SPE and Hotelling's T^2 charts are proposed, and their versions are compared based on the adaptive-dynamic principal component analysis (AD-PCA). They demonstrated the behavior of Hotelling's T^2 and SPE in a simulation study, and then illustrated it with a real case study on a decentralized wastewater treatment (WWT) system. They found that by combining SPE and Hotelling's T^2 monitoring statistics after dimension reduction, MSAD-PCA was superior to AD-PCA for monitoring nonlinear, nonstationary, and serially autocorrelated multistate multivariate processes.

To address both nonlinear and non-normal situations in high-dimensional processes, Lee, et al. [61] proposed a Variational Auto-Encoder (VAE)-based control chart as a monitoring method that combines T^2 and SPE statistics. Through experiments using both real data from a thin-film transistor liquid-crystal display (TFT-LCD) process and simulated data, the effectiveness and applicability of the proposed VAE-based control chart were established. As it can quickly reduce the original data to a multivariate normal distribution, implementing the proposed VAE chart is expected to decrease undesirable false alarms and misdetections in process control. The proposed VAE-based chart outperformed the existing latent variable-based charts in the simulation setting and case study related to the TFT-LCD manufacturing process. To detect faults in complex industrial processes, Yu and Zhang [13] proposed a new Deep Neural Network (DNN) called Manifold Regularized

Stacked Autoencoders (MRSAE) for dimension reduction and then used Hotelling's T^2 and SPE based control charts. They applied the proposed method to the TEP and fed-batch fermentation penicillin (FBFP) processes and compared the performance of MRSAE with other typical DNNs, stacked denoising autoencoders (SDAE), PCA, KPCA, and LGPCA. The findings show that when compared to other methods, the MRSAE-based fault detection method exhibits a higher FDR and lower FAR, which suggests that the proposed method performs better than the other competitive methods. Moreover, Shaohui, et al. [68] proposed Hotelling's T^2 charts based on the PCA mix algorithm using the bootstrap control limit. The proposed method is compared with the conventional PCA-based Hotelling's T^2 control chart, and the findings show that the proposed method performs better in detecting anomalies in the network. Yu and Liu [79] recently used a deep neural network model called 1DRCAE for dimension reduction. A combined SPE and Hotelling's T^2 chart was proposed and compared with charts based on PCA, KPCA, LGPCA, SDAE, stack sparse autoencoder (SSAE), and PCA-support vector machine (PCA-SVM). Furthermore, the proposed chart was implemented on the TE and FBFP processes to highlight the importance of this study.

2.1.2 The MEWMA Charts

In SPC literature, several studies have been designed to implement MEWMA charts for monitoring high-dimensional process parameter(s) after dimension reduction using a variety of dimension-reduction techniques. For example, Yu [41] employed a dimension reduction and feature extraction strategy based on the LPP. An EWMA statistic was developed by combining two effective statistics, T^2 and SPE, for the bearing performance degradation assessment after the information was extracted using the LPP method. This

multivariate control chart is based on the bearing performance quantification index. First, they calculated the combined index Q statistic, which is the sum of the SPE and T^2 weighted against their respective control bounds. They then calculated the EWMA statistic based on the Q statistic. Experimental testing of the bearing testbeds was conducted to determine the effectiveness of the proposed LPP-EWMA. The experimental results showed that it could recognize slight bearing degradation at an early stage and reveal the degradation propagation of the bearing performance over its entire life. Zhang, et al. [26] used random ensemble projections to decompose a high-dimensional space into several subspaces. Then, for every subspace, a local monitoring scheme is designed based on the spatial rank test to detect local changes, and the monitoring results of all subspaces are combined for final decision-making. The proposed spatial rank-based EWMA chart with ensemble random projections is presented as an RPSR chart. A comparative study was designed between the proposed and existing charts, such as the CUSUM chart, named MCUSUM [90], the anti-rank chart named AnRank [91], and the Kernel Hibert chart named RKHS [43]. The RPSR chart performed consistently better than the MCUSUM, RKHS, and AnRank charts in the presence of correlations between different variables and had efficient detection power for sparse process changes and robust performance for non-normal distributions. Furthermore, the findings were validated using case studies on handwritten numerical identification data and semiconductor manufacturing process data.

Recently, Gómez, et al. [45] used a recursive tensor recovery (RTR) model to capture the spatiotemporal structure and convert high-dimensional streaming data into a low-dimensional subspace. In addition, they introduced a tensor sequential sampling (TSS) technique that balances exploration and exploitation by combining two probability

functions to specify where each acquisition time is sampled. The proposed monitoring structure integrates the RTR approach with the EWMA control chart. The performance of the proposed chart under the TSS technique was compared with charts dependent on local CUSUM statistics [84] (referred to as TRAS) and the spatial-adaptive sampling method [92] (referred to as SASAM). The findings revealed that the proposed method outperformed the existing methods owing to its additional feature of capturing spatial and temporal information from the data. Furthermore, two case studies using solar and soil temperature data were adopted to validate the simulated findings.

2.1.3 The MCUSUM Charts

Chen, et al. [59] proposed a set of reduced-dimension (RD) control charts that perform dimension reduction during the data acquisition process using spatial scanning to monitor the means of the high-dimensional process. A control chart was constructed for the RD observation vectors for each scan cluster and only the local covariance was utilized. The MCUSUM chart was proposed based on the LR and T^2 statistics. The charts based on all variables are referred to as the F-MCUSUM charts, including the LR-F-MCUSUM and T^2 -F-MCUSUM charts, while the reduced dimension charts are referred to as the RD-MCUSUM charts, including the LR-RD-MCUSUM and T^2 -RD-MCUSUM charts. They analytically compared the F-MCUSUM and RD-MCUSUM charts in terms of their detection performances using Average Run Length (ARL). The simulation results match the ARL measures very well. The findings show that the RD-MCUSUM charts are powerful and easy to implement compared to the F-MCUSUM charts, particularly in the case of a high-dimensional and ill-conditioned covariance matrix. A case study on water

quality monitoring demonstrates the effectiveness of the proposed method in real applications.

2.1.4 Other Charts

In addition to conventional methods, such as Hotelling's T^2 , MEWMA, or MCUSUM-based control charts, other measure-based charts have also been proposed for monitoring high-dimensional data. For example, Ning and Tsung [80] proposed a method based on a Local Outlier Factor (LOF) scheme for monitoring mixed-type high-dimensional data. Based on a dataset from a logistics company, the LOF method was compared with a simplicial depth statistic method. The results illustrated the superiority of the LOF over the simplicial depth statistic when the data have multiple clusters. Bae, et al. [81] proposed a nonparametric monitoring method based on data depth, which was used for dimension reduction in high-dimensional problems. When data vary from the normality assumption, several depth measures, notably the Tukey depth, are particularly effective for SPC purposes.

Kan and Yang [52] designed a new network-generalized likelihood ratio (NGLR) chart to effectively detect the change from the underlying process. The proposed methodology is implemented and evaluated for real-world applications in ultraprecision machining and biomanufacturing processes. An ARL comparison between NGLR and Hotelling's T^2 charts was performed, and the results showed that the proposed approach effectively characterized and monitored variations in the complex structures of time-varying image data. Wang, et al. [39] proposed a unique-threshold multivariate PCA approach for multichannel profile monitoring. The method is based on two dimension-reduction steps: first, it uses functional PCA to extract a reasonably large number of features from the in-

control state, and then employs soft-thresholding approaches to choose additional significant features from the out-of-control state that capture profile information. They reported the performance of their suggested threshold PCA approach with three choices for the soft-thresholding parameter. They used an actual forging manufacturing process dataset to test their proposed threshold PCA methodology. Their numerical analyses showed that soft thresholding could significantly enhance the PCA's performance.

Zhang, et al. [82] developed monitoring methods based on the FPCA and MFPCA to monitor multiple sensor profiles. The proposed monitoring methods were compared with monitoring methods based on UMPCA, multilinear PCA (MPCA), and VPCA in terms of ARL and standard deviations of the run length (SDRL). The findings reveal that the proposed monitoring methods are better than the existing ones under consideration, and similar results are observed in the case study of manufacturing process data. Zhang, et al. [83] proposed a novel SMFPCA for modelling multi-channel profiles. SMFPCA can flexibly describe the correlation structure of multiple or even high-dimensional profiles with distinct features, and achieve sparse and easily interpretable PCA scores. The proposed SMFPCA chart was compared with charts based on the MFPCA, UMPCA, MPCA, and VPCA in terms of ARL and SDRL. The findings revealed that the proposed SMFPCA chart has a higher efficiency than the other charts because it can correctly model the weak inter-profile correlation structure of multichannel profiles. Furthermore, the simulation findings were validated by implementing charts on the manufacturing system data. Moreover, for binary asymmetrical data with multicollinearity among independent variables, Kim, et al. [49] presented regression-based $\bar{r}$ -control charts combining principal component analysis methods (PCA and FPCA) and binary response regression models (a

generalized linear model with probit and logit link functions and a neural network regression model). Using both simulated and real breast cancer data, they performed an efficiency comparison among the proposed methods: PCA-based binary response regression model-based statistical control r-charts, FPCA-based binary response regression model-based statistical control r-charts with the GLM with probit and logit link function models, and a neural network model. The results showed that the neural-network-based r-chart is relatively effective in monitoring simulated and real binary response data.

2.2 High-Dimensional Control Charts using Variable Selection Techniques

In literature, variable selection methods are often used in high-dimensional process monitoring to select a subset of variables (also known as features or predictors) from many variables in a dataset. Choosing only relevant variables for monitoring in a high-dimensional process is crucial because the number of variables is often significantly larger than the number of observations. Variable selection aims to reduce the dimensionality of the dataset and increase fault detection accuracy. Various variable selection techniques, such as the step-down procedure, LASSO-based procedure, conditional sum of square-based procedure, adaptive variable selection, adaptive LASSO, forward-selection technique, penalized likelihood, and integrated multiway principal component analysis, are currently used for high-dimensional process monitoring.

2.2.1 The Hotelling's T^2 Charts

In the past, very few studies have been designed in which variable selection techniques were used to lower the dimensions of high-dimensional process data, and Hotelling's T^2 chart was implemented to monitor the parameter(s) of the process. For example, Kim, et al. [37] used a variable selection technique called the adaptive step-down (ASD) procedure and Hotelling's T^2 chart to identify the changed elements of the mean vector. Furthermore, the diagnostic procedure is compared with existing procedures such as MYT procedures [93], step-down procedures [94], and LASSO-based procedures [34] using correctness rates (CR) and expected error rates (EER). The findings show that the proposed method improves the diagnostic capability of identifying mean-shifted variables when a shift occurs only in a small number of variables. Moreover, the simulation results show that the proposed method performs better than the MYT and step-down procedures. When the shift levels are not small, the proposed approach is better than the LASSO-based procedure.

2.2.2 The MEWMA Charts

After using the variable selection technique, the MEWMA chart has been used in various studies on high-dimensional process monitoring. For example, based on the Bayesian information criterion (BIC), Zou, et al. [34] suggest a practical LASSO-based diagnostic process that combines BIC with a popular adaptive LASSO variable-selection technique. The MEWMA control chart was then used to monitor a real dataset from a white wine production process. It was discovered that it offers a useful substitute for multivariate process diagnostics. The proposed method is particularly appealing in practice because of the piecewise linearity of the LASSO solution path, which reduces the overall computational effort. Furthermore, Zou, et al. [48] propose a LASSO-based EWMA chart

or LEWMA chart for Phase II analysis, which is a variable-selection-based multivariate control scheme. The performance of the LEWMA chart is compared with that of the generalization of the MEWMA chart for multivariate profiles. The simulation findings demonstrate that the proposed LASSO-based method has a reasonable diagnostic capacity for identifying shifted parameters in all scenarios considered. The proposed LEWMA chart also shows good performance when applied to a real-data example related to logistics services.

Sangahn [56] used the conditional sum of the square-based variable selection method on high-dimensional data and applied the MEWMA control chart, that is, the CVS-MEWMA chart. Both a simulation and case study for monitoring the LNG process were performed to evaluate the performance of the proposed chart. The proposed CVS-MEWMA chart is compared with existing multivariate SPC charts (Shewhart, Hotelling's T^2 , and MEWMA) and recent variable selection-based control charts (VSMSPC and VS-MEWMA). The findings reveal that the proposed chart outperforms competitive charts. To monitor sparse mean shifts, Zhao, et al. [40] suggested variable selection techniques by proposing two enhanced MEWMA control charts, resulting in the adaptive VS-MEWMA (AVS-MEWMA) and adaptive LEWMA (ALEWMA) schemes. Based on expected weighted run length (EWRL) and ARL, the proposed AVS-MEWMA and ALEWMA control charts are found to be more stable when compared with the existing VS-MEWMA and LEWMA charts. Furthermore, metallic aluminum data from a chemical reaction process were used in real-world applications to demonstrate the superiority of the proposed adaptive control charts.

2.2.3 The MCUSUM Charts

Abdella, et al. [95] presented a variable-selection-based multivariate cumulative sum (VS-MCUSUM) chart to monitor the mean. The conventional MCUSUM, Hotelling's T^2 chart, and a variable selection-based chart (VS-EWMA) were used to compare the performance of the proposed chart. Unlike other existing charts with variable selection techniques, the proposed chart performs well for small process shifts when the parameters are identified. The superiority of the suggested VS-MCUSUM chart was also evaluated by implementing it on hexagonal blot manufacturing data.

2.2.4 Other Charts

Some alternative measure-based charts have also been proposed for monitoring high-dimensional data compared to traditional structures, such as Hotelling's T^2 , MEWMA, or MCUSUM control charts. For example, Wang and Jiang [69] presented a variable-selection-based Multivariate SPC (MSPC) approach known as the VS-MSPC chart for process monitoring and fault diagnosis. Potential out-of-control variables were eliminated using a forward-selection technique and suspect variables were monitored using a multivariate control chart. The efficiency of the VS-MSPC chart was demonstrated by simulated studies and a real example using Vertical Density Profiles (VDP) data, which showed better performance as compared to Hotelling's T^2 . Yan, et al. [55] proposed a VS-based control chart using penalized likelihood with a Gaussian Mixture Model (GMM) to handle high dimensionality and multimodality. A simulation experiment was conducted to evaluate the performance of the proposed GMM-VS control chart using ARL as a measure of performance. In a high-dimensional environment, the suggested GMM-VS control chart outperformed the traditional T^2 , GMM-based T^2 (GMM- T^2), and VS-MSPC charts

and demonstrated strong fault-detection ability. To overcome the limitations of Multiway Principal Component Analysis (MPCA) for high-dimensional data, Peres, et al. [20] proposed the Pareto Variable Selection (PVS)-MPCA method, which integrates variable selection with MPCA. The proposed PVS-MPCA approach deals with variable selection before constructing an MPCA for dimension reduction. The performance of the proposed PVS-MPCA method is compared with that of the classical MPCA based on Cumulative Error Rate (CER), which combines the FAR and Missed Detection Rate (MDR). A case study on the chocolate conching batch process was conducted to demonstrate the efficiency of the proposed method. Recently, Guo, et al. [64] introduced a Bayesian Spike-Slab Composite Decomposition (BSSCD) technique to decompose high-dimensional signals into normal and abnormal bases. Furthermore, a Thompson Sampling Bayesian Spike-Slab Composite Decomposition change detection TS-BSSCD(O) was proposed for anomaly detection. The proposed TS-BSSCD(O) was implemented on a solar flare detection dataset, and it outperformed the other methods under consideration by achieving efficient online anomaly detection.

2.3 High-Dimensional Control Charts without the use of Dimension Reduction and Variable Selection Techniques

Dimension reduction and variable selection are useful techniques for reducing high-dimensional data to a lower-dimensional space. However, these techniques provide insufficient information when the variables are highly correlated (known as multicollinearity). In addition, both methods are computationally intensive and time-consuming. These shortcomings have motivated researchers to select techniques that do not involve dimension reduction or variable selection. Over the years, many studies have

been designed in SPC literature to address high-dimensional process monitoring without using dimension reduction or variable selection techniques. A comprehensive review of SPC studies is presented in the following subsections.

2.3.1 The Hotelling's T^2 Charts

In order to monitor the shift in the mean for high-dimensional normally distributed multistream processes, Ahmadi-Javid and Ebadi [14] proposed a two-step technique: Hotelling and chi-square charts in the first step and a Shewhart chart for each stream in the second step. For both the independent and dependent streams, the necessary control limits and performance measures, such as statistical power and ARLs, are precisely and effectively computed. One benefit of their approach, from a practical standpoint, is that when the process is under control, only one chart is required. A thorough numerical analysis demonstrates that the proposed method generally outperforms the existing method discussed by [96]. Ullah, et al. [54] proposed a James-Stein shrinkage estimator and created n independent T^2 statistics using the leave-one-out resampling procedure of the n baseline samples. The implications of the proposed method on both simulation and metabolomics datasets show that the proposed approach outperforms the classical Hotelling's T^2 approach. Moreover, Ahmad and Ahmed [6] proposed a modified version of the T^2 statistic that is useful for high-dimensional data monitoring. The modification is straightforward, operates under mild conditions, and offers a unified framework for the upper control limits for both phases of the control charts as well as for individual observations and subgroup cases.

2.3.2 The MEWMA Charts

Huang, et al. [43] introduced a new Reproducing Kernel Hilbert Space (RKHS) -based control chart named the KMD-EWMA chart. This chart can detect a broader range of variations in processes beyond the typical changes in mean or variance identified by conventional SPC methods. This method is unique in that it can be applied to complex high-dimensional processes that do not follow a Gaussian distribution. The proposed method was compared with conventional Hotelling's T^2 and MEWMA charts using ARL as a performance evaluation measure. Additionally, the proposed KMD-EWMA control chart was demonstrated to be effective in practical applications. Furthermore, Qi, et al. [42] proposed a one-sided MEWMA chart for high-dimensional data. The proposed chart implementation on the simulated data and aircraft maintenance database revealed that the one-sided MEWMA scheme performed better than other existing schemes. Wang, et al. [18] proposed an EWMA-based chart called the CP chart, based on monitoring a high-dimensional Poisson dataset. The suggested chart is implemented on a Printed Circuit Board (PCB) packaging process, and the results demonstrate that the performance of the CP chart is better than that of the MP-CUSUM chart. Du and Zou [47] created a dynamic multiple-testing approach for high-dimensional data streams with the control of FDR-based on the EWMA scheme. Simulation studies and an illustration using a hard disk drive monitoring system showed that the proposed method can deliver satisfactory monitoring performance.

Alaeddini, et al. [33] introduced a spatiotemporal outlier detection technique to enhance the image-based monitoring system's performance based on Partial Least Squares Discriminant Analysis (PLSDA) and a control statistic using the area Delaunay

Triangulation (ADT) of the SPEs. They utilized an EWMA control chart to integrate spatial details with the temporal influence of potential outliers. The proposed method uses PLSDA, including Kernel PLSDA (KPLSDA), to evaluate images of various product types, including those with possible distortions. The proposed KPLSDA_ADT technique was compared with five other outlier detection methods: K-Nearest Neighbor-Mean Squared Errors (KNN_MSE), Linear Autoregressive Model-Mean Squared Errors (LAM_MSE), KPLSDA Mean Squared Errors (KPLSDA_MSE), KPLSDA Maximum Error (KPLSDA_MAX_E), and KPLSDA top K errors (KPLSDA_MAX_KE), and results revealed the superiority of the proposed method. Li and Tsung [35] suggested a new approach called the MVP chart, which combines a robust high-dimensional covariance matrix test with the MEWMA method to monitor variability using individual observations. A comparative study of the MVP chart's performance was designed with other competing charts such as the MaxNorm, LASSO-MEWMC (LMC), and Penalized Likelihood Ratio (PLR). The findings indicate that the proposed MVP chart is highly effective when process shifts result in changes in variance components, but less effective when shifts occur only in the off-diagonal components of the covariance matrix, especially with small dimensions. However, the MVP chart's performance is not severely affected by the high number of variables compared to the other charts, and its superiority increases as the number of variables increases. The MVP chart is particularly suitable for detecting changes in the covariance matrix in the higher dimensions that occur in blocks.

Feng, et al. [66] proposed a test statistic integrated into an MEWMA chart for Phase II process monitoring. Calculating the Hotelling's T^2 statistics on low-dimensional sub-vectors and combining them is the fundamental concept. A simulation study is designed to

assess the performances of the proposed High-Dimensional EWMA (HDEWMA) control chart (paired HDEWMA (PHD-EWMA)) and its enhanced version, HDEWMA* (PHD-EWMA*), with some counterparts of the IEWMA and MEWMA procedures. The proposed methods performed well on the real-time traffic data, and the results confirmed that PHD-EWMA performed better than PHD-EWMA*. Kim, et al. [2] suggested a new method that used L2 norm regularization to "shrink" all mean process estimates towards zero based on penalized likelihood. This method, Ridge-based Multivariate SPC (RMSPC), was extended to the Adaptive RMSPC, the EWMA version of RMSPC (RMEWMA), and VS-based RMSPC/RMEWMA charts. They compared the performance of the RMSPC chart with VSMSPC and Hotelling's T^2 for different mean-shift scenarios using the out-of-control Average Run Length (ARL) metric. They also tested the proposed method in a high-dimensional context with many features ($p = 1000$ and $p = 5000$) and compared the ARL performance of Adaptive RMEWMA (ARMEWMA) with those of MEWMA, LEWMA, and VSMEWMA. The proposed charts represent a case study of the milling process for assembling complex parts in several devices, focusing on the detection of small mean shifts in high-dimensional processes with strongly correlated quality measurements. The proposed approach was found to be effective in online monitoring and detecting small mean changes in high-dimensional processes compared with conventional multivariate SPC and recent VS-based charts. Ahmad, et al. [85] proposed a MEWMA chart based on a Fast Region-based Convolutional Network method. The proposed algorithm offers several advantages over traditional techniques for process monitoring. For example, feature extraction is not required when high-dimensional data are available, because the networks can be fed directly from raw images. Furthermore, it has proven to

significantly improve fault detection accuracy, without the algorithm requiring many images to train the network.

Li, et al. [9] proposed a new control chart called HAMEWMA, which is a nonparametric control chart based on the Hamiltonian path. This study focuses on contrasting HAMEWMA with three types of self-starting control charts: the SSEWMA [97], SREWMA [98] and DFEWMA [99]. The findings revealed that the HAMEWMA control chart showed better resilience and monitoring efficiency than conventional multivariate nonparametric control charts. Fang, et al. [60] employed Water Vapor Flux Divergence (WVFD) data to demonstrate the Data-Driven Pairwise EWMA (DPEWMA) chart for monitoring extreme weather events. The proposed method utilizes a powerful pairwise method using the EWMA to exploit the informative pairwise correlation between any two dimensions of a dataset. In contrast to conventional multivariate methods for monitoring binary data, such as NEWMA and MEWMA, the proposed approach is easy to implement and highly efficient in monitoring High-Dimensional Binary Data Streams (HBDS). Kim, et al. [46] introduced a multivariate Bayesian EWMA (BEWMA) chart and compared it with existing charts, such as MEWMA and VSMEWMA. Different comparisons were made, and the results showed that BEWMA outperformed MEWMA and VSMEWMA in all shift scenarios with fewer sparse setups. The effectiveness of the proposed approach was further demonstrated through a case study involving a continuously stirred tank heater. Maboudou-Tchao, et al. [51] proposed Least Squares Support Vector Data Description (LS-SVDD) and Support Vector Data Description (SVDD)-based charts called LSMEWMA and SMEWMA charts. The findings reveal that the proposed methods are better than the MEWMA [74], RMEWMA [2] and L_0 MEWMA [100] charts.

2.3.3 The MCUSUM Charts

Liu, et al. [84] proposed a CUSUM chart based on the sum of TRAS (Top-r-based Adaptive Sampling) statistics for online process change detection. The proposed scheme is scalable and can detect a variety of mean shifts in all directions, as long as each data stream follows a univariate normal distribution. Zou, et al. [22] proposed a new control chart based on a powerful goodness-of-fit (GOF) test by integrating information from multiple data streams using the CUSUM scheme. A simulation study and case study on the semiconductor manufacturing process demonstrated that the proposed method is superior to other competitive techniques. Furthermore, multiple studies have been designed based on a single-stage technique to detect faults in high-dimensional data streams, which has limitations in finding solutions to high-dimensional data monitoring problems. Therefore, Li [3] proposed a two-stage CUSUM procedure in which the first stage involves the IC ARL determining the control limit, which regulates how frequently false alarms are expected when the system is in the IC. In the second stage, the Per-Comparison Error Rate (PCER) determines the control limit and regulates the number of false alarms that can be accepted while identifying the OC data streams. This enables users to develop personalized monitoring schemes based on how frequently they want to expect false alarms when the system is in the IC and how many false alarms they can tolerate when identifying OC data streams. The proposed method performed better than the LT procedure proposed by [101, 102]. A simulation study and case study on a real dataset are presented to demonstrate that the two-stage CUSUM procedure maintains control of the global PCER in all settings. Moala, et al. [65] discuss the use of both MCUSUM and MEWMA control charts to address a high-dimensional Spatial–Temporal Autoregressive Moving Average (STARMA)

process. They used a five-parameter parsimonious STARMA model to simulate the monthly rates of vehicle robberies and obtained control limits that met the desired level of ARL. This study compares the performance of both control charts in terms of ARL and finds that they have similar performance, although MCUSUM is slightly better in some situations.

Li [86] utilized the order statistics of local monitoring statistics to propose a new category of global monitoring statistics (G_t) using the Cumulative Sum (CUSUM) statistic. The proposed monitoring statistics can effectively detect various abnormal situations owing to the unique application of order statistics. To compare the efficacy of the proposed monitoring statistic (G_t), a simulation study was conducted using the global monitoring statistics proposed by Zou, et al. [22] (G_t^Z) and a thresholding-based global monitoring statistic (G_t^L). The computational time of the proposed monitoring scheme based on G_t is similar to that of the G_t^L -based monitoring methods. Conversely, the computational time of the monitoring scheme based on G_t^Z is nearly ten times the computational time of the proposed monitoring scheme. Ye, et al. [29] recently introduced a nonparametric asynchronous monitoring (NAM) framework for monitoring high-dimensional asynchronous and heterogeneous data streams online. The proposed framework is unique because it allows for different sampling intervals between data streams and arbitrary distributions for measurements, making it more flexible than previous studies that relied on restrictive assumptions, such as synchronous measurements and parametric distributions. Local Cumulative Sum (CUSUM) statistics based on quantile-based ordered interval indicators are constructed to generate a single global monitoring statistic by combining the sum of the top- r local CUSUM statistics. The proposed NAM algorithm was

evaluated based on two metrics, maximum ATS and average ATS, and was compared with the same NAM framework but with two different compensation strategies (data-forward and statistic-forward) and the MRS-KNN method to demonstrate the superiority of their proposed methodology. To illustrate the real-world applicability of the NAM framework, two publicly available datasets, the semiconductor manufacturing process and TEP, were used, and it was shown that the proposed NAM algorithm outperformed the benchmark methods in detecting quick shifts.

2.3.4 Other Charts

In SPC literature, many studies have been conducted using neither the dimension-reduction/variable-selection method nor conventional structures such as Hotelling's T^2 , MEWMA, or MCUSUM control charts. For example, Zhang and Albin [70] proposed the χ^2 control chart method, which is effective in detecting outliers while preserving most non-outlier profiles. The authors conducted both a simulation and case study analysis using the VDP dataset to evaluate the performance of the proposed chart, using the average percentage (and standard deviation) as the evaluation measure, and found satisfactory results in detecting faults. Hu and Yuan [16] proposed a method for monitoring batch processes using tensor factorization and the TLPP. Unlike existing methods, such as multiway principal component analysis (MPCA) and MLPP, which represent batch data as a vector in high-dimensional space, TLPP avoids information loss by not vectorizing the data. The effectiveness of the TLPP was tested using benchmark fed-batch penicillin fermentation process data and two industrial fed-batch processes of penicillin and cephalosporin fermentation, and compared with MPCA and MLPP monitoring techniques. The results were shown on SPE charts, and it was found that the TLPP detected faults

earlier than the MLPP and MPCA. In addition, the results for different numbers of reference batches in the benchmark fed-batch penicillin fermentation examples demonstrated that TLPP had a much lower false alarm rate than MLPP and MPCA, especially with smaller reference datasets. Chen and Nemhard [71] proposed an HD control chart for monitoring profiles. Their approach relies on the Adaptive Neyman (AN) test for coefficients and the Discrete Fourier Transform (DFT) algorithm. The chart can track profiles without necessitating the fitting of regression models, and is effective for monitoring both linear and nonlinear profiles with autocorrelated stationary noise.

Yu [57] presented a method for identifying and diagnosing faults in chemical processes by using a Nonlinear Kernel Gaussian Mixture Model (NKGMM) based chart. This approach was used to monitor a simulated wastewater treatment process with multiple modes of operation, and its performance was compared to GMM and Independent Component Analysis (ICA) based charts. Two erroneous scenarios, step change and drift error, were tested, and the results showed that the NKGMM method excelled in detecting and diagnosing faults in the simulated multimode wastewater treatment process. Li, et al. [87] presented a high-dimensional change-point detection method called the HC chart to detect changes in high-dimensional data. Their approach can detect early shifts in measurements regardless of dimensionality and reduces the average run length. Based on the ARL and diagnostic accuracy, the proposed HC chart is superior to the ZH control chart [103].

Zou, et al. [21] introduced an approach to detect outliers in a baseline dataset. They used penalized regression to analyze high-dimensional profile vectors, and transformed the model into a specialized regression model. To detect abnormal profile observations, two Penalized Profile Outlier detection (PPOD) algorithms, PPOD-R and PPOD-C, were

proposed. The performance of PPOD-R was evaluated by comparing the average percentage of Type-I and Type-II errors with the χ^2 chart. Additionally, the study presented a comparative analysis of the performances of PPOD-R and PPOD-C under different correlated scenarios. The proposed method was evaluated using simulations and real-world case studies on an industrial etching process for semiconductor manufacturing. The results demonstrate that the proposed approach performs exceptionally well in identifying outliers across various situations. Kan and Yang [11] presented a dynamic network methodology for the profile monitoring of high-dimensional imaging streams. Using experimental data related to biomanufacturing and machining imaging profiles, the proposed approach captured complicated image profiles and offered an effective online control chart for monitoring them. Turkoz, et al. [88] proposed a new method for identifying fault variables that does not rely on any specific distribution of observations. The paper aimed to address the drawbacks of existing approaches, such as the impact of the correlation between variables, computational complexities in high-dimensional data, and the number of changed variables. The proposed Distribution Free Adaptive Step-Down (DFASD) method for fault identification combines the SVDD-based test statistic with an ASD approach. The performance of the proposed approach was evaluated and compared with that of other methods, such as ASD, K^2 decomposition, and HNS decomposition, using both normal and non-normal datasets.

Jang, et al. [62] proposed three real-time contrast-based control charts (FWRF, GWRF, and MWRF) for the variance monitoring of both normal and non-normal high-dimensional data. Using ARL as an evaluation measure, the results revealed that FWRF, MWRF, and GWRF worked well with low-dimensional data, high-dimensional data, and bivariate t-and

gamma distributions, respectively. Wang, et al. [17] presented a survey of data visualization techniques for detecting faults in the chemical industry and introduced a framework called time-explicit Kiviat diagrams, which are multivariate plots with time dimensions. The study reviewed the development of this type of diagram and examined its application in three types of chemical processes: continuous, batch, and periodic. Finally, they adapted traditional fault detection methods for periodic processes, including Dynamic PCA (DPCA) and Multiway PCA (MPCA), and found that the time-explicit Kiviat diagram approach performed better in most cases.

Shu and Fan [23] developed an Interpoint Distance-based control chart (IPD chart) to track mean vectors in high-dimensional environments. Monte Carlo simulations were used to evaluate the performance of the chart, which was then compared with the CQ, PA, and WPL charts, which were based on the T_{CQ} , T_{PA} , and T_{WPL} tests, respectively. Finally, the proposed IPD chart is implemented on a dataset from a semiconductor manufacturing process. Yan, et al. [44] proposed a monitoring method based on the sequential likelihood ratio of a Spatio-Temporal Smooth Sparse Decomposition (ST-SSD) test to monitor the process. The study began by testing the capability of ST-SSD in estimating the functional mean and showed the estimated functional mean using Reproducing Kernel (RK) and Roughness Minimization (RM) models. The performance of the proposed approach was compared with those of Hotelling's T^2 , LASSO, and CUSUM methods. It was revealed that the RK and RM models performed better than other benchmarks because RK is more precise in modeling the temporal trend, whereas RM has a slower computation time owing to its higher modeling complexity. Finally, the proposed methods were demonstrated in three real-world case studies, including steel rolling, solar activity, and tonnage signal

monitoring, wherein they effectively identified the time and location of process changes and detected anomalies. He, et al. [63] introduced a distance-based multivariate process control chart wherein Support Vector Machines (SVM) were utilized, known as the D-SVM chart. The proposed chart's performance is compared with the random forest-based Real-Time Contrasts (RTC-RF) and One-Class SVM (OC-SVM) charts, using the ARL for shifts in both the mean and variance. The study's results indicated that the performance of the D-SVM chart is similar to that of the RTC-RF chart when only shifts occur in the mean of the process. However, in the case of shifts in both the mean and variance, the D-SVM chart outperformed the RTC-RF chart. In addition, a case study of a camera performance measurement dataset collected from a mobile phone manufacturing process was conducted to demonstrate how the proposed D-SVM chart worked.

Guo, et al. [89] proposed a method based on the RTC technique to detect defects in industrial image data by considering both the spatial and temporal aspects. The Steady-State Median Run Length (SSMRL) was used to evaluate the detection capability of their approach and calculate the Relative Mean Index (RMI) to assess the overall performance of the control chart for shifts with varying magnitudes. Finally, the results showed that the proposed method not only accurately estimated the occurrence of defects, but also identified their location and size. Abdella, et al. [24] proposed an Adaptive LASSO Thresholding-norm chart (ALT-norm chart) to detect various types of shifts. The proposed approach can accurately detect shifts of different magnitudes and patterns. Unlike other monitoring techniques that rely on shift-pattern information to construct control charts, the design of an ALT norm chart is not affected by such information gaps. Kim, et al. [7] introduced a new control chart, the RPLR chart, based on the ridge-penalized likelihood

ratio, to monitor changes in the covariance matrix. The effectiveness of the RPLR chart was first evaluated based on the ARL, and then compared with the PLR chart developed by Li, et al. [104]. The RPLR chart was tested on simulated and industrial data, and it was found that its design was not affected by the lack of information about the shift pattern, unlike the existing PLR and LMEWMC charts.

Turkoz, et al. [38] introduced a computationally efficient Bayesian Support Vector Data Description (BSVDD) approach for identifying faults that overcome the limitations imposed by the normality assumption. The performance of BSVDD and its comparison with T^2 decomposition with the ASD procedure, K^2 (the HNS decomposition), and SVDD-based procedure, are evaluated in terms of EER and Fault Correctness Ratio (FCR) measures. Experiments with various datasets showed that BSVDD is more robust for detecting faults in non-normal data, particularly those with irregular patterns, and is thus useful in practice when the distribution is unknown. Li, et al. [25] presented a quick computational method for diagnosing High-dimensional Data Streams (HDS) by framing the problem as a multiple testing one and controlling the weighted missed discovery rate (wMDR) at a satisfactory level. The marginal data-driven wMDR-control procedure (MDW) was proposed and compared with the marginal oracle wMDR-control procedure (MOW) and LASSO-based diagnostic procedure (LEB). The study demonstrated the effectiveness of the proposed method through simulations and semiconductor manufacturing process data. Martinez, et al. [67] introduced the One-Class Peeling (OCP) method, a versatile approach that combines statistical and machine learning techniques to identify multiple outliers in multivariate data. The OCP can be implemented in a Phase I estimation without requiring the estimation of covariance, and it is particularly appropriate

for high-dimensional datasets with a substantial fraction of outliers. To demonstrate the efficiency of the OCP method, it was applied to webpage hit data of the National Football League (NFL) on Wikipedia. Furthermore, a simulation study is designed, and Type I error percentage is used to compare the performance of OCP versus the Robust Minimum Diagonal Product (RMDP) method. In most cases, the classification accuracy and outlier detection rate of the OCP method exceed those of the RMDP method. Abdella, et al. [8] assessed the effectiveness of the proposed T-COV chart by examining its signal probability and comparing it with the Conditional Entropy (CE) control chart. The T-COV was implemented on spur gear production data and simulation data. The results demonstrate that the T-COV control chart performs better than the CE control chart under different shift levels in the covariance matrix of high-dimensional processes.

Maboudou-Tchao [53] applied 2nd-order tensors to high-dimensional data and employed the Support Tensor Vector Data Description (STVDD) chart. The effectiveness of the proposed chart was evaluated against that of the SVDD chart, using the ARL, Relative Average Run Length (RARL), and Extra Quadratic Loss (EQL) measures. The performance of the chart was tested on two real datasets, and the results indicated that the proposed chart is a viable option for effectively monitoring high-dimensional data. In addition, the proposed chart is distribution-free and useful when the in-control parameters are unknown. Fan, et al. [50] introduced a nonparametric Phase I control chart called the SLED chart, which is based on the change-point formulation and tests for a change in the leading eigenvalue of the difference between the two covariance matrices. The proposed SLED chart was applied to a diagnostic breast cancer dataset, and the results showed that the SLED chart had greater detection power than the existing CLX and LC control charts.

Kim and Turkoz [32] proposed a Bayesian Sequential Update (BSU) chart and compared it with the MEWMA, VS-MEWMA, and Kalman update-based charts based on the ARL. The BSU chart detected a small shift faster than the other methods. Merlo, et al. [36] introduced a distribution-free method by combining the Mahalanobis distances with the Mann-Whitney statistic to create either MW(D) or MW(I) control charts, depending on the chosen monitoring scheme. This approach is particularly useful for high-dimensional processes, in which distance measures are ranked and compared using two-sample test statistics. As the MW(D) chart showed the best overall performance, it was compared with the HDSOR-W control chart (combined version of origin (OR) in the Wilcoxon statistic (W)), using the median run length and Conditional Average Run Length (CARL) as evaluation measures. Xiang, et al. [28] discussed the challenge of accurately detecting faults in high-dimensional correlated data streams by using large-scale multiple testing. A data-driven LIS-based diagnostic procedure was proposed under a Hidden Markov Model (HMM) dependence framework that can control Missed Discovery Exceedance (MDX) while ensuring high reliability and efficiency. The proposed method was compared with existing diagnostic approaches and was found to be effective in detecting abnormal mean shifts. For joint monitoring of location and scale parameters of the multivariate and high-dimensional process, Mukherjee and Marozzi [27] suggested Shewhart-type nonparametric control charts without requiring the normality assumption. The median run length was used as a measure of performance, and simulations and real data from the semiconductor manufacturing process were used to demonstrate the performance of the proposed method.

2.4 Research Questions

After conducting a comprehensive review of the high-dimensional control charts, it has been noticed that all the control charts in the literature focused on one-sample high-dimensional data. Hence, the following research questions are raised:

- How to construct control charts for two-sample high-dimensional data?
- How to develop memoryless monitoring structures (Shewhart environment) that can capture large shifts in two-sample high-dimensional data? Will the new charts be able to overcome the limitation of conventional Hotelling's T^2 chart for higher dimensions?
- What advantages can be gained by constructing two-sample high-dimensional memory-based monitoring schemes based on memory setup, such as Exponentially Weighted Moving Average (EWMA) and Cumulative Sum (CUSUM), over the memoryless Shewhart environment?
- How to implement memoryless and memory-based high-dimensional control charts for two-sample cases in real-life data?
- Will the two-sample high-dimensional control chart under CUSUM setup outperform the two-sample high-dimensional chart under EWMA setup?

This thesis aims to answer the above questions using both theoretical and computational activities. R statistical package is used to serve the purpose of simulation and case study.

CHAPTER 3

HIGH-DIMENSIONAL CONTROL CHARTS UNDER SHEWHART ENVIRONMENT

High-dimensional data refers to data with more features or variables than the number of observations. For instance, in a manufacturing process, the number of process variables, features, or dimensions, such as pressure, temperature, chemical concentrations, etc., can be notably larger than the produced chips or observations. Multivariate control charts are crucial in monitoring and detecting anomalies in industrial processes dealing with two-sample data. Traditional Hotelling's T^2 chart provides satisfactory performance as long as the dimension of the two-sample data remains lower. However, high-dimensional data with two-sample is obvious in manufacturing processes. Therefore, the conventional chart fails to detect variation in the process mean when the data dimension increases. This study aims to propose new control charts that can perform with superiority to detect variations in the mean for both low and high-dimensional data. In this regard, DR, BS, and SD charts are proposed. Both simulation and real-life data are used to verify the better detection ability of the charts. Different performance measures are used to study the run length profile. The performance of the charts is compared to each other under different multivariate distributional setups, such as under multivariate normal, multivariate t , and multivariate gamma. DR and BS chart perform the same; however, the DR chart performs better than the SD chart for low-dimensional data under multivariate normal distribution. On the

contrary, when the dimension increases, the ability to shift detection is found better with the SD chart under multivariate t and gamma distributions. Further, to highlight the importance of the stated proposal, proposed charts are implemented to surveillance grease damage in wind turbines' bearings.

The remaining part of the chapter is as follows. In Section 3.1, the two-sample Hotelling's T^2 chart is introduced, and the proposed charts are developed in Section 3.2. Then the Section 3.3 describes the designs required for the simulation study. In Section 3.4, the performance of the proposed charts for both lower and higher-dimensional data is presented. At the end, a case study is provided in Section 3.5.

3.1 Two-Sample Hotelling's T^2 Chart

For two-sample case (k refers to sample number, m represents sample size of first sample, n represents sample size of second sample, and p is used to determine dimensions), let x_{ijk} , $i = 1, 2, \dots, m$ or $i = 1, 2, \dots, n$, $j = 1, 2, \dots, p$, and $k = 1, 2$ are the p -dimensional multivariate normal random vectors with mean vector $\mu = (\mu_1, \dots, \mu_p)'$ and unknown common positive variance-covariance matrix Σ . The two-sample Hotelling's T^2 statistic is computed as follows:

$$T^2 = (\bar{x}_1 - \bar{x}_2)' \left\{ S_p \left(\frac{1}{m} + \frac{1}{n} \right) \right\}^{-1} (\bar{x}_1 - \bar{x}_2) \quad (3.1)$$

Where S_p is the pooled variance-covariance matrix. The upper control limit is given by:

$$UCL = \frac{p(m+n-2)}{m+n-p-1} F_{\alpha, p, (m+n-p-1)}, \text{ where } F_{\alpha, p, (m+n-p-1)} \text{ is the upper } \alpha^{th} \text{ percentile of}$$

the F distribution with parameters p and $(m + n - p - 1)$. The lower control limit is zero, and the chart detects a signal when T^2 exceeds the UCL .

3.2 Proposed Two-Sample High-dimensional Charts

For two-sample case (k refers to sample number, m represents sample size of first sample, n represents sample size of second sample, and p is used to determine dimensions), let x_{ijk} , $i = 1, 2, \dots, m$ or $i = 1, 2, \dots, n$, $j = 1, 2, \dots, p$, and $k = 1, 2$ are the p -dimensional multivariate normal random vectors with mean vector $\mu = (\mu_1, \dots, \mu_p)'$ and unknown common positive variance-covariance matrix Σ . In the two-sample case, when the number of variables is more than observations, the two-sample statistic T_1 [105] can be written as:

$$T_1 = \frac{\frac{mn}{m+n}(\bar{x}_1 - \bar{x}_2)' D_S^{-1} (\bar{x}_1 - \bar{x}_2) - \frac{(m+n)p}{m+n-2}}{\left[2 \left(\text{tr } R^2 - \frac{p^2}{m+n} \right) c_{p,m+n} \right]^{\frac{1}{2}}}, \quad (3.2)$$

where the diagonal matrix (D_S) of sample variances (S) is defined by

$$D_S = \text{diag}(s_{11}, \dots, s_{pp}), \quad (3.3)$$

The sample correlation matrix (R) is defined as follows:

$$R = D_S^{-\frac{1}{2}} S D_S^{-\frac{1}{2}}, \quad (3.4)$$

The adjustment coefficient $c_{p,m+n}$ can be chosen by

$$c_{p,m+n} = 1 + \frac{\text{tr } R^2}{p^{3/2}}. \quad (3.5)$$

The lower control limit (LCL) and upper control limit (UCL) of the statistic T_1 for a given false alarm α , can be obtained from the following probability limits:

$$P(T_1 \leq LCL) = \int_{t_1=0}^{LCL} P(T_1 = t_1) \leq \frac{\alpha}{2}, \quad (3.6)$$

$$P(T_1 \geq UCL) = \int_{t_1=UCL}^{\infty} P(T_1 = t_1) \leq \frac{\alpha}{2}. \quad (3.7)$$

It is to be noted that the test statistic T_1 was proposed by Srivastava and Du [105].

Therefore, the control chart based on statistic T_1 is named as SD chart.

The DR chart is proposed based on the statistic T_2 [106, 107], which can be expressed as follows:

$$T_2 = \frac{\frac{mn}{m+n} (\bar{x}_1 - \bar{x}_2)' (\bar{x}_1 - \bar{x}_2)}{\text{tr } S}. \quad (3.8)$$

The lower control limit (LCL) and upper control limit (UCL) of the test statistic T_2 for a given false alarm α , can be obtained from the following probability limits:

$$P(T_2 \leq LCL) = \int_{t_2=0}^{LCL} P(T_2 = t_2) \leq \frac{\alpha}{2}, \quad (3.9)$$

$$P(T_2 \geq UCL) = \int_{t_2=UCL}^{\infty} P(T_2 = t_2) \leq \frac{\alpha}{2}. \quad (3.10)$$

Further, the BS chart is proposed based on statistic T_3 [108], which is given by:

$$T_3 = \frac{\frac{mn}{m+n} (\bar{x}_1 - \bar{x}_2)' (\bar{x}_1 - \bar{x}_2) - \text{tr } S}{\left[\frac{2(m+n)(m+n+1)}{(m+n-1)(m+n+2)} \left(\text{tr } S^2 - \frac{(\text{tr } S)^2}{m+n} \right) \right]^{\frac{1}{2}}}. \quad (3.11)$$

The lower control limit (LCL) and upper control limit (UCL) of the BS chart for a given false alarm α , can be obtained from the following probability limits:

$$P(T_3 \leq LCL) = \int_{t_3=0}^{LCL} P(T_3 = t_3) \leq \frac{\alpha}{2}, \quad (3.12)$$

$$P(T_3 \geq UCL) = \int_{t_3=UCL}^{\infty} P(T_3 = t_3) \leq \frac{\alpha}{2}. \quad (3.13)$$

It is to be noted that all proposed three charts, named SD, DR, and BS charts signal an out-of-control (OOC) point when T_1 , T_2 , and T_3 falls outside of their control limits, respectively. Otherwise, the process is considered to be in a stable or in-control (IC) situation. The statistics T_1 , T_2 , and T_3 are computed using “SHT” R package version 0.1.8 [109]. The package can handle to biasedness of the estimators during the computation stage of the statistics.

3.3 Design of Simulation Study

Extensive simulations have been carried out using 1000000 iterations to obtain the control limits under different distributions and dependency structures. Following Stoumbos and Sullivan [110], the following distributions are considered: (i) multivariate normal with unit variances and covariances equal to 0, denoted as MVN_0 ; (ii) multivariate normal with unit variances and covariances equal to 0.5, denoted as MVN_1 ; (iii) multivariate t with 5 degrees of freedom, denoted as MVT ; (iv) multivariate gamma with shape parameter 5 and scale parameter 1, denoted as MVG. For both MVT and MVG, unit variances and covariances are considered equal to 0.5. For lower dimensions such as $p = 10, 30$, and 50, equal sample size ($m = 30, n = 30$) is considered for both samples. With pre-specified ARL_0 such as 168, 200, 370, and 500, Table 3 presents the control limits for the proposed DR, BS, and SD charts for each of four cases, such as MVN_0 , MVN_1 , MVT, and MVG. To assess the performance of the proposed charts with Hotelling’s T^2 chart under different shifts ($\delta = 0$,

0.1, 0.2, 0.3, 0.4, 0.5, 0.75, 1) in the second sample observations, a simulation study with 10,000 iterations, pre-specified $ARL_0 = 200$, and under four different distribution cases is designed. The comparative findings are presented in Table 4.

For higher dimensions such as $p = 100, 300, \text{ and } 500$, the sample size for two samples is considered as (a) $m = p \times 30\%, n = p \times 30\%$ and (b) $m = p \times 30\%, n = p \times 45\%$. A simulation study of 1000000 iterations has been performed to provide control limits under the distribution cases (i) – (iv) for two combinations of sample size (a) – (b). Further, on the same pre-fixed ARL_0 choices that are used for lower dimensions, the control limits of the proposed charts for the higher dimensions under four different distribution cases are presented in Table 5. Finally, the comparative analysis of the proposed charts based on the run length profile with 10,000 simulations, $ARL_0 = 200$, shifts ($\delta = 0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.75, 1$), sample sizes (a) – (b), and distribution cases (i) – (iv) have been provided in Table 6.

It is to be noted that the run length profile of Hotelling's T^2 chart for higher dimensions is not possible due to the singularity problem of the variance-covariance matrix, S_p in Equation (3.1). Further, Bai and Saranadasa [108] showed that the T_3 and T_1 statistics have the similar performance in testing the means of two-sample high-dimensional data. Similarly, we found in initial findings that the BS chart has the same run length profile as the DR chart, also seen in the case study presented in Section 3.5. Therefore, for brevity, we have excluded the results of the BS chart.

3.4 Results and Discussions

In this study, Average Run Length (ARL) and Standard Deviation of Run Length (SDRL) along with Q_1 (1st quartile), Q_2 (2nd quartile or median), and Q_3 (3rd quartile) of run length are considered as performance measures. Average Run Length (ARL) is interpreted as the average number of points until a point is out-of-control. Generally, ARL is categorized as in-control Average Run Length (ARL_0) and out-of-control Average Run Length (ARL_1). ARL_0 represents the average number of points until a point is out-of-control when the process is stable while ARL_1 indicates the average number of points until a point is out-of-control when the process is unstable or shifted. The Standard Deviation of Run Length (SDRL) measures the variability or dispersion of run lengths. It can evaluate the performance of the charts by measuring the consistency and precision of detecting out-of-control conditions. A larger SDRL indicates more variation in the run lengths, while a smaller SDRL suggests a reliable control chart. The quartiles Q_1 , Q_2 , and Q_3 provide the 25th, 50th, and 75th percentiles of run length, respectively. Although ARL, SDRL, Q_1 , Q_2 , and Q_3 are reported as the evaluation measures of the charts; for the sake of brevity, only ARL will be considered to decide about the charts' performance. With a pre-defined ARL_0 , a smaller ARL_1 is desired that indicates more suitable control chart to detect mean shift in the process monitoring.

3.4.1 Performance of the Proposed Charts under Lower Dimensions

Based on the run length profile, the performance comparison between the proposed DR and SD charts, and the existing two-sample Hotelling's T^2 chart under lower dimensions is described as follows:

- The control limits for the proposed charts for $ARL_0 = 168, 200, 370, \text{ and } 500$ are presented in Table 3. However, only $ARL_0 = 200$ will be used to conduct comparative analysis of the charts.
- In Table 4, under the distribution structure MVN_0 , for $p = 10$, Hotelling's T^2 chart has shown better detection ability as its values of ARL_1 for different shifts are least than the proposed DR and SD charts. However, for $p = 30$, our proposed DR chart performed the best for different shifts. For $p = 50$, the values of ARL_1 of Hotelling's T^2 chart are smallest for small shifts such as $\delta = 0.1, 0.2, \text{ and } 0.3$, however for large shifts such as $\delta = 0.4, 0.5, 0.75, \text{ and } 1$, our proposed DR chart has shown the best detection ability. The smallest values of ARL_1 are shown in bold font in Table 4.
- Under the distribution structure MVN_1 , the detection ability of the proposed DR chart for both $p = 10$ and 30 is better for different shift values as compared to the other charts, except shift $\delta = 0.1$ where the ARL_1 values of Hotelling's T^2 , DR, and SD charts are reported as 197.7, 198.7, 201.3 respectively (cf. Table 4). However, for $p = 50$, $\delta = 0.2$, Hotelling's T^2 chart gives lowest ARL_1 (193.7); for the same dimension with $\delta = 1$, the proposed DR chart obtains the smallest ARL_1 (45.7) than the other charts (cf. Table 4). For all other shifts with dimension $p = 50$, our proposed SD chart provides the smallest values of ARL_1 compared to other charts.
- With the distribution structure MVT, only for $p = 10, \delta = 0.1$; $p = 30, \delta = 0.1, 0.2$, Hotelling's T^2 chart presents small ARL_1 values (cf. Table 4). For all the other shifts and dimensions, $p = 10 \text{ and } 30$, the proposed DR chart has shown better detection ability. For the dimension $p = 50$, the proposed SD chart reports the smallest ARL_1 values (cf. Table 4) showcasing its better performance than the other charts.

- The performance of our proposed DR chart is the best for $p = 10$ and 30 under the distribution MVG for all shifts except $\delta = 0.1$, $p = 30$ where proposed SD chart gives small ARL_1 value (193.9) (cf. Table 4). This proposed SD chart also shows the best detection ability against all shifts for the dimension $p = 50$ as it provides the smallest values of ARL_1 than the other charts (cf. Table 4).

Table 3: Control limits of the proposed charts under lower dimensions with $m = 30$ and $n = 30$

p	Distribution	ARL_0	T^2		DR		BS		SD	
			LCL	UCL	LCL	UCL	LCL	UCL	LCL	UCL
10	MVN_0	168	0.0000	34.6147	0.1893	2.7104	-1.8057	3.8034	-1.6325	3.5429
		200	0.0000	35.4851	0.1813	2.7601	-1.8236	3.9178	-1.6493	3.6392
		370	0.0000	38.5743	0.1571	2.9337	-1.8756	4.2997	-1.6966	3.9984
		500	0.0000	40.0987	0.1469	3.0289	-1.9000	4.5052	-1.7175	4.1809
	MVN_1	168	0.0000	34.6147	0.1517	3.5181	-1.5198	4.5390	-1.3013	3.9109
		200	0.0000	35.4851	0.1452	3.6130	-1.5325	4.7019	-1.3127	4.0554
		370	0.0000	38.5743	0.1248	3.9287	-1.5720	5.2864	-1.3467	4.5577
		500	0.0000	40.0987	0.1168	4.0773	-1.5889	5.5599	-1.3637	4.7885
	MVT	168	2.2520	36.1876	0.1530	3.4606	-1.4646	4.2174	-1.2591	3.5795
		200	2.1611	36.9421	0.1463	3.5441	-1.4783	4.3769	-1.2716	3.7143
		370	1.8737	39.7212	0.1268	3.8413	-1.5222	4.9097	-1.3094	4.1819
		500	1.7592	41.0688	0.1182	3.9931	-1.5417	5.1770	-1.3261	4.4038
	MVG	168	2.2386	36.1468	0.1586	3.3116	-1.5119	4.1550	-1.3326	3.5850
		200	2.1449	36.9390	0.1520	3.3928	-1.5253	4.2980	-1.3445	3.7032
		370	1.8617	39.6374	0.1310	3.6739	-1.5674	4.8079	-1.3833	4.1599
		500	1.7319	40.8993	0.1224	3.8081	-1.5861	5.0561	-1.4022	4.3506
30	MVN_0	168	0.0000	155.8179	0.4335	1.8681	-2.1818	3.3361	-2.0530	3.2062
		200	0.0000	159.6027	0.4261	1.8887	-2.2119	3.4199	-2.0818	3.2882
		370	0.0000	173.1906	0.3987	1.9638	-2.3147	3.7094	-2.1778	3.5724
		500	0.0000	179.9851	0.3856	2.0028	-2.3644	3.8598	-2.2207	3.7118
	MVN_1	168	0.0000	155.8179	0.3640	2.2594	-1.9148	3.8164	-1.7340	3.4862
		200	0.0000	159.6027	0.3567	2.2997	-1.9394	3.9391	-1.7558	3.5988
		370	0.0000	173.1906	0.3312	2.4302	-2.0194	4.3428	-1.8240	3.9693
		500	0.0000	179.9851	0.3189	2.4933	-2.0541	4.5305	-1.8575	4.1436
	MVT	168	22.4132	162.9715	0.3644	2.2345	-1.7228	3.3413	-1.5515	2.9934
		200	21.9699	166.4957	0.3571	2.2706	-1.7455	3.4361	-1.5733	3.0867
		370	20.4900	178.9737	0.3304	2.4017	-1.8185	3.7892	-1.6381	3.4123
		500	19.8021	185.8269	0.3202	2.4641	-1.8513	3.9524	-1.6675	3.5615
	MVG	168	21.8473	165.8428	0.3774	2.1809	-1.8618	3.5422	-1.7356	3.2251
		200	21.3931	169.3532	0.3696	2.2130	-1.8859	3.6444	-1.7569	3.3167
		370	19.8307	182.4737	0.3437	2.3355	-1.9658	4.0269	-1.8297	3.6454
		500	19.1683	188.9602	0.3312	2.3979	-2.0037	4.1937	-1.8659	3.8008
50	MVN_0	168	0.0000	1677.1218	0.5342	1.6483	-2.3164	3.2148	-2.1962	3.0964
		200	0.0000	1757.3746	0.5264	1.6644	-2.3533	3.2997	-2.2308	3.1786
		370	0.0000	2066.0483	0.5022	1.7212	-2.4750	3.5805	-2.3479	3.4582
		500	0.0000	2232.8894	0.4902	1.7450	-2.5347	3.7013	-2.3970	3.5925
	MVN_1	168	0.0000	1677.1218	0.4632	1.9182	-2.0753	3.5742	-1.9055	3.3051
		200	0.0000	1757.3746	0.4559	1.9435	-2.1048	3.6753	-1.9313	3.3963
		370	0.0000	2066.0483	0.4304	2.0312	-2.2015	4.0161	-2.0186	3.7278
		500	0.0000	2232.8894	0.4188	2.0747	-2.2477	4.1853	-2.0609	3.8659
	MVT	168	100.8269	1945.7424	0.4656	1.9029	-1.7577	2.9602	-1.5992	2.6791
		200	98.5918	2035.9290	0.4583	1.9273	-1.7851	3.0446	-1.6250	2.7533
		370	91.5676	2392.5281	0.4344	2.0202	-1.8679	3.3498	-1.7017	3.0370
		500	88.3677	2573.3137	0.4237	2.0618	-1.9079	3.4933	-1.7404	3.1650
	MVG	168	98.6260	1999.2274	0.4781	1.8638	-2.0016	3.3263	-1.8862	3.0569
		200	96.4214	2095.5778	0.4704	1.8877	-2.0309	3.4200	-1.9153	3.1431
		370	88.9886	2448.5387	0.4445	1.9667	-2.1285	3.7210	-2.0098	3.4257
		500	85.7206	2641.6604	0.4339	2.0043	-2.1778	3.8779	-2.0506	3.5542

Table 4: ARL (SDRL) comparison of Hotelling's T^2 and proposed charts with m = 30, n = 30

Dist.	δ	p = 10			p = 30			p = 50		
		T^2	DR	SD	T^2	DR	SD	T^2	DR	SD
MVN ₀	0	200.3 (199.6) 58,141,279	200.5 (200.3) 58,141,276	203.6 (203.7) 59,144,280	199 (198.2) 57,137,278	196.1 (197.2) 57,136,269	197.2 (201.9) 57,134,269	197.4 (196.4) 58,136,272	203.8 (206.6) 59,142,277	203.3 (200.8) 60,141,280
	0.1	196.5 (199.3) 57,134,266	198.3 (198.7) 58,138,274	202 (201.3) 59,138,281	200.7 (200.9) 59,139,275	191.6 (190.6) 55,133,266	192.7 (194) 55,135,262	198.2 (195.3) 57,139,277	199.4 (199) 58,138,276	202.1 (202.1) 57,137,282
	0.2	196 (194) 57,137,269	201.3 (203.6) 58,141,279	207.3 (202.3) 60,147,290	195.7 (198) 56,134,270	189.6 (186.7) 56,133,263	190.6 (188) 56,133,265	191.2 (193.8) 55,129,267	199.6 (198.7) 58,139,276	199.8 (199.5) 57,137,279
	0.3	188.3 (187.4) 54,133,259	197.4 (194) 57,139,274	201.2 (200.3) 57,139,282	187.7 (188) 53,130,258	186.3 (189) 52,127,258	189.7 (190.9) 53,130,265	185 (182.3) 55,130,258	189.9 (188.7) 55,134,263	189.3 (189.5) 55,131,261
	0.4	180.7 (183.1) 50,124,249	192.6 (193.4) 55,134,266	197 (198.6) 57,138,270	182.3 (181) 54,127,253	178 (178.1) 52,125,246	182.9 (183.3) 53,127,253	176.7 (175.5) 51,122,247	169 (169.3) 49,115,235	169 (167.7) 48,118,236
	0.5	174.8 (174.1) 51,122,241	189.5 (192.3) 54,128,262	193.1 (195.6) 56,131,265	177 (178) 52,121,243	171.6 (172.4) 49,116,238	179.4 (180.4) 51,122,247	166.9 (166.2) 48,117,231	140.8 (140.7) 41,99,192	147 (148.8) 43,101,202
	0.75	156.4 (154.2) 45,109,218	170.3 (171.3) 49,118,236	180.7 (179.6) 52,126,250	164.5 (164.2) 47,114,229	142 (143.4) 42,99,194	158.9 (159.8) 46,109,218	144.9 (142.6) 42,102,201	72.3 (72) 21,51,100	78.4 (76.6) 23,55,109
	1	143.9 (143.5) 42,100,198	144.2 (145.7) 41,98,200	164.3 (163.3) 48,114,228	156.9 (156.2) 45,110,217	102.4 (102.5) 29,70,141	131.1 (131.3) 38,88,182	120.6 (118.7) 35,84,169	28.9 (29) 9,20,40	35.3 (35.2) 10,24,49
MVN ₁	0	200.3 (199.6) 58,141,279	199.5 (200.6) 58,139,273	200.8 (200.8) 58,139,280	198.3 (197.3) 58,136,279	200.6 (199.5) 58,141,278	203.7 (206.4) 59,139,278	199.5 (201.5) 56,138,276	200.5 (199.8) 59,139,279	201.5 (202.7) 58,138,280
	0.1	197.7 (200.6) 57,135,267	198.7 (199.8) 58,137,276	201.3 (204.5) 58,138,278	202.5 (203.6) 59,140,278	199.5 (199) 58,138,277	201.2 (200.9) 59,140,277	196.2 (194.1) 57,137,269	199.9 (202.5) 57,136,278	193.5 (193) 56,137,267
	0.2	199.3 (196.9) 58,139,275	196.1 (199.4) 55,135,270	197.3 (201.3) 55,134,274	200 (200.5) 57,139,276	195.7 (197.1) 56,136,270	199.3 (198) 59,139,275	193.7 (198.9) 54,133,268	195 (194.6) 56,134,273	196.6 (195.5) 57,137,272
	0.3	195.8 (194.6) 57,138,270	186.3 (187) 52,129,257	196.1 (196.4) 56,135,271	194 (195) 56,134,267	181.8 (183) 53,124,251	195.7 (197.5) 57,133,271	188.8 (189.6) 53,128,263	187 (187.2) 53,129,257	184.7 (186) 52,129,257
	0.4	188.8 (191.5) 53,130,259	177.8 (179.3) 51,124,245	192.1 (194.2) 55,132,263	191.1 (191.3) 55,132,265	171.2 (172.4) 49,118,236	187 (187.6) 53,130,260	174.1 (175.4) 50,119,241	173.2 (171.6) 51,122,240	170.5 (170.4) 49,116,237
	0.5	191 (189.9) 55,134,264	168.1 (170) 47,116,233	185.6 (187.3) 53,128,256	187.7 (186) 54,129,262	159.5 (156.6) 47,111,220	178.5 (178.3) 51,123,247	166.6 (164.8) 49,116,230	153.2 (151.9) 44,107,213	151 (148.5) 44,106,208
	0.75	177.8 (175.7) 51,123,246	139.3 (140) 41,96,193	176.8 (178.7) 50,121,245	179.4 (174.8) 53,125,251	122.2 (121) 36,85,168	152.1 (149.2) 46,107,213	140.6 (139.5) 41,98,196	93.3 (91) 27,65,130	91 (89.8) 26,64,127
	1	164.2 (162.8) 48,115,227	111.2 (111.3) 32,77,155	156.9 (158.2) 46,107,217	170.3 (170.5) 49,118,236	85.8 (84.9) 26,60,118	125.5 (125.6) 37,87,172	118.6 (116.6) 36,83,164	45.7 (44.9) 13,32,63	47.8 (47.4) 14,33,67
MVT	0	200.4 (198.7) 58,141,278	203.5 (202.3) 60,141,280	202.8 (202.3) 60,140,280	197.8 (198.2) 58,139,272	203.4 (202.6) 59,140,283	208.2 (206.1) 61,144,292	196.1 (198) 56,136,270	199.4 (201) 58,136,275	200.6 (200) 57,140,277
	0.1	201.3 (204.4) 57,137,282	206.3 (204.7) 61,144,282	206.9 (205.4) 59,145,288	198.3 (198.2) 58,138,274	200.4 (199.8) 56,139,280	206.8 (206.2) 60,142,287	201.7 (197) 61,142,279	199.6 (199.8) 58,137,278	198.9 (198.4) 58,139,275
	0.2	198.4 (196.8) 58,139,276	198.3 (194.6) 58,139,276	202 (199.8) 59,141,281	195.9 (198.3) 57,135,271	198.3 (199) 58,136,272	202.1 (199.2) 60,141,282	196.6 (196) 56,136,272	196.3 (192.3) 58,139,272	195.3 (194.7) 56,135,270
	0.3	196.8 (197.1) 58,135,271	192.1 (192) 55,133,268	196.2 (196.1) 57,136,273	201.1 (199.3) 57,142,280	194.5 (189.8) 57,138,272	202.4 (200.2) 58,140,280	202.9 (207.4) 58,139,280	191.7 (190) 55,133,269	189.2 (190.6) 55,131,260
	0.4	198.4 (199.3) 58,138,271	187.7 (185.9) 54,131,260	193.7 (190) 57,135,269	196.7 (197) 57,135,273	184.7 (183) 54,128,255	192.7 (193.8) 54,135,267	204.9 (202.3) 59,141,288	183.3 (183.1) 53,128,255	179 (180.9) 52,123,246
	0.5	194 (191.7) 57,136,267	180.5 (180.3) 51,126,249	193.4 (193) 56,134,266	196 (196.8) 56,135,272	175.1 (175.6) 49,120,245	183.9 (181.1) 53,128,257	207.4 (206.3) 62,144,288	172.3 (172.4) 48,119,243	163.9 (164.1) 47,112,226
	0.75	193.6 (193.8) 56,135,266	157.8 (158.3) 46,110,217	176.2 (176.6) 51,121,246	194.2 (195) 57,134,267	143.2 (144.1) 42,100,197	158.5 (157.9) 46,111,221	208.2 (208.3) 60,143,286	130.1 (132) 37,91,179	118.2 (118.8) 34,83,162
	1	187.7 (187.5) 54,129,261	131.7 (132.2) 38,90,181	159.7 (159.1) 47,111,221	195.2 (198.8) 56,134,267	113.3 (113.3) 32,78,156	132.6 (132.7) 38,91,184	192.8 (191) 56,136,268	79.4 (77.8) 23,55,110	73.3 (72.7) 21,50,102
MVG	0	199.9 (199) 59,140,276	196.8 (191.3) 58,139,277	196.8 (197.2) 56,137,275	198.1 (197.9) 58,137,271	197.3 (198) 56,136,276	202.6 (202) 58,141,277	198.6 (197.6) 57,138,273	194.8 (193.2) 57,136,272	197.4 (199.9) 55,135,274
	0.1	202.2 (201.1) 60,140,282	193.6 (197.8) 55,131,268	195.9 (195.8) 57,136,270	200.2 (198.2) 59,141,278	195.1 (194.5) 56,134,272	193.9 (190.5) 58,136,270	202.5 (201.8) 58,139,280	197.1 (195.7) 56,137,274	196.4 (196.9) 57,136,271
	0.2	200.3 (198.1) 57,141,278	192.7 (194.7) 54,135,266	195.5 (195.9) 55,138,271	200.2 (203.6) 56,138,276	188.8 (189.3) 54,130,259	192.3 (192.7) 55,133,264	204 (203.2) 60,142,282	189.4 (188.6) 53,130,265	189.4 (189.9) 53,130,263
	0.3	199.8 (197) 60,142,275	183.4 (179) 54,130,256	188.6 (185.6) 56,130,262	196.8 (195.1) 55,138,276	178.1 (176.3) 50,125,251	182.2 (181.5) 54,125,252	205.8 (205) 60,142,285	184.8 (188.1) 54,126,257	171.6 (174.8) 49,117,234
	0.4	195.7 (193.4) 57,136,275	172.4 (172.4) 50,119,239	180.2 (179) 52,125,250	201.1 (199.2) 59,140,278	170.5 (171.1) 50,118,237	176.5 (174.3) 52,124,246	212.3 (212.3) 60,148,298	169.1 (163.5) 50,119,239	154.8 (153.7) 45,107,214
	0.5	196.3 (198.3) 57,137,273	160.7 (160.6) 46,110,224	174.2 (174.6) 50,121,240	197 (197.7) 56,135,270	153 (149.1) 45,109,212	164.4 (164.7) 47,114,225	214.8 (215) 62,148,296	147.4 (144.5) 43,103,207	129.3 (128.7) 37,90,179
	0.75	188.5 (189.4) 55,130,259	130.7 (129.8) 38,90,182	152.2 (152.8) 43,105,213	198 (195.4) 58,137,274	117.7 (115.7) 34,82,162	129.7 (128.5) 38,91,179	213.5 (210.2) 62,149,292	85.4 (85.6) 25,59,118	71.5 (72.1) 20,49,100
	1	187.3 (186.9) 55,130,258	103.1 (100.6) 31,73,143	126.1 (125.7) 37,88,176	197.1 (199.1) 57,137,271	80.4 (80.3) 23,55,112	100.3 (99.6) 29,70,137	199.8 (198.3) 58,139,275	41.9 (41.4) 12,29,58	35.3 (34.4) 11,25,49

The first row of each cell shows ARL (SDRL), second row of each cell shows the values of 25th, 50th, and 75th percentiles, respectively. Bold numbers mean the smallest ARL for each shift in different dimension size.

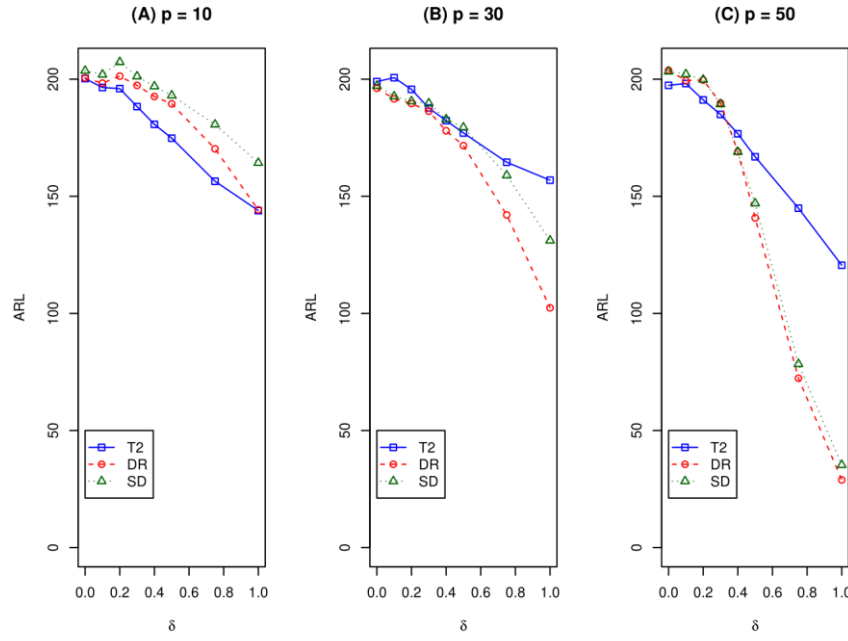


Figure 4: ARL comparison of Hotelling's T^2 and proposed charts with $m = 30$, $n = 30$ for MVN_0

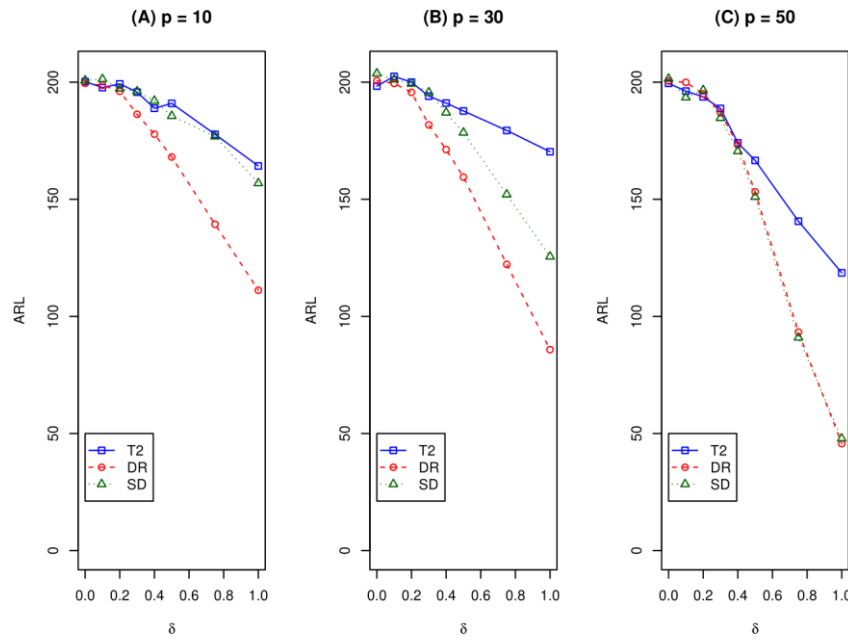


Figure 5: ARL comparison of Hotelling's T^2 and proposed charts with $m = 30$, $n = 30$ for MVN_1

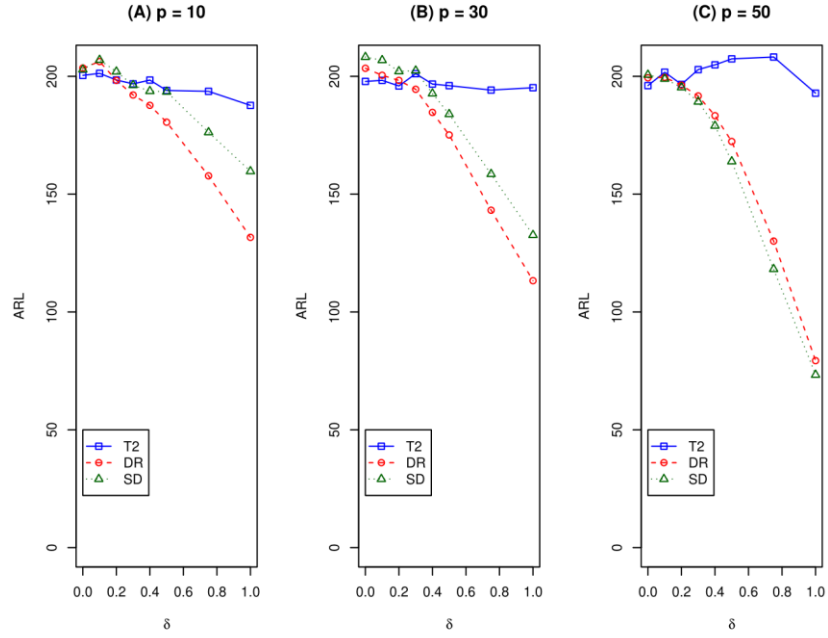


Figure 6: ARL comparison of Hotelling's T^2 and proposed charts with $m = 30$, $n = 30$ for MVT

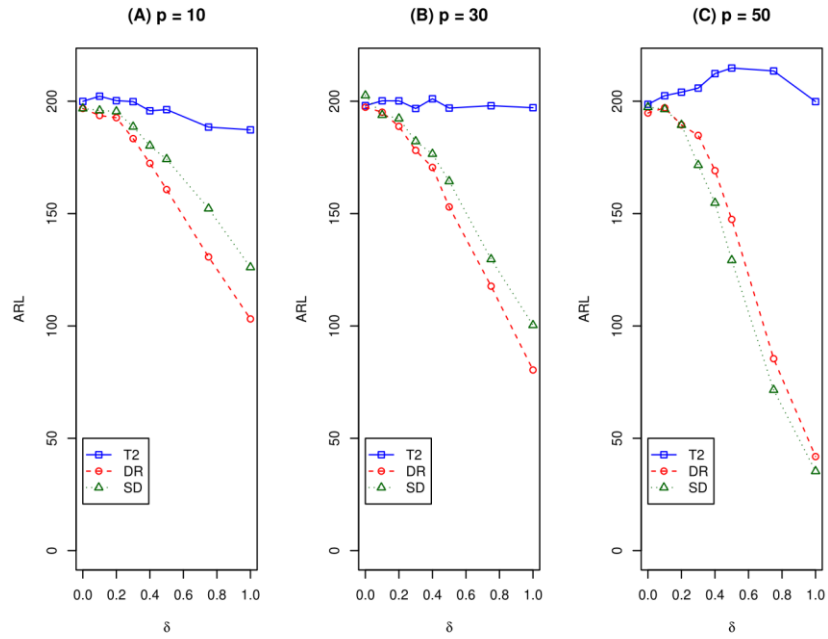


Figure 7: ARL comparison of Hotelling's T^2 and proposed charts with $m = 30$, $n = 30$ for MVG

For the lower dimensions of $p = 10, 30$, and 50 , only one combination of sample size $m = 30, n = 30$ is considered. Figure 4 shows that the proposed DR chart performs the best in

general for MVN_0 , however, only for the dimension $p = 10$, Hotelling's T^2 performs the best, which is the only case when the Shewhart type chart performs better than the proposed charts. Hence, for the remaining comparison, either of the proposed charts will be found better. Figure 5 depicts that for MVN_1 , DR chart's performance is better for $p = 10$ and 30 , however, SD chart performs better for $p = 50$. From Figure 6, it can be noticed that for MVT with $p = 50$, SD chart's performance is more satisfactory than others, whereas DR chart gives early detection for $p = 10$ and 30 . As displayed in Figure 7, with $p = 50$, the performance of SD chart for MVG is even finer than its performance for MVT.

3.4.2 Performance of the Proposed Charts under Higher Dimensions

As aforementioned, two-sample Hotelling's T^2 chart can not be implemented for higher dimensions, e.g., $p = 100$ and 300 . Therefore, under a higher dimension environment, the run length profile performance is compared between the proposed DR and SD charts.

- The control limits for the proposed charts for $ARL_0 = 168, 200, 370, \text{ and } 500$ are presented in Table 5. However, only $ARL_0 = 200$ will be used to conduct comparative analysis of the charts. For example, for $ARL_0 = 200$, with $p = 100$ ($m = 30, n = 30$), distribution = MVN_1 , the LCL and UCL of DR chart are 0.5796 and 1.6214 , respectively. Similarly, with $p = 300$, distribution = MVT, the LCL and UCL of SD chart are -1.6393 and 2.0354 , respectively.
- As shown in Table 6, under the distribution structure MVN_0 , for $p = 100$ ($m = n = 30\%$, and $m = 30\%, n = 45\%$), DR chart has shown better detection ability as its values of ARL_1 for different shifts are least than that of the SD charts. Similarly, for the same distribution structure with $p = 300$ ($m = n = 30\%$, and $m = 30\%, n =$

45%), the proposed DR chart performed the best as it provides smallest values of ARL_1 for different shifts. Though the values of ARL_1 of SD chart are smallest for small shifts such as $\delta = 0.1$ ($p = 100, m = n = 30\%$; $p = 300, m = 30\%, n = 45\%$), and $\delta = 0.1$, and 0.2 ($p = 100, m = 30\%, n = 45\%$); for large shifts the DR chart has shown the best detection ability. The smallest values of ARL_1 are shown in bold font in Table 6.

- Under the distribution structure MVN_1 , the detection ability of the DR chart for both $p = 100$ and 300 is better for different shift values as compared to the SD chart, except shift $\delta = 0.1$ ($p = 300, m = n = 30\%$), $\delta = 0.2$ and 0.3 ($p = 100, m = 30\%, n = 45\%$), and $\delta = 0.1$ and 0.2 ($p = 300, m = 30\%, n = 45\%$) where the ARL_1 values of the SD charts are reported as 186.7, 189.5, 171.3, 189.9, and 109.1, respectively (cf. Table 6).
- For all combinations of sample sizes, distributions, and dimensions $p = 100$, and 300 , with the distribution structure MVT, only for some higher shifts such as $\delta = 0.75$ and 1 , DR chart presents small ARL_1 values. For all the other shifts with all combinations of samples sizes and dimensions, $p = 100$ and 300 , the SD chart has shown better detection ability as compared to the DR chart (cf. Table 6).
- SD chart's performance is the best for both combinations of samples sizes, and with $p = 100$ and 300 under the distribution MVG for all shifts except $\delta = 0.1, p = 100$ ($m = 30\%, n = 45\%$) where DR chart gives small ARL_1 value (204.2) (cf. Table 6).

Table 5: Control limits of the proposed charts under higher dimensions

Sample Size	p	Distribution	ARL ₀	DR		BS		SD	
				LCL	UCL	LCL	UCL	LCL	UCL
$m = p \times 30\%, n = p \times 30\%$	100	MVN ₀	168	0.6520	1.4393	-2.4448	3.0832	-2.3096	2.9489
			200	0.6458	1.4496	-2.4872	3.1539	-2.3512	3.0215
			370	0.6261	1.4855	-2.6256	3.4023	-2.4890	3.2702
			500	0.6158	1.5028	-2.6963	3.5285	-2.5530	3.3896
		MVN ₁	168	0.5862	1.6057	-2.2564	3.3128	-2.0862	3.0857
			200	0.5796	1.6214	-2.2893	3.3968	-2.1207	3.1605
			370	0.5577	1.6747	-2.4092	3.6961	-2.2307	3.4375
			500	0.5473	1.6998	-2.4665	3.8385	-2.2783	3.5790
		MVT	168	0.5891	1.5981	-1.7034	2.4726	-1.5531	2.2343
			200	0.5827	1.6139	-1.7314	2.5397	-1.5792	2.2976
			370	0.5620	1.6662	-1.8307	2.7757	-1.6703	2.5096
			500	0.5527	1.6936	-1.8781	2.8791	-1.7129	2.6102
		MVG	168	0.5995	1.5748	-2.1617	3.1123	-2.0530	2.8863
			200	0.5930	1.5903	-2.1943	3.1978	-2.0859	2.9588
			370	0.5721	1.6382	-2.3016	3.4615	-2.1907	3.2019
			500	0.5621	1.6606	-2.3591	3.5844	-2.2473	3.3268
	300	MVN ₀	168	0.7888	1.2403	-2.5814	2.9347	-2.4503	2.7924
			200	0.7850	1.2454	-2.6259	2.9975	-2.4940	2.8543
			370	0.7719	1.2628	-2.7847	3.2135	-2.6512	3.0640
			500	0.7653	1.2720	-2.8674	3.3206	-2.7204	3.1691
		MVN ₁	168	0.7413	1.3233	-2.4494	3.0616	-2.2930	2.8695
			200	0.7366	1.3316	-2.4917	3.1410	-2.3321	2.9458
			370	0.7214	1.3593	-2.6366	3.4026	-2.4656	3.1902
			500	0.7140	1.3714	-2.7023	3.5154	-2.5297	3.2958
		MVT	168	0.7424	1.3233	-1.7536	2.1706	-1.6091	1.9861
			200	0.7380	1.3307	-1.7869	2.2241	-1.6393	2.0354
			370	0.7224	1.3567	-1.8984	2.4068	-1.7431	2.2079
			500	0.7156	1.3682	-1.9505	2.4819	-1.7943	2.2763
		MVG	168	0.7500	1.3082	-2.4221	2.9884	-2.2842	2.7997
			200	0.7455	1.3154	-2.4675	3.0579	-2.3270	2.8622
			370	0.7305	1.3395	-2.6087	3.2923	-2.4651	3.0932
			500	0.7240	1.3522	-2.6721	3.4108	-2.5317	3.2029
	500	MVN ₀	168	0.8345	1.1832	-2.6117	2.8924	-2.5000	2.7715
			200	0.8313	1.1869	-2.6622	2.9517	-2.5454	2.8332
			370	0.8212	1.2008	-2.8250	3.1689	-2.7049	3.0367
			500	0.8163	1.2075	-2.9009	3.2735	-2.7760	3.1304
		MVN ₁	168	0.7935	1.2441	-2.5262	2.9871	-2.3901	2.8237
			200	0.7898	1.2496	-2.5713	3.0529	-2.4319	2.8846
			370	0.7767	1.2688	-2.7322	3.2889	-2.5813	3.1067
			500	0.7706	1.2780	-2.8059	3.4022	-2.6456	3.2163
		MVT	168	0.7943	1.2428	-1.7610	2.0641	-1.6326	1.9081
			200	0.7905	1.2483	-1.7968	2.1168	-1.6646	1.9562
			370	0.7774	1.2674	-1.9140	2.2811	-1.7796	2.1081
			500	0.7719	1.2756	-1.9698	2.3631	-1.8294	2.1917
		MVG	168	0.8012	1.2327	-2.5067	2.9328	-2.3837	2.7733
			200	0.7974	1.2377	-2.5554	2.9992	-2.4272	2.8410
			370	0.7851	1.2559	-2.7114	3.2301	-2.5742	3.0626
			500	0.7797	1.2642	-2.7779	3.3331	-2.6419	3.1584
$m = p \times 30\%, n = p \times 45\%$	100	MVN ₀	168	0.6529	1.4372	-2.4422	3.0742	-2.3104	2.9315
			200	0.6467	1.4472	-2.4815	3.1443	-2.3529	2.9998
			370	0.6262	1.4831	-2.6286	3.3916	-2.4874	3.2412
			500	0.6157	1.4998	-2.6965	3.5074	-2.5507	3.3600
		MVN ₁	168	0.5874	1.6061	-2.2522	3.3122	-2.0802	3.0838
			200	0.5809	1.6230	-2.2864	3.4046	-2.1125	3.1633
			370	0.5592	1.6772	-2.4083	3.7103	-2.2254	3.4511
			500	0.5494	1.7024	-2.4599	3.8491	-2.2766	3.5832
		MVT	168	0.5777	1.6313	-1.7836	2.6404	-1.6247	2.3864
			200	0.5714	1.6477	-1.8126	2.7114	-1.6508	2.4507
			370	0.5494	1.7068	-1.9131	2.9562	-1.7442	2.6695

		MVG	500	0.5397	1.7317	-1.9570	3.0707	-1.7848	2.7795
			168	0.5990	1.5781	-2.1856	3.1718	-2.0569	2.9382
			200	0.5928	1.5926	-2.2198	3.2469	-2.0901	3.0094
			370	0.5712	1.6439	-2.3376	3.5355	-2.2051	3.2759
	300	MVN ₀	500	0.5618	1.6684	-2.3888	3.6856	-2.2576	3.4023
			168	0.7897	1.2402	-2.5698	2.9363	-2.4544	2.8048
			200	0.7857	1.2455	-2.6167	2.9996	-2.5007	2.8627
			370	0.7726	1.2631	-2.7788	3.2138	-2.6553	3.0735
		MVN ₁	500	0.7666	1.2716	-2.8476	3.3188	-2.7211	3.1770
			168	0.7407	1.3234	-2.4548	3.0666	-2.3065	2.8843
			200	0.7360	1.3313	-2.5009	3.1393	-2.3476	2.9524
			370	0.7204	1.3583	-2.6464	3.3977	-2.4846	3.1965
		MVT	500	0.7135	1.3705	-2.7119	3.5123	-2.5489	3.3074
			168	0.7306	1.3413	-1.8677	2.3408	-1.7247	2.1536
			200	0.7258	1.3490	-1.9039	2.3949	-1.7559	2.2069
			370	0.7098	1.3772	-2.0202	2.5920	-1.8655	2.3816
		MVG	500	0.7024	1.3900	-2.0719	2.6875	-1.9126	2.4681
			168	0.7496	1.3100	-2.4348	3.0172	-2.3008	2.8405
			200	0.7454	1.3174	-2.4785	3.0908	-2.3426	2.9083
			370	0.7311	1.3429	-2.6146	3.3393	-2.4768	3.1286
	500	MVN ₀	500	0.7246	1.3548	-2.6782	3.4529	-2.5392	3.2445
			168	0.8342	1.1833	-2.6175	2.8932	-2.5165	2.7843
			200	0.8308	1.1874	-2.6726	2.9586	-2.5679	2.8436
			370	0.8202	1.2005	-2.8390	3.1671	-2.7276	3.0458
		MVN ₁	500	0.8152	1.2070	-2.9172	3.2687	-2.8031	3.1421
			168	0.7940	1.2439	-2.5208	2.9843	-2.3930	2.8344
			200	0.7905	1.2491	-2.5652	3.0496	-2.4397	2.9030
			370	0.7770	1.2693	-2.7300	3.2957	-2.5872	3.1256
		MVT	500	0.7714	1.2778	-2.7992	3.4003	-2.6514	3.2234
			168	0.7849	1.2585	-1.8858	2.2361	-1.7605	2.0773
			200	0.7809	1.2645	-1.9253	2.2884	-1.7952	2.1273
			370	0.7673	1.2863	-2.0448	2.4742	-1.9083	2.3020
		MVG	500	0.7614	1.2953	-2.1010	2.5594	-1.9579	2.3826
			168	0.8010	1.2335	-2.5155	2.9526	-2.4012	2.8098
			200	0.7971	1.2390	-2.5644	3.0221	-2.4457	2.8711
			370	0.7852	1.2558	-2.7163	3.2351	-2.5870	3.0739
			500	0.7796	1.2644	-2.7883	3.3409	-2.6558	3.1803

Table 6: ARL (SDRL) comparison of DR and SD charts with $m = p \times 30\%$, $n = p \times 30\%$ and $m = p \times 30\%$, $n = p \times 45\%$

Sample Size	Dist.	δ	p = 100		p = 300		Sample Size	p = 100		p = 300	
			DR	SD	DR	SD		DR	SD	DR	SD
$m = p \times 30\%$, $n = p \times 30\%$	MVN ₀	0	199.8 (196.1) 59,141,277	198.8 (194.7) 58,140,277	200.5 (199.7) 57,140,279	200.6 (198.9) 57,140,279	$m = p \times 30\%$, $n = p \times 45\%$	204.3 (203.1) 59,142,285	199.7 (201.7) 59,138,273	199.8 (206.2) 56,135,272	199.4 (197.2) 60,138,275
		0.1	202.7 (203.4) 58,140,280	201 (200.2) 58,141,278	170.8 (172.2) 49,116,235	171.6 (171.5) 50,118,237		196.6 (194.7) 56,138,275	196.4 (195.3) 58,137,272	182.7 (181.5) 53,129,252	177.4 (176.6) 53,121,246
		0.2	180.6 (180.4) 53,126,250	181.2 (179.9) 52,126,254	51.5 (50.1) 15,36,72	52.1 (51.3) 15,36,72		183.6 (182.2) 55,128,256	182.8 (181.7) 52,128,256	78.9 (79.5) 23,55,108	79.5 (79.4) 23,55,110
		0.3	136 (138.4) 40,93,185	139.8 (140.4) 40,97,194	9 (8.6) 3,6,12	9.4 (8.9) 3,7,13		155.2 (155.1) 44,108,215	157 (156.8) 47,110,214	18.6 (18.3) 6,13,26	19.2 (18.9) 6,13,26
		0.4	80.8 (79.5) 24,56,113	85.1 (85.1) 25,59,117	2.3 (1.7) 1,2,3	2.4 (1.8) 1,2,3		111.6 (110.2) 33,78,155	116.3 (116.5) 34,81,159	4.6 (4.1) 2,3,6	4.7 (4.2) 2,3,6
		0.5	42 (42.2) 13,29,58	45.9 (46.1) 14,31,64	1.2 (0.4) 1,1,1	1.2 (0.5) 1,1,1		70.3 (70.1) 21,48,97	73.4 (73.9) 21,51,101	1.7 (1.1) 1,1,2	1.8 (1.2) 1,1,2
		0.75	7.6 (7) 3,5,10	9.1 (8.6) 3,6,12	1 (0) 1,1,1	1 (0) 1,1,1		16.9 (16.1) 5,12,24	18.9 (18.7) 6,13,26	1 (0) 1,1,1	1 (0) 1,1,1
		1	2.1 (1.5) 1,2,3	2.6 (2.1) 1,2,3	1 (0) 1,1,1	1 (0) 1,1,1		4.7 (4.1) 2,3,6	5.7 (5.1) 2,4,8	1 (0) 1,1,1	1 (0) 1,1,1
	MVN ₁	0	197.6 (197.3) 58,138,272	199 (199.2) 58,139,273	197.5 (193.8) 57,139,275	197.1 (194.4) 57,138,276		196.3 (198.3) 56,136,270	195.2 (195.3) 56,134,270	198.7 (195.7) 58,140,276	197.8 (195.5) 57,139,274
		0.1	199.1 (195.2) 59,138,280	199.8 (198.6) 58,138,279	188.2 (190.2) 54,131,257	186.7 (189.8) 53,128,258		199.5 (196.3) 58,142,277	200.4 (198.3) 59,139,275	191.7 (189.8) 55,133,272	189.9 (188.8) 55,131,267
		0.2	190.8 (190.2) 55,131,264	189.7 (192) 55,131,260	86.4 (84.4) 26,61,120	87.1 (86.6) 26,61,121		190 (189.6) 54,132,266	189.5 (189) 56,132,261	109.8 (108.2) 33,78,151	109.1 (108.3) 33,77,150
		0.3	168.7 (164.7) 48,119,236	170.6 (165.5) 51,120,242	20.1 (19.3) 6,14,28	20.5 (19.7) 6,14,28		177.7 (175.5) 52,125,246	171.3 (170) 50,120,239	34.9 (34.3) 11,24,48	35 (34) 11,25,48
		0.4	116.8 (118) 34,81,159	121 (119.9) 36,85,166	4.9 (4.3) 2,3,7	5.1 (4.5) 2,4,7		138.7 (139) 41,95,192	139.3 (136.2) 42,99,192	9.8 (9.3) 3,7,14	10.1 (9.5) 3,7,14
		0.5	72.8 (73.7) 21,50,99	76.6 (77.7) 23,53,105	1.7 (1.1) 1,1,2	1.8 (1.2) 1,1,2		99.2 (100) 29,68,136	101.3 (100.2) 29,71,140	3.4 (2.9) 1,2,4	3.4 (2.9) 1,2,4
		0.75	17.5 (17) 5,12,24	20.2 (19.7) 6,14,28	1 (0) 1,1,1	1 (0) 1,1,1		31.7 (31.4) 9,22,44	34.5 (34.1) 10,24,48	1 (0.2) 1,1,1	1 (0.2) 1,1,1
		1	4.6 (4.1) 2,3,6	5.8 (5.3) 2,4,8	1 (0) 1,1,1	1 (0) 1,1,1		9.7 (9.2) 3,7,13	11.6 (11.3) 4,8,16	1 (0) 1,1,1	1 (0) 1,1,1
	MVT	0	198.5 (197.6) 58,138,276	199.1 (198.7) 59,138,277	199.5 (197.4) 57,139,281	201.3 (205.4) 57,138,277		196 (194.3) 56,139,271	194.8 (193.2) 56,137,269	201.3 (200.8) 59,141,280	200.5 (202.5) 58,139,276
		0.1	195.9 (195.9) 55,136,273	195.6 (196.9) 55,135,271	199.5 (199.9) 59,140,275	196.7 (196.7) 57,135,271		195.6 (194.1) 56,135,270	195 (194.2) 57,137,272	198.2 (198.7) 57,137,272	199.9 (195.7) 58,141,280
		0.2	191.2 (189.7) 55,135,266	190.2 (188.8) 55,133,264	137.3 (138.7) 40,95,189	133.4 (133.1) 39,93,185		195.3 (195.3) 57,134,270	194.8 (193.4) 57,134,271	159.9 (160.4) 46,110,223	156.9 (159.9) 44,107,214
		0.3	185.5 (185.4) 52,127,259	182.5 (182.2) 52,125,254	54.7 (53.3) 16,38,77	53.3 (52) 16,37,74		190.9 (191.1) 55,130,265	187.7 (192.7) 53,127,258	86.9 (85.3) 25,61,120	81.5 (81.9) 24,56,112
		0.4	154.3 (151.5) 46,108,215	151.9 (149) 45,107,209	17 (16.4) 5,12,24	17 (16.1) 5,12,24		170.6 (170.4) 48,117,237	167.1 (163.7) 48,116,235	34.6 (34.2) 10,24,48	32.3 (32.1) 10,23,44
		0.5	119.2 (116.8) 35,83,167	117.3 (116.5) 34,82,163	5.5 (4.9) 2,4,7	5.7 (5.2) 2,4,8		144.4 (142.3) 42,101,201	142.3 (140.1) 42,100,199	13 (12.6) 4,9,18	12.2 (11.8) 4,9,17
		0.75	42 (41.2) 13,29,59	42.7 (42.4) 12,30,60	1.1 (0.4) 1,1,1	1.2 (0.5) 1,1,1		68.8 (67.5) 20,48,95	67.4 (66.5) 20,47,93	1.9 (1.3) 1,1,2	2 (1.4) 1,1,2
		1	12.8 (12.1) 4,9,17	14.2 (13.6) 4,10,19	1 (0) 1,1,1	1 (0.1) 1,1,1		27.6 (26.8) 8,19,38	28.2 (27.3) 8,20,39	1 (0.2) 1,1,1	1.1 (0.3) 1,1,1
	MVG	0	195 (196.9) 56,134,270	197.1 (197.4) 57,137,273	200.5 (201.4) 56,138,280	196.6 (195.6) 56,137,275		199.7 (196) 59,140,278	201.7 (202.5) 57,137,280	204.3 (203.2) 60,144,282	201.3 (202.1) 58,140,277
		0.1	198.5 (197.7) 57,138,278	198.5 (198.1) 58,138,275	187 (187.6) 54,129,258	182.9 (181.7) 53,126,252		204.2 (202.1) 60,143,284	204.4 (202) 61,143,282	202.5 (203.6) 57,138,284	200.5 (201.3) 59,139,277
		0.2	188.8 (191.5) 54,129,261	184.7 (184.1) 55,126,259	78.9 (78.1) 22,56,109	73.3 (72.4) 21,51,102		211.9 (209.1) 63,148,294	209.9 (209.8) 62,148,290	119.2 (119.1) 34,83,166	114.4 (116) 39,79,157
		0.3	163.6 (164.1) 47,114,225	155.4 (154.3) 46,109,216	17.7 (17.2) 6,12,24	16.4 (15.9) 5,11,22		200.6 (201.6) 59,138,275	189.9 (188.4) 55,132,267	34.6 (34) 10,24,48	32.3 (32.4) 9,22,44
		0.4	115.3 (115.2) 33,79,159	101.8 (101.5) 29,71,140	4.3 (3.8) 2,3,6	3.9 (3.3) 1,3,5		162.4 (162.3) 48,111,224	152.7 (152.9) 44,106,210	9.3 (8.8) 3,7,13	8.5 (8) 3,6,11
		0.5	68.8 (68.2) 21,48,95	58.7 (58.3) 18,41,81	1.6 (1) 1,1,2	1.5 (0.9) 1,1,2		116.9 (116.7) 34,82,161	103.8 (103.2) 31,73,143	3 (2.5) 1,2,4	2.9 (2.3) 1,2,4
		0.75	15.2 (14.9) 5,11,21	13.2 (12.7) 4,9,18	1 (0) 1,1,1	1 (0) 1,1,1		34.3 (34) 10,24,48	29 (28.8) 9,20,40	1 (0.1) 1,1,1	1 (0.1) 1,1,1
		1	3.9 (3.4) 2,3,5	3.7 (3.2) 1,3,5	1 (0) 1,1,1	1 (0) 1,1,1		9.6 (9) 3,7,13	8.7 (8.3) 3,6,12	1 (0) 1,1,1	1 (0) 1,1,1

The first row of each cell shows ARL (SDRL), second row of each cell shows the values of 25th, 50th, and 75th percentiles, respectively. Bold numbers mean the smallest ARL for each shift in different dimension size.

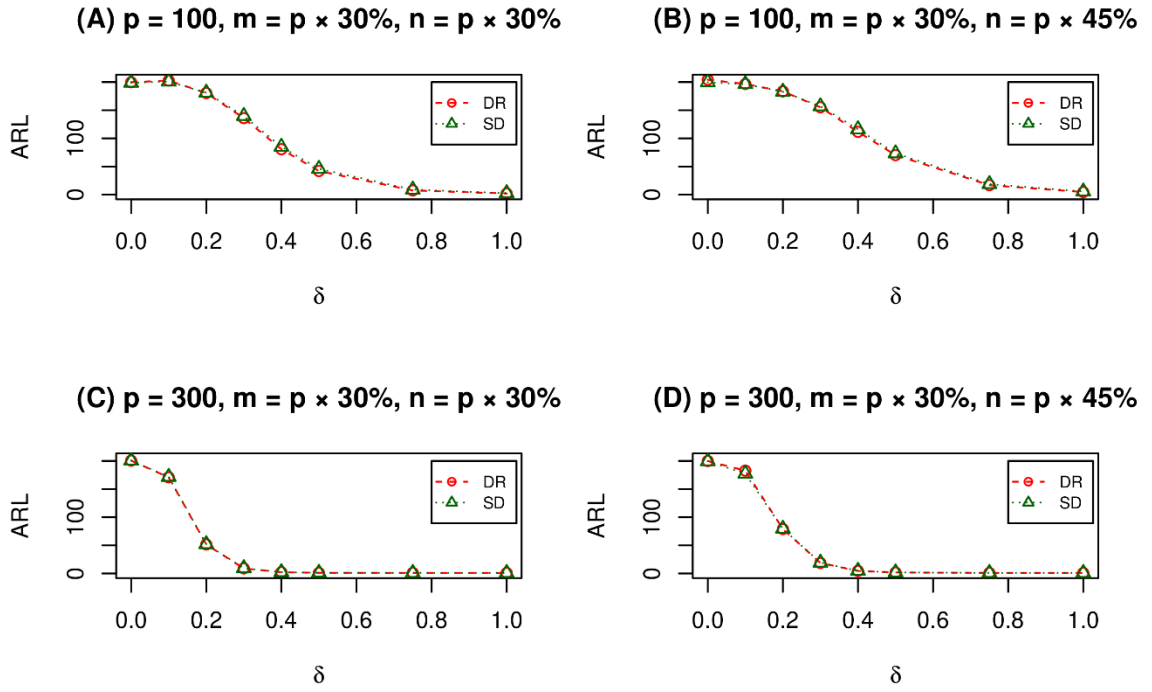


Figure 8: ARL comparison of DR and SD charts with 4 combinations of dimensions and sample sizes for MVN_0

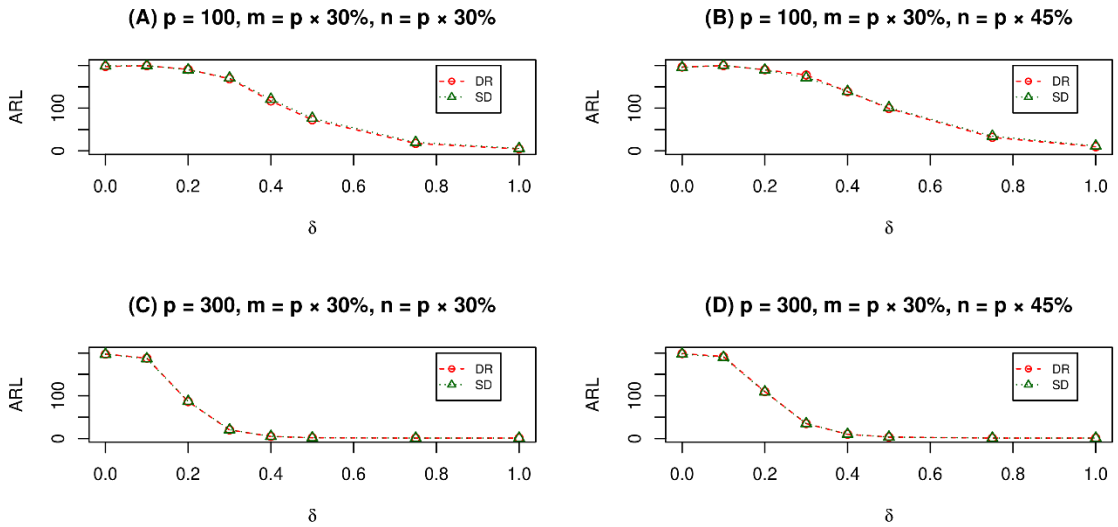


Figure 9: ARL comparison of DR and SD charts with 4 combinations of dimensions and sample sizes for MVN_0

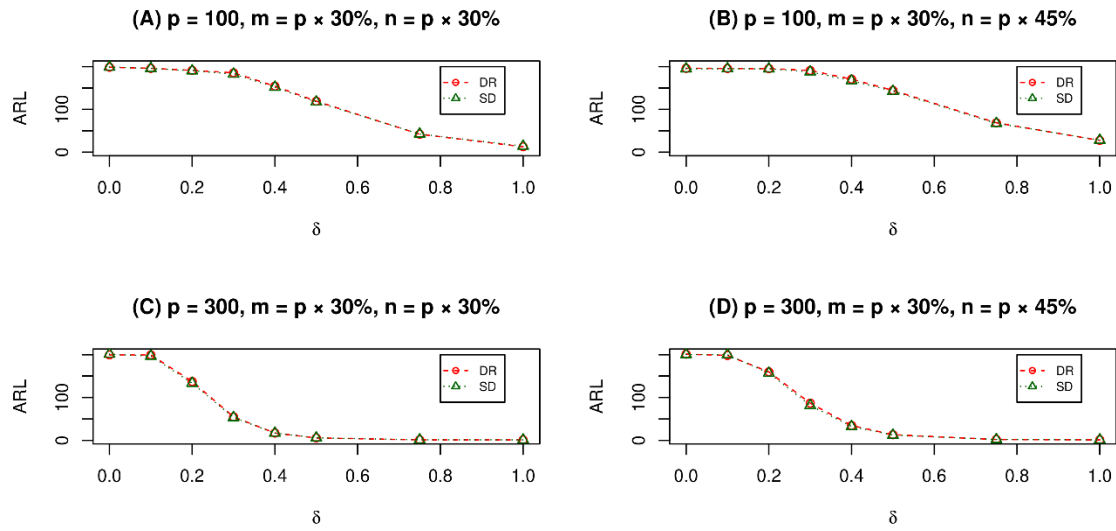


Figure 10: ARL comparison of DR and SD charts with 4 combinations of dimensions and sample sizes for MVT

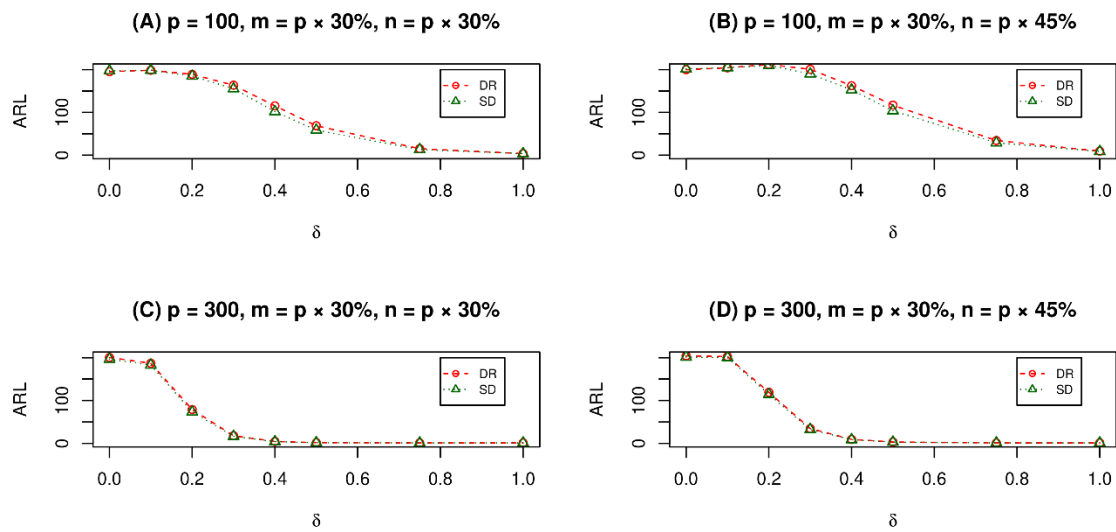


Figure 11: ARL comparison of DR and SD charts with 4 combinations of dimensions and sample sizes for MVG

The above figures illustrated the performance of the two proposed Shewhart based control charts. Due to the close difference between the ARL values of the two proposed charts, the

figures show that their performance is similar to each other. To study the exact difference between the two proposed Shewhart control charts, one may refer to Table 6.

3.5 A Case Study about Bearings of Wind Turbines

Being a reliable, inherent in nature, and environment-friendly energy source, wind power has become the fastest-growing renewable energy globally, paving the path toward a future reliant on clean energy [111]. Wind turbines are created to transform the wind's kinetic energy into mechanical or electrical energy. Two popular designs of wind turbines are shown in Figure 12. The main bearings, gearbox bearings, generator bearings, blade bearings, and yaw bearings are vital elements in large-scale wind turbines; these bearings play a critical role in converting the wind's kinetic energy into electrical energy [112]. Hundreds of wind turbines, each with their individual main bearing, can be found in large-scale wind farms, whereas smaller wind farms typically comprise only a handful of turbines. Grease is a semi-solid lubricating substance composed of base oil, thickeners, and additives, used to provide long-lasting lubrication, reduce friction, and protect against wear and corrosion in various industrial and mechanical applications including the bearings in wind turbines. Grease damage is one of the important factors that may cause deterioration to the bearings of wind turbines. The aim of this case study is to apply the proposed control charts to efficiently detect anomaly in the wind turbines main bearing caused by grease damage.

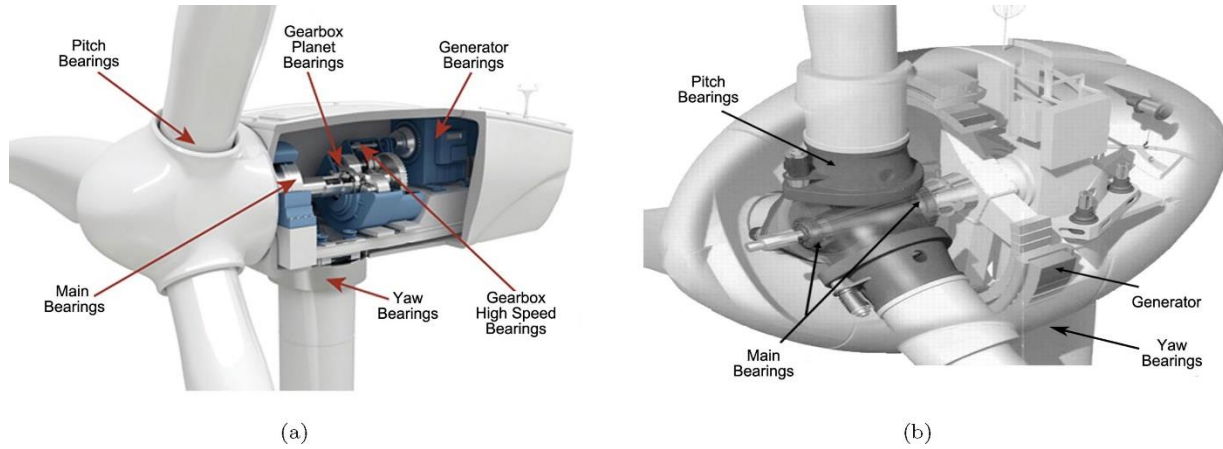


Figure 12: Wind turbine configuration (A) gearbox operated wind turbine [113] and (B) direct-drive wind turbine [114]

This section will elaborate on the real-life application of our proposed DR, BS, and SD charts to monitor shifts in mean of high-dimensional data with two samples. With this purpose, true grease damage dataset for 6 months [115] is used. The dataset can also be found online at the following link: <https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/J3IUVR>.

The dataset has 25,920 rows (observations) and 123 columns (features or dimensions). Each column represents a wind turbine number, indicating a total of 123 wind turbines of the main bearing. Each row represents the measure of true grease damage recorded every 10 minutes for each of the 123 turbines. Hence, the dataset consists of 6 months of observations for the 123 wind turbines.

Using the “EnvCpt” R package version 1.1.3 [116], the change point of the data is found at 17,280. The package fits all the possible models, which are based on a variety of trend, change-point, and autocorrelation models, on 25,920 row means, and found that the change point occurs after the first 4 months, shown in Figure 13. Hence, the first 17,280 rows (4 months observations) are considered as IC data. This IC data is divided into 288 samples,

with 60 as the sample size. Each sample is grouped into two sub-samples by considering equal sample size as $m = n = 30$. Using this high-dimensional data with two samples and considering $ARL_0 = 200$, the control limits of the proposed DR, BS, and SD charts are obtained as (156.59, 194.3), (111.04, 137.99), and (5.20, 38.23), respectively. Then, the test statistic values for the three proposed charts, resulting in 288 IC values for each chart, are obtained.

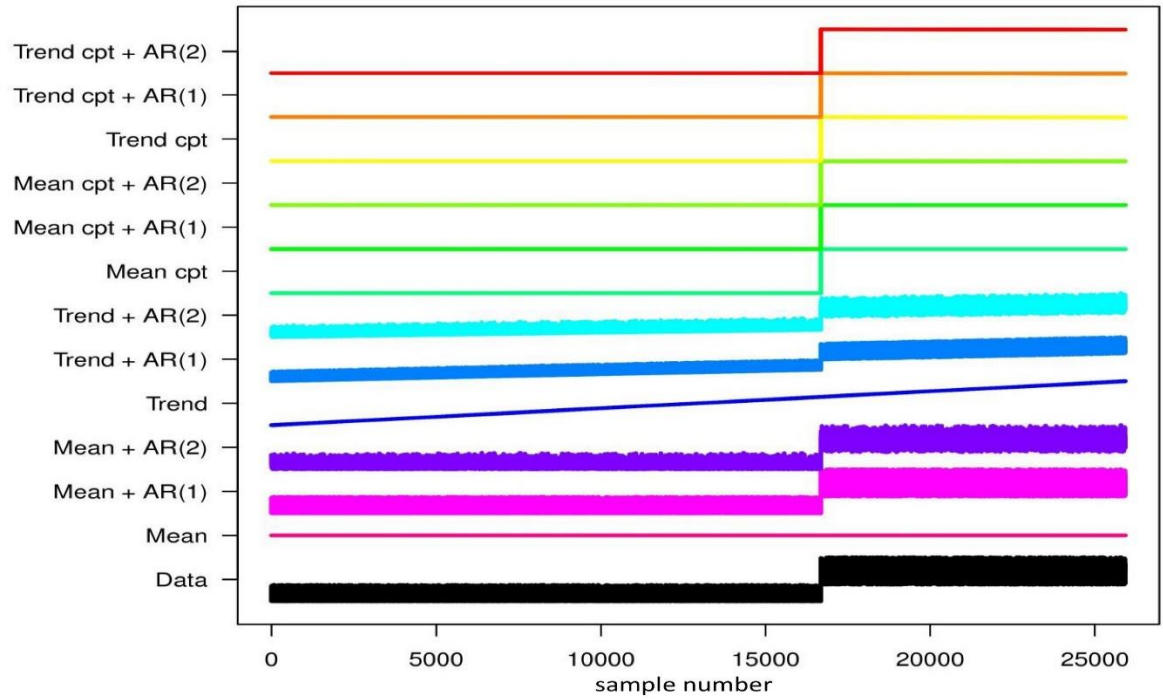


Figure 13: Change point analysis by fitting different models

Similar to the simulation study, the OOC data is recorded where the mean shift will occur in the second sample while the first sample will remain in-control. The first sample of the OOC data consists of 8,640 in-control observations taken from rows indexed 8,641 to 17,280 (2nd and 3rd month observations), while the second sample also consists of 8,640

observations taken from rows indexed 17,281 to 25,920 (last 2 months observations). Both samples have equal sample sizes, i.e., $m = n = 30$; the total sample size is 60. This arrangement of data provided 288 OOC values to find test statistics for DR, BS, and SD charts. Figure 14 depicts the charts' 288 IC and 288 OOC test statistic values along with their control limits. The IC and OOC values are separated by a vertical line in each of the three plots in the figure, where in-control region is presented in pink color while the white-shaded area is used as OOC region. Two horizontal green lines in each plot represent UCL and LCL for the corresponding charts. The plotting statistics of the charts are shown in blue points when they are in-control state and red points when they are in out-of-control state. For each of the charts, two false alarms have been noticed. The performance of the DR and BS charts are the same, as both charts detect 159 OOC values, while the SD chart detects 56 OOC values. Hence, compared to the SD chart, DR and BS charts are more efficient to study grease degradation behavior in wind turbines' bearings.

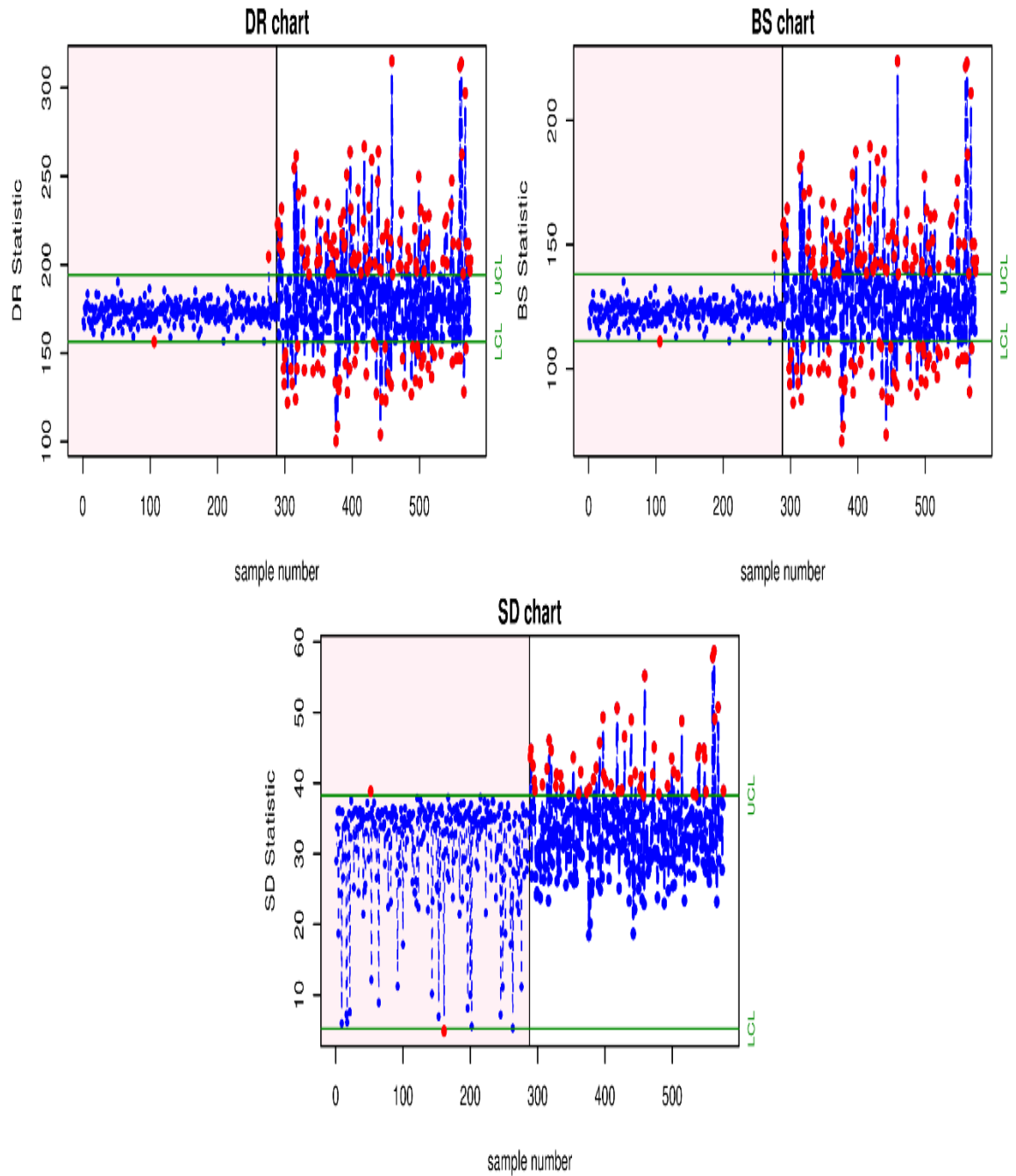


Figure 14: Graphical displays of plotting statistics for DR, BS, and SD charts corresponding to their control limits

CHAPTER 4

HIGH-DIMENSIONAL CONTROL CHARTS UNDER

EWMA ENVIRONMENT

An increase in the capacity of data storage and computational strength allows dealing with big data analysis. In such data analysis, when the number of variables surpasses the number of observations, it refers to high-dimensional data. Application of Statistical Process Control (SPC) on high-dimensional data is tedious. As an important tool of SPC techniques, a control chart is used to detect shifts in the process. However, it is not straightforward to construct control charts with high-dimensional data. A good number of studies focused on feature reduction techniques such as dimension reduction and variable selection to deal with high-dimensional data with one sample. But when it comes to two-sample high-dimensional data, more robust techniques rather than feature extraction methods are required to handle the data with efficiency. With this purpose, three Shewhart type charts named as DR, BS, and SD charts were proposed in the previous chapter. These charts showed good performance to detect large shifts in the second sample of high-dimensional data. A problem arises when these charts fail to detect small shifts in the process efficiently. To tackle such a problem, three memory-based EWMA control charts named DREWMA, BSEWMA, and SDEWMA are proposed. Both the simulation and real high-dimensional data are used to present the performance of the proposed charts. The proposed memory charts are found to be superior to the memoryless charts.

The remaining part of the chapter is as follows. In Section 4.1, the proposed two-sample high-dimensional charts are developed. Section 4.2 describes the designs required for the simulation study. In Section 4.3, the performance of the proposed charts is presented. The application of the proposed charts is provided in Section 4.4.

4.1 Proposed Two-Sample High-dimensional Charts

In this section, we discuss the methodologies of the three proposed charts. For a two-sample case (k refers to sample number, m represents the sample size of the first sample, n represents the sample size of the second sample, and p is used to determine dimensions), let x_{ijk} , $i = 1, 2, \dots, m$ or $i = 1, 2, \dots, n$, $j = 1, 2, \dots, p$, and $k = 1, 2$ are the p -dimensional multivariate normal random vectors with mean vector $\mu = (\mu_1, \dots, \mu_p)'$ and unknown common positive variance-covariance matrix Σ . In the two-sample case, when the number of variables is more than observations, the two-sample statistic T_1 [117] can be written as:

$$T_1 = \frac{\frac{mn}{m+n}(\bar{x}_1 - \bar{x}_2)' D_S^{-1} (\bar{x}_1 - \bar{x}_2) - \frac{(m+n)p}{m+n-2}}{\left[2 \left(\text{tr } R^2 - \frac{p^2}{m+n} \right) c_{p,m+n} \right]^{\frac{1}{2}}}, \quad (4.1)$$

where the diagonal matrix (D_S) of sample variances (S) is defined by

$$D_S = \text{diag}(s_{11}, \dots, s_{pp}), \quad (4.2)$$

The sample correlation matrix (R) is defined as follows:

$$R = D_S^{-\frac{1}{2}} S D_S^{-\frac{1}{2}}, \quad (4.3)$$

The adjustment coefficient $c_{p,m+n}$ can be chosen by

$$c_{p,m+n} = 1 + \frac{\text{tr } R^2}{p^{3/2}}. \quad (4.4)$$

Now, the T_1 statistic obtained from two-sample high-dimensional data is integrated to EWMA statistic as follows:

$$Z_{ij(SDEWMA)} = \lambda T_{1ij} + (1 - \lambda)Z_{ij-1(SDEWMA)} \quad (4.5)$$

The lower control limit (LCL) and upper control limit (UCL) of the statistic $Z_{ij(SDEWMA)}$ for a given false alarm α , can be obtained from the following limits:

$$LCL_{(SDEWMA)} = \mu_{0T_1} - L_1 \sqrt{\frac{\lambda}{2 - \lambda}} \sigma_{0T_1} \quad (4.6)$$

$$UCL_{(SDEWMA)} = \mu_{0T_1} + L_2 \sqrt{\frac{\lambda}{2 - \lambda}} \sigma_{0T_1} \quad (4.7)$$

where, μ_{0T_1} and σ_{0T_1} are the mean and standard deviation of the statistic T_1 . The values of L_1 and L_2 can be determined empirically, and $0 < \lambda \leq 1$. It is to be noted that the test statistic T_1 was proposed by Srivastava and Du [117]. Therefore, the control chart based on statistic T_1 is named as SD chart, and the chart based on T_1 and EWMA is named as SDEWMA chart.

The control chart on the basis of the statistic T_2 [106, 107] is named as DR chart. While the DREWMA chart is proposed based on the statistic T_2 and EWMA chart. The statistic T_2 can be expressed as follows:

$$T_2 = \frac{\frac{mn}{m+n} (\bar{x}_1 - \bar{x}_2)' (\bar{x}_1 - \bar{x}_2)}{\text{tr } S}. \quad (4.8)$$

The statistic for DREWMA chart is as follows:

$$Z_{ij(DREWMA)} = \lambda T_{2ij} + (1 - \lambda)Z_{ij-1(DREWMA)} \quad (4.9)$$

The lower control limit (LCL) and upper control limit (UCL) of the statistic $Z_{ij(DREWMA)}$ for a given false alarm α , can be obtained from the following limits:

$$LCL_{(DREWMA)} = \mu_{0T_2} - L_1 \sqrt{\frac{\lambda}{2-\lambda}} \sigma_{0T_2} \quad (4.10)$$

$$UCL_{(DREWMA)} = \mu_{0T_2} + L_2 \sqrt{\frac{\lambda}{2-\lambda}} \sigma_{0T_2} \quad (4.11)$$

where, μ_{0T_2} and σ_{0T_2} are the mean and standard deviation of the statistic T_2 . The values and L_1 and L_2 can be determined empirically, and $0 < \lambda \leq 1$.

The control chart on the basis of the statistic T_3 [108] is named as BS chart. Further, the BSEWMA chart is proposed based on EWMA chart and the statistic T_3 , which is given by:

$$T_3 = \frac{\frac{mn}{m+n} (\bar{x}_1 - \bar{x}_2)' (\bar{x}_1 - \bar{x}_2) - \text{tr } S}{\left[\frac{2(m+n)(m+n+1)}{(m+n-1)(m+n+2)} \left(\text{tr } S^2 - \frac{(\text{tr } S)^2}{m+n} \right) \right]^{\frac{1}{2}}} \quad (4.12)$$

The statistic for BSEWMA chart is as follows:

$$Z_{ij(BSEWMA)} = \lambda T_{3ij} + (1-\lambda) Z_{ij-1(BSEWMA)} \quad (4.13)$$

The lower control limit (LCL) and upper control limit (UCL) of the BSEWMA chart for a given false alarm α , can be obtained as follows:

$$LCL_{(BSEWMA)} = \mu_{0T_3} - L_1 \sqrt{\frac{\lambda}{2-\lambda}} \sigma_{0T_3} \quad (4.14)$$

$$UCL_{(BSEWMA)} = \mu_{0T_3} + L_2 \sqrt{\frac{\lambda}{2-\lambda}} \sigma_{0T_3} \quad (4.15)$$

where, μ_{0T_3} and σ_{0T_3} are the mean and standard deviation of the statistic T_3 . The values and L_1 and L_2 can be determined empirically, and $0 < \lambda \leq 1$.

It is to be noted that all proposed three charts, named SDEWMA, DREWMA, and BSEWMA charts signal an out-of-control (OOC) point when $Z_{ij(SDEWMA)}$, $Z_{ij(DREWMA)}$, and $Z_{ij(BSEWMA)}$ falls outside of their control limits, respectively. Otherwise, the process is considered to be in a stable or in-control (IC) situation.

4.2 Design of Simulation Study

Simulations have been carried out in three steps. In the first step, the values of means and standard deviations of T_1 , T_2 , and T_3 statistics are obtained under different distributions and dependency structures using 100,000 iterations. In the second step, for the pre-obtained means and standard deviations and pre-defined $ARL_0 = 200$ under different distributions and dependency structures, the values of L_1 and L_2 are found using 10,000 iterations. The third step consists of ARL_1 studies based on 10,000 iterations using the values of means, standard deviations, L_1 and L_2 for different combinations of distributions and dependencies. The following combinations of distributions and dependencies [110] are taken into account: (i) multivariate normal with unit variances and covariances equal to 0, denoted as MVN_0 ; (ii) multivariate normal with unit variances and covariances equal to 0.5, denoted as MVN_1 ; (iii) multivariate t with 5 degrees of freedom, denoted as MVT ; (iv) multivariate gamma with shape parameter 5 and scale parameter 1, denoted as MVG. For both MVT and MVG, unit variances and covariances are considered equal to 0.5.

Two types of high dimensions, such as $p = 100$ and 300 , are used. The sample size for two samples is considered as (a) $m = p \times 30\%$, $n = p \times 30\%$ and (b) $m = p \times 30\%$, $n = p \times 45\%$. A simulation study of 100,000 iterations has been performed to provide means and standard deviations and then L_1 and L_2 using 10000 iterations under the distribution cases (i) – (iv) for the two combinations of sample size (a) – (b) for the

proposed DREWMA and SDEWMA charts; these values are reported in Table 7. BSEWMA chart has not been considered for simulation study as it gives similar results as DREWMA chart, which is proved at the end with a case study. Finally, the comparative analysis of the proposed charts based on $ARL_0 = 200$, shifts ($\delta = 0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.75, 1$), sample sizes (a) – (b), and distribution cases (i) – (iv) have been provided in Table 8.

Further, Bai and Saranadasa [108] showed that the T_3 and T_1 statistics have the similar performance in testing the means of two-sample high-dimensional data. Similarly, we found in initial findings that the BSEWMA chart has the same run length profile as the DREWMA chart, also seen in the case study presented in Section 4.4. Therefore, for brevity, we have excluded the results of the BSEWMA chart.

4.3 Results and Discussions

The run length refers to the number of successive points before the out-of-control (OOC) state of a process. As described in the previous chapter, the average number of points until an out-of-control (OOC) signal is identified is known as Average Run Length (ARL). This ARL can be termed as ARL_0 when the process is in a stable state, while if the process is not in a stable state, then the term ARL is specified by ARL_1 . Starting with a preset ARL_0 , lower values of ARL_1 indicate earlier shift detection ability of a control chart. It is desired to obtain a lower ARL_1 values to get the shift signal early. Hence, the performance of the proposed charts is mainly evaluated based on the ARL values. However, other measures such as Standard Deviation of Run Length (SDRL) along with Q_1 (1st quartile), Q_2 (2nd quartile or median), and Q_3 (3rd quartile) of run length are also recorded with the ARL

values. These measures provide standard deviation, 25th percentile, 50th percentile, and 75th percentile of the run length, respectively.

4.3.1 Performance of the Proposed Charts

In the context of high-dimensional data with varying dimensions, such as $p = 100$ and 300 , this study conducts performance evaluation of the proposed charts based on run length profile.

- Table 7 provides the mean and standard deviation, L_1 , and L_2 for the proposed charts for $ARL_0 = 200$ under the distribution cases (i) – (iv) for the two combinations of sample size (a) – (b). For example, for $ARL_0 = 200$, with $p = 100$ ($m = 30, n = 45$), distribution = MVN_0 , the mean and standard deviation, L_1 , and L_2 of DREWMA chart are -0.00209, 0.99921, 2.49000, and 2.77500, respectively. In the same way, with $p = 300$ ($m = 90, n = 90$), distribution = MVN_1 , the mean and standard deviation, L_1 , and L_2 of SDEWMA chart are -0.00350, 0.93590, 2.50000, and 2.80000, respectively.
- As shown in Table 8, under the distribution structure MVN_0 , for $p = 100$ ($m = n = p \times 30\%$, and $m = p \times 30\%, n = p \times 45\%$) DREWMA chart has shown better detection ability as its values of ARL_1 for different shifts are least than that of the SDEWMA charts. Similarly, for the same distribution structure with $p = 300$ ($m = n = p \times 30\%$, and $m = p \times 30\%, n = p \times 45\%$), the proposed DREWMA chart performed the best as it provides smallest values of ARL_1 for different shifts. Though the values of ARL_1 of SDEWMA chart are smallest for some small shifts such as $\delta = 0.1, 0.2, \text{ and } 0.3$ ($p = 100, m = n = p \times 30\%$), and $\delta = 0.1,$

and 0.2 ($p = 300, m = 30\%, n = 45\%$); in general for large shifts the DREWMA chart has shown the best detection ability for MVN_0 . The smallest values of ARL_1 are shown in bold font in Table 8.

- Under the distribution structure MVN_1 , the detection ability of the DREWMA chart for both $p = 100$ and 300 is better for different shift values as compared to the SDEWMA chart, except shift $\delta = 0.1, 0.2$ ($p = 100, m = n = p \times 30\%$), and $\delta = 0.1$ ($p = 300, m = p \times 30\%, n = p \times 45\%$) where the ARL_1 values of the SDEWMA charts are reported as 198.6, 169.5, and 160.5, respectively (cf. Table 8).
- DREWMA chart has shown better detection ability as compared to the SDEWMA chart for both dimensions $p = 100$, and 300 with sample sizes of $m = n = p \times 30\%$, under the distribution structure of MVT, as DREWMA chart presents smaller ARL_1 values than the SDEWMA chart for the mentioned combinations of dimensions and sample sizes. Interestingly, the SDEWMA chart performs superior for the same setup when the change takes place in the sample sizes as $m = p \times 30\%, n = p \times 45\%$ (cf. Table 8).
- Under the distribution structure of MVG, other than the shift $\delta = 0.1$, dimension $p = 300$ ($m = p \times 30\%, n = p \times 30\%$), the DREWMA chart provides smaller ARL_1 value of 139.9. For all other combinations of shifts, dimensions, and sample sizes SDEWMA perform better than DREWMA (cf. Table 8).

Table 7: Mean, standard deviation, L_1 and L_2 for DREWMA and SDEWMA charts with $m = n = p \times 30\%$ and $m = p \times 30\%, n = p \times 45\%$

Sample Size	Distribution	$ARL_0 = 200$	p = 100		p = 300	
			DREWMA	SDEWMA	DREWMA	SDEWMA
m=p×30%, n=p×30%	MVN_0	μ_0	0.00020	-0.00292	-0.00418	-0.00464
		σ_0	1.00005	0.95140	1.00067	0.95197
		L_1	2.50000	2.50000	2.54000	2.54000
		L_2	2.79000	2.76800	2.72000	2.72000
	MVN_1	μ_0	0.00200	-0.00144	-0.00292	-0.00350
		σ_0	1.00057	0.92868	0.99933	0.93590
		L_1	2.42400	2.41500	2.50000	2.50000
		L_2	2.89000	2.88500	2.77000	2.80000
	MVT	μ_0	0.00023	-0.00679	0.00148	0.00013
		σ_0	0.72113	0.65516	0.68771	0.62989
		L_1	2.42000	2.45000	2.53000	2.53000
		L_2	2.91700	2.95000	2.78000	2.81700
	MVG	μ_0	-0.00058	-0.01588	-0.00172	-0.00510
		σ_0	0.94721	0.88686	0.97998	0.92185
		L_1	2.43000	2.65000	2.52000	2.50000
		L_2	2.89000	2.85000	2.78000	2.80000
m=p×30%, n=p×45%	MVN_0	μ_0	-0.00209	-0.00459	-0.00294	-0.00357
		σ_0	0.99921	0.94907	1.00081	0.95609
		L_1	2.49000	2.48500	2.53500	2.54000
		L_2	2.77500	2.79000	2.73500	2.72000
	MVN_1	μ_0	0.00194	-0.00069	-0.00059	-0.00114
		σ_0	1.00232	0.92876	1.00450	0.94417
		L_1	2.41500	2.41000	2.51000	2.48000
		L_2	2.86000	2.87500	2.75000	2.75000
	MVT	μ_0	-0.00167	-0.00614	0.00392	0.00211
		σ_0	0.77365	0.70027	0.74289	0.68294
		L_1	2.40500	2.41400	2.52500	2.53000
		L_2	2.93000	2.91000	2.78500	2.78000
	MVG	μ_0	0.00293	-0.00692	0.00138	-0.00113
		σ_0	0.96638	0.90169	0.98804	0.93338
		L_1	2.41000	2.42000	2.52000	2.50000
		L_2	2.86000	2.83100	2.75000	2.74000

Table 8: ARL (SDRL) comparison of DREWMA and SDEWMA charts with $m = n = p \times 30\%$ and $m = p \times 30\%, n = p \times 45\%$

Sample Size	Dist.	δ	p = 100		p = 300		Sample Size	p = 100		p = 300	
			DREWMA	SDEWMA	DREWMA	SDEWMA		DREWMA	SDEWMA	DREWMA	SDEWMA
$m = p \times 30\%, n = p \times 30\%$	MVN ₀	0	197.2 (194.3) 60,138,272	191.5 (188.4) 57,134,264	197 (197.2) 59,136,271	197.6 (197.9) 59,138,269	$m = p \times 30\%, n = p \times 45\%$	194.3 (187.5) 60,138,266	195.3 (190.8) 60,138,268	196.9 (191.1) 60,139,271	196.4 (190.1) 62,139,270
		0.1	196.9 (192.6) 58,139,273	189.2 (185.8) 57,133,259	109.4 (105.4) 35,76,150	109.9 (104.7) 35,77,150		192.6 (184.6) 59,136,272	193.6 (184.8) 59,137,273	144.1 (139.3) 45,103,195	139.2 (135.3) 43,98,192
		0.2	144.2 (139.1) 45,102,198	139.4 (132.6) 43,99,194	14.8 (10.8) 7,12,19	15 (11) 7,12,19		162.5 (156.7) 51,113,223	165.1 (160.1) 51,114,227	25.3 (20.6) 11,19,34	25 (20.7) 10,19,33
		0.3	64.7 (59.6) 22,46,89	63.9 (59.3) 21,46,88	4.4 (2) 3,4,5	4.5 (2.1) 3,4,6		94.5 (91.1) 30,67,128	97.5 (94.3) 31,69,132	6.8 (3.7) 4,6,8	6.8 (3.8) 4,6,9
		0.4	27 (22.6) 11,20,36	27.2 (23.2) 11,20,36	2.3 (0.8) 2,2,3	2.4 (0.8) 2,2,3		46 (41.6) 16,33,62	48.1 (43.5) 17,35,65	3.3 (1.3) 2,3,4	3.3 (1.3) 2,3,4
		0.5	12.7 (8.9) 6,10,17	13.4 (9.6) 7,11,17	1.6 (0.5) 1,2,2	1.6 (0.5) 1,2,2		22.2 (18.4) 9,17,29	23.2 (19.2) 10,17,31	2.1 (0.6) 2,2,2	2.1 (0.7) 2,2,2
		0.75	4.1 (1.8) 3,4,5	4.3 (2) 3,4,5	1 (0) 1,1,1	1 (0) 1,1,1		6.4 (3.5) 4,6,8	6.9 (4) 4,6,9	1 (0.2) 1,1,1	1 (0.2) 1,1,1
		1	2.2 (0.8) 2,2,3	2.4 (0.9) 2,2,3	1 (0) 1,1,1	1 (0) 1,1,1		3.2 (1.3) 2,3,4	3.5 (1.4) 2,3,4	1 (0) 1,1,1	1 (0) 1,1,1
	MVN ₁	0	198.8 (197.3) 61,137,272	196.9 (194.6) 60,137,271	196.7 (191.8) 60,139,271	204 (199.7) 62,144,281		194.1 (190.4) 58,136,267	197.4 (193.6) 58,138,273	201.5 (197.7) 61,140,277	192.2 (189.9) 58,134,264
		0.1	202.3 (198.9) 61,142,275	198.6 (195.7) 59,140,270	137.8 (130.9) 43,97,190	144.9 (137.3) 46,103,200		194.7 (191.7) 60,136,269	197 (193.7) 60,138,269	163.1 (161.7) 49,114,224	160.5 (155.6) 51,113,220
		0.2	169.8 (165.7) 53,119,234	169.5 (166.6) 52,117,232	24.3 (19.9) 10,18,32	25.4 (21) 11,19,34		174.8 (170.6) 54,124,242	178.8 (174.7) 56,126,246	39 (34.9) 14,28,53	40.1 (35.8) 15,29,54
		0.3	97.4 (92) 31,70,134	98.5 (92.5) 32,71,135	6.5 (3.5) 4,6,8	6.7 (3.7) 4,6,8		124.8 (121.5) 39,87,171	127.7 (121.7) 40,90,177	10.1 (6.5) 5,8,13	10.3 (6.7) 6,9,13
		0.4	45.2 (40.1) 17,33,61	46.9 (42) 17,35,64	3.1 (1.2) 2,3,4	3.2 (1.2) 2,3,4		67.7 (63) 23,48,94	69.9 (65.6) 23,50,95	4.5 (2.1) 3,4,6	4.5 (2.1) 3,4,6
		0.5	21.7 (17) 10,17,29	22.7 (18) 10,18,30	2 (0.6) 2,2,2	2.1 (0.6) 2,2,2		35.9 (31.3) 13,27,49	37.9 (33.5) 14,28,51	2.7 (1) 2,3,3	2.7 (1) 2,3,3
		0.75	5.9 (3.1) 4,5,7	6.3 (3.4) 4,6,8	1 (0.1) 1,1,1	1 (0.2) 1,1,1		9.7 (6.3) 5,8,12	10.4 (7.1) 6,9,13	1.2 (0.4) 1,1,1	1.3 (0.4) 1,1,2
		1	3 (1.1) 2,3,4	3.2 (1.3) 2,3,4	1 (0) 1,1,1	1 (0) 1,1,1		4.4 (2.2) 3,4,5	4.8 (2.4) 3,4,6	1 (0) 1,1,1	1 (0) 1,1,1
	MVT	0	196 (195.1) 59,134,268	211.8 (209.2) 63,148,291	200.6 (195.3) 61,140,277	206.9 (203.3) 62,144,287		199.2 (194.2) 61,140,274	197.4 (192) 61,138,270	197.6 (194) 60,138,273	195.3 (191) 59,137,268
		0.1	193.8 (186.6) 60,138,266	209.3 (204.5) 64,147,289	175.4 (171.3) 53,124,238	186.4 (183.4) 56,130,256		202.3 (198.2) 62,141,283	198.3 (193.1) 55,129,252	183.8 (182) 55,129,252	181.3 (177.1) 55,128,248
		0.2	181.5 (179.6) 55,125,250	194.3 (194.2) 60,136,262	55.2 (49.8) 19,40,76	57.5 (52.5) 20,41,79		195.2 (191.5) 59,137,269	192.3 (189.6) 57,134,264	86.4 (82.6) 28,62,116	83.5 (79.4) 27,59,114
		0.3	139 (133.4) 43,98,193	145.6 (139.1) 46,103,201	14.5 (10.4) 7,12,19	14.7 (10.9) 7,12,19		166.4 (161.1) 51,118,231	158.9 (154.8) 49,111,218	25.8 (21.4) 11,19,34	25.1 (20.9) 10,19,33
		0.4	83.2 (78.1) 27,59,114	86.9 (82.1) 28,63,119	6 (3.1) 4,5,7	6 (3.1) 4,5,7		121.6 (117.4) 38,86,168	115.3 (111.3) 37,81,159	9.9 (6.4) 5,8,13	9.6 (6.2) 5,8,12
		0.5	45.8 (41.2) 17,34,62	45.9 (41.5) 16,33,62	3.4 (1.4) 2,3,4	3.4 (1.4) 2,3,4		78.2 (73.5) 25,56,106	73.9 (69.8) 24,52,101	5.1 (2.6) 3,5,6	5.1 (2.6) 3,5,6
		0.75	12 (8.3) 6,10,15	12.3 (8.5) 6,10,16	1.5 (0.5) 1,1,2	1.6 (0.6) 1,2,2		22.6 (18.6) 10,17,30	21.8 (17.7) 9,17,29	2.1 (0.7) 2,2,2	2.1 (0.7) 2,2,2
		1	5.2 (2.7) 3,5,6	5.4 (2.8) 3,5,7	1.1 (0.2) 1,1,1	1.1 (0.2) 1,1,1		9 (5.8) 5,8,11	8.9 (5.6) 5,7,11	1.3 (0.4) 1,1,2	1.3 (0.4) 1,1,2
	MVG	0	196.1 (189.7) 61,137,273	209.6 (210.6) 66,138,287	205.1 (202.5) 62,144,279	203.2 (200.6) 62,141,278		196.8 (192.1) 59,136,276	195.4 (188.9) 59,136,272	203.9 (203.9) 61,143,278	196.6 (196.1) 59,138,269
		0.1	200.4 (198.8) 60,138,278	175.9 (167.1) 55,131,245	139.9 (135.3) 44,99,191	140.4 (136.7) 45,99,194		204.2 (201.2) 63,143,282	199.1 (197.8) 60,136,271	162.2 (158.1) 50,113,225	157.6 (151.4) 50,111,215
		0.2	167.1 (164.6) 50,117,229	132 (133.4) 40,87,178	22.7 (18) 10,18,30	21.7 (17.2) 10,17,29		192.6 (188) 59,136,266	183.3 (177.5) 56,131,253	38.2 (33.3) 14,28,52	35.6 (31.1) 13,26,48
		0.3	91.8 (87.4) 29,65,126	66.7 (63) 22,49,92	6.1 (3.2) 4,5,8	5.9 (3) 4,5,7		134.5 (130.8) 43,94,184	119.2 (117.1) 38,83,163	9.5 (5.9) 5,8,12	9.1 (5.7) 5,8,12
		0.4	41.1 (36.7) 15,30,55	28.6 (24.3) 12,21,39	3 (1.1) 2,3,4	2.9 (1) 2,3,3		71 (66.4) 24,50,97	61.5 (56.8) 21,44,84	4.2 (1.8) 3,4,5	4.1 (1.8) 3,4,5
		0.5	19.5 (15.2) 9,15,26	14.9 (11.5) 7,12,20	1.9 (0.6) 2,2,2	1.9 (0.6) 2,2,2		35.5 (31.1) 14,26,47	30.1 (25.6) 12,23,40	2.6 (0.9) 2,2,3	2.5 (0.9) 2,2,3
		0.75	5.5 (2.8) 3,5,7	4.7 (2.3) 3,4,6	1 (0.1) 1,1,1	1 (0.1) 1,1,1		9.2 (5.8) 5,8,12	8.3 (5.2) 5,7,11	1.2 (0.4) 1,1,1	1.2 (0.4) 1,1,1
		1	2.8 (1) 2,3,3	2.6 (0.9) 2,2,3	1 (0) 1,1,1	1 (0) 1,1,1		4.2 (1.9) 3,4,5	4 (1.8) 3,4,5	1 (0) 1,1,1	1 (0) 1,1,1

The first row of each cell shows ARL (SDRL), and the second row of each cell shows the values of 25th, 50th, and 75th percentiles, respectively. Bold numbers mean the smallest ARL for each shift in different dimension sizes.

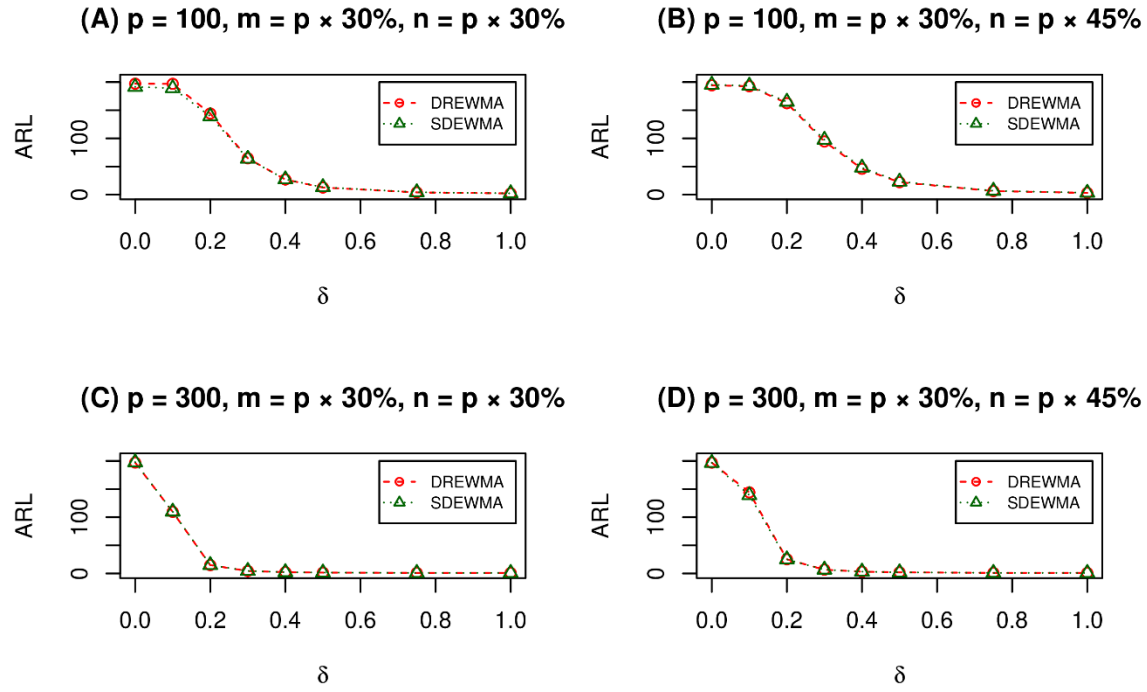


Figure 15: ARL comparison of DREWMA and SDEWMA charts with 4 combinations of dimensions and sample sizes for MVN_0

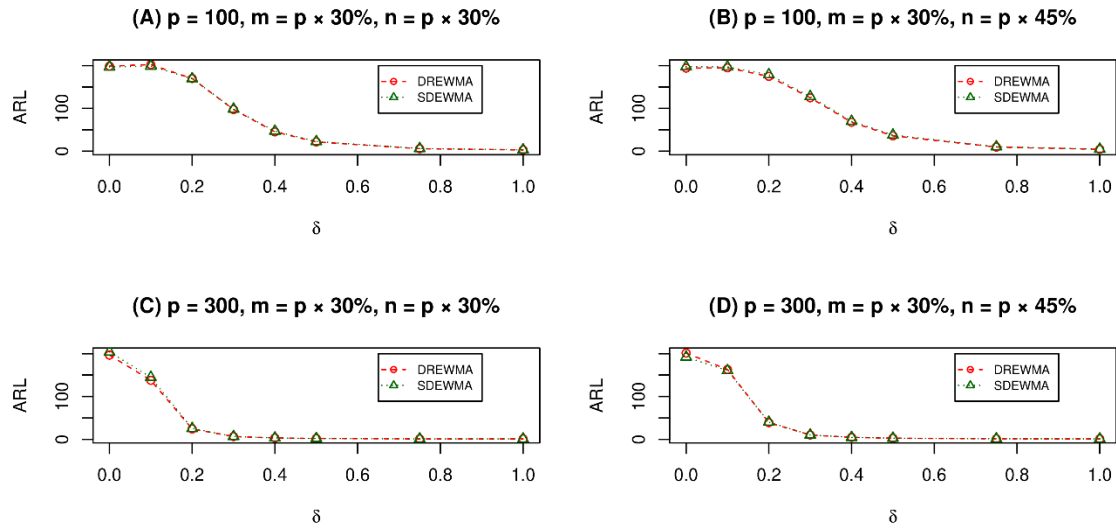


Figure 16: ARL comparison of DREWMA and SDEWMA charts with 4 combinations of dimensions and sample sizes for MVN_1

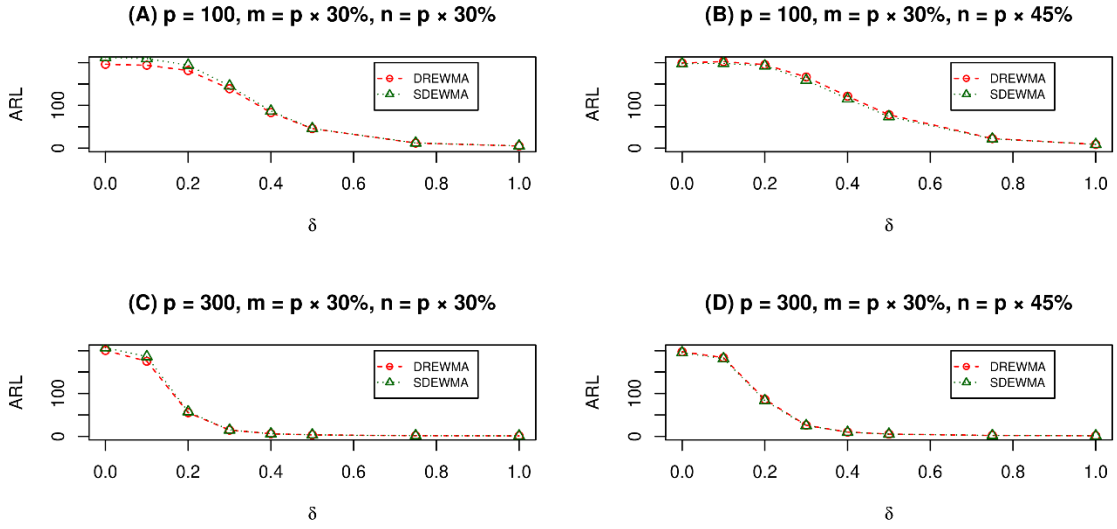


Figure 17: ARL comparison of DREWMA and SDEWMA charts with 4 combinations of dimensions and sample sizes for MVT

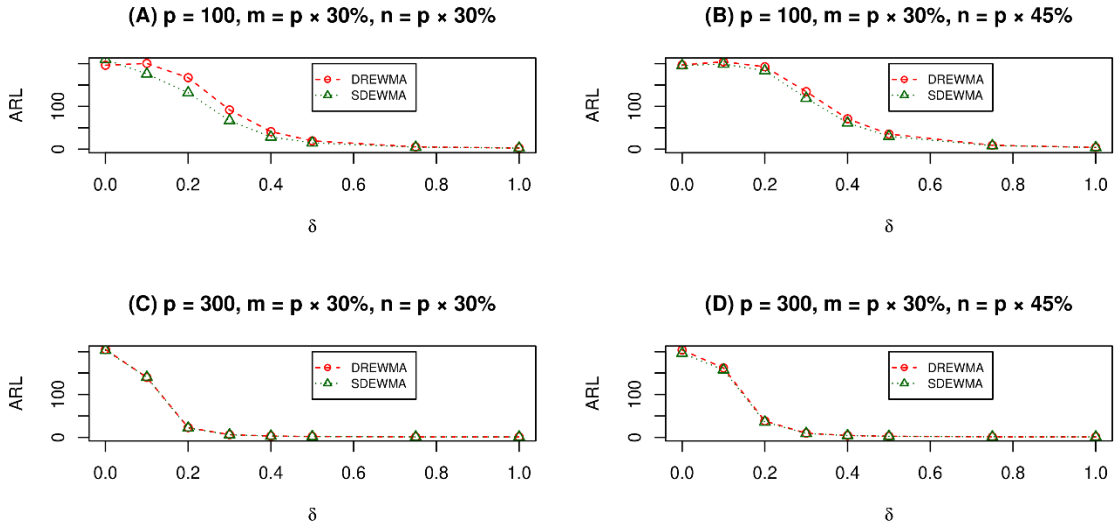


Figure 18: ARL comparison of DREWMA and SDEWMA charts with 4 combinations of dimensions and sample sizes for MVG

All the above figures compared the two EWMA based control charts for different combinations of dimensions, sample sizes, and distribution structures. Except a few of them, the ARL values of DREWMA and SDEWMA are closer to each other highlighting that their performance for the high-dimensional data is almost similar to each other. However, Table 8 tells the true difference between the proposed memory charts.

4.3.2 Performance Comparison of Memory Charts with Memoryless Charts

In this section, the performance of the proposed memory-based DREWMA and SDEWMA charts are compared with previously proposed memoryless DR and SD charts, as no such comparable chart is available in literature. The comparison is discussed below.

- The smallest values of ARL_1 are provided in bold font in Table 9. Overall, the performance of the memory-based DREWMA chart is better than the memoryless DR chart for detecting the shifts in the mean of second sample when the distribution structure is MVN_0 , for $p = 100$ ($m = n = p \times 30\%$). The only exception occurs for higher shift of $\delta = 1$, when the value of ARL_1 of DR chart is found to be smaller than that of DREWMA, which confirms that the DR chart as a Shewhart type chart may perform better for large shift. However, for all shifts ($\delta = 0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.75, 1$), with the dimension of $p = 100$ ($m = n = p \times 30\%$), and under distribution structure of MVN_0 , SDEWMA chart detects the shifts earlier than the SD chart which is under Shewhart environment.
- As the ARL_1 values of the memory charts are found to be smaller than that of the memoryless charts (cf. Table 9), the performance of both the DREWMA and SDEWMA charts are found to be better than the DR and SD charts, respectively under the distribution structure MVN_1 , with dimension $p = 100$ ($m = p \times 30\%$, $n = p \times 45\%$) for all shifts.
- With the dimension of $p = 300$ ($m = n = p \times 30\%$) under the MVT distribution structure, both DR and SD charts give smaller ARL_1 values than the memory charts for

higher shifts, such as $\delta = 0.75$ and 1. However, for small shifts such as $\delta = 0, 0.1, 0.2, 0.3, 0.4$, and 0.5, as expected, the memory charts – DREWMA and SDEWMA can detect a shift in the location of the second sample earlier than their memoryless counterparts (cf. Table 9).

- For the MVG distribution structure with high dimensions, such as $p = 300$, sample size of first sample $m = p \times 30\%$, and sample size of second sample $n = p \times 45\%$, the proposed memory charts perform better than the proposed memoryless charts for small shifts. While for large shifts, the memoryless charts perform either better or equal to the memory-based charts (cf. Table 9).

Table 9: Comparison of DR with DREWMA charts, and SD with SDEWMA charts

Sample Size	Dist.	δ	p = 100				Sample Size	Dist.	p = 300			
			DR	DREWMA	SD	SDEWMA			DR	DREWMA	SD	SDEWMA
$m = p \times 30\%, n = p \times 30\%$	MVN ₀	0	199.8 (196.1) 59,141,277	197.2 (194.3) 60,138,272	198.8 (194.7) 58,140,277	191.5 (188.4) 57,134,264	$m = p \times 30\%, n = p \times 30\%$	MVT	199.5 (197.4) 57,139,281	200.6 (195.3) 61,140,277	201.3 (205.4) 57,138,277	206.9 (203.3) 62,144,287
		0.1	202.7 (203.4) 58,140,280	196.9 (192.6) 58,139,273	201 (200.2) 58,141,278	189.2 (185.8) 57,133,259			199.5 (199.9) 59,140,275	175.4 (171.3) 53,124,238	196.7 (196.7) 57,135,271	186.4 (183.4) 56,130,256
		0.2	180.6 (180.4) 53,126,250	144.2 (139.1) 45,102,198	181.2 (179.9) 52,126,254	139.4 (132.6) 43,99,194			137.3 (138.7) 40,95,189	55.2 (49.8) 19,40,76	133.4 (133.1) 39,93,185	57.5 (52.5) 20,41,79
		0.3	136 (138.4) 40,93,185	64.7 (59.6) 22,46,89	139.8 (140.4) 40,97,194	63.9 (59.3) 21,46,88			54.7 (53.3) 16,38,77	14.5 (10.4) 7,12,19	53.3 (52) 16,37,74	14.7 (10.9) 7,12,19
		0.4	80.8 (79.5) 24,56,113	27 (22.6) 11,20,36	85.1 (85.1) 25,59,117	27.2 (23.2) 11,20,36			17 (16.4) 5,12,24	6 (3.1) 4,5,7	17 (16.1) 5,12,24	6 (3.1) 4,5,7
		0.5	42 (42.2) 13,29,58	12.7 (8.9) 6,10,17	45.9 (46.1) 14,31,64	13.4 (9.6) 7,11,17			5.5 (4.9) 2,4,7	3.4 (1.4) 2,3,4	5.7 (5.2) 2,4,8	3.4 (1.4) 2,3,4
		0.75	7.6 (7) 3,5,10	4.1 (1.8) 3,4,5	9.1 (8.6) 3,6,12	4.3 (2) 3,4,5			1.1 (0.4) 1,1,1	1.5 (0.5) 1,1,2	1.2 (0.5) 1,1,1	1.6 (0.6) 1,2,2
		1	2.1 (1.5) 1,2,3	2.2 (0.8) 2,2,3	2.6 (2.1) 1,2,3	2.4 (0.9) 2,2,3			1 (0) 1,1,1	1.1 (0.2) 1,1,1	1 (0.1) 1,1,1	1.1 (0.2) 1,1,1
$m = p \times 30\%, n = p \times 45\%$	MVN ₁	0	196.3 (198.3) 56,136,270	194.1 (190.4) 58,136,267	195.2 (195.3) 56,134,270	197.4 (193.6) 58,138,273	$m = p \times 30\%, n = p \times 45\%$	MVG	204.3 (203.2) 60,144,282	203.9 (203.9) 61,143,278	201.3 (202.1) 58,140,277	196.6 (196.1) 59,138,269
		0.1	199.5 (196.3) 58,142,277	194.7 (191.7) 60,136,269	200.4 (198.3) 59,139,275	197 (193.7) 60,138,269			202.5 (203.6) 57,138,284	162.2 (158.1) 50,113,225	200.5 (201.3) 59,139,277	157.6 (151.4) 50,111,215
		0.2	190 (189.6) 54,132,266	174.8 (170.6) 54,124,242	189.5 (189) 56,132,261	178.8 (174.7) 56,126,246			119.2 (119.1) 34,83,166	38.2 (33.3) 14,28,52	114.4 (116) 39,79,157	35.6 (31.1) 13,26,48
		0.3	177.7 (175.5) 52,125,246	124.8 (121.5) 39,87,171	171.3 (170) 50,120,239	127.7 (121.7) 40,90,177			34.6 (34) 10,24,48	9.5 (5.9) 5,8,12	32.3 (32.4) 9,22,44	9.1 (5.7) 5,8,12
		0.4	138.7 (139) 41,95,192	67.7 (63) 23,48,94	139.3 (136.2) 42,99,192	69.9 (65.6) 23,50,95			9.3 (8.8) 3,7,13	4.2 (1.8) 3,4,5	8.5 (8) 3,6,11	4.1 (1.8) 3,4,5
		0.5	99.2 (100) 29,68,136	35.9 (31.3) 13,27,49	101.3 (100.2) 29,71,140	37.9 (33.5) 14,28,51			3 (2.5) 1,2,4	2.6 (0.9) 2,2,3	2.9 (2.3) 1,2,4	2.5 (0.9) 2,2,3
		0.75	31.7 (31.4) 9,22,44	9.7 (6.3) 5,8,12	34.5 (34.1) 10,24,48	10.4 (7.1) 6,9,13			1 (0.1) 1,1,1	1.2 (0.4) 1,1,1	1 (0.1) 1,1,1	1.2 (0.4) 1,1,1
		1	9.7 (9.2) 3,7,13	4.4 (2.2) 3,4,5	11.6 (11.3) 4,8,16	4.8 (2.4) 3,4,6			1 (0) 1,1,1	1 (0) 1,1,1	1 (0) 1,1,1	1 (0) 1,1,1

The first row of each cell shows ARL (SDRL), and the second row of each cell shows the values of 25th, 50th, and 75th percentiles, respectively. Bold numbers mean the smallest ARL for each shift in different dimension sizes.

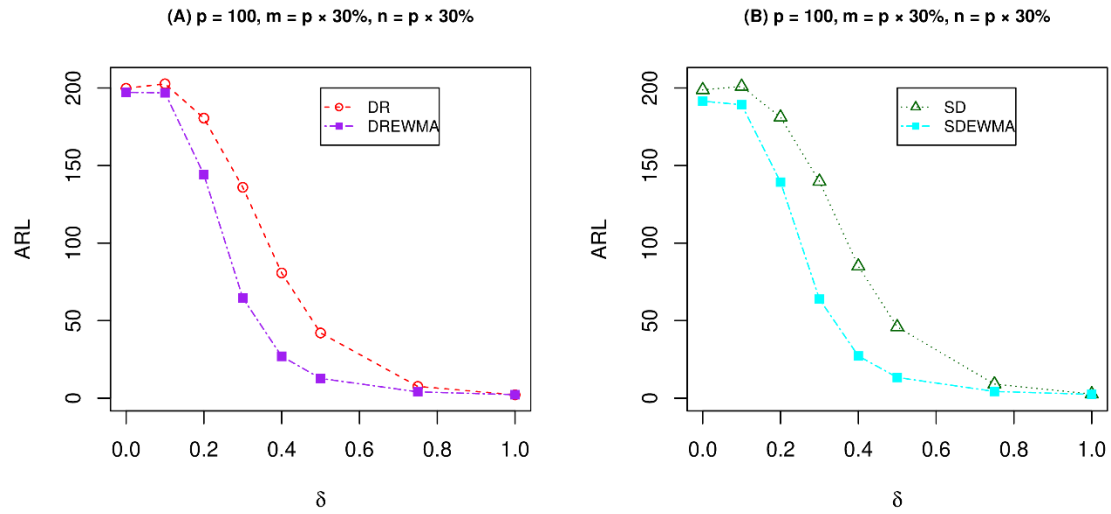


Figure 19: Comparison of (A) DR with DREWMA charts, and (B) SD with SDEWMA charts for MVN_0

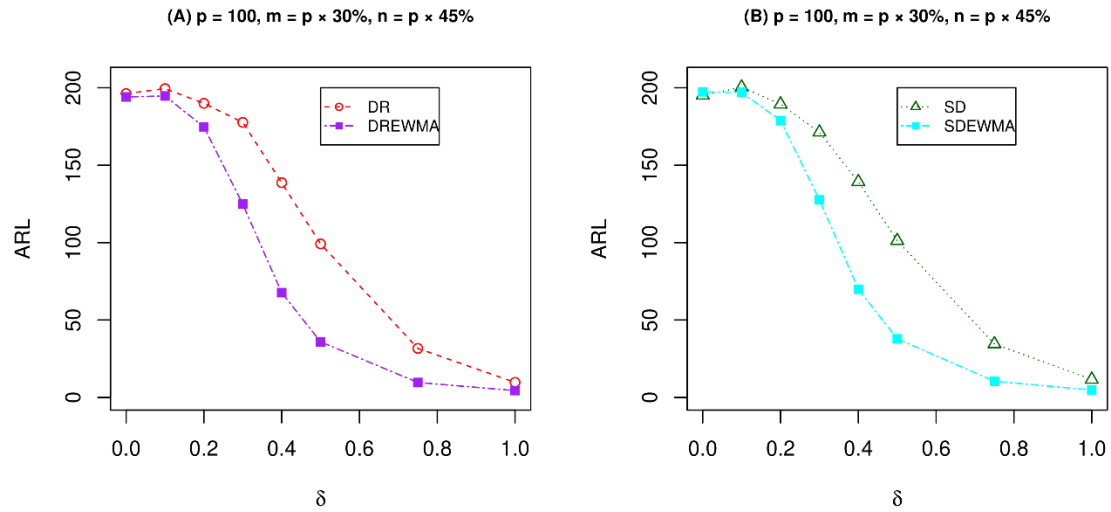


Figure 20: Comparison of (A) DR with DREWMA charts, and (B) SD with SDEWMA charts for MVN_1

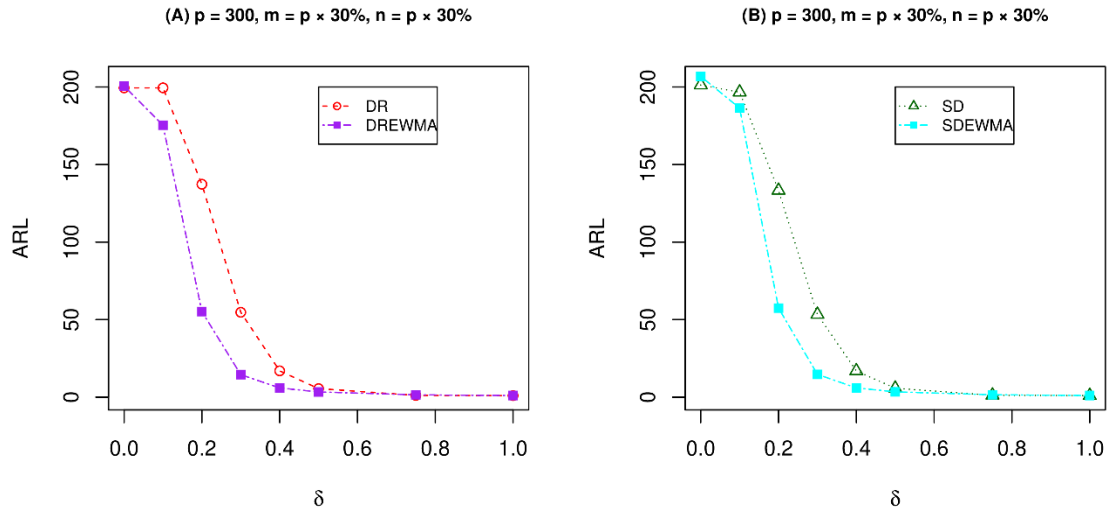


Figure 21: Comparison of (A) DR with DREWMA charts, and (B) SD with SDEWMA charts for MVT

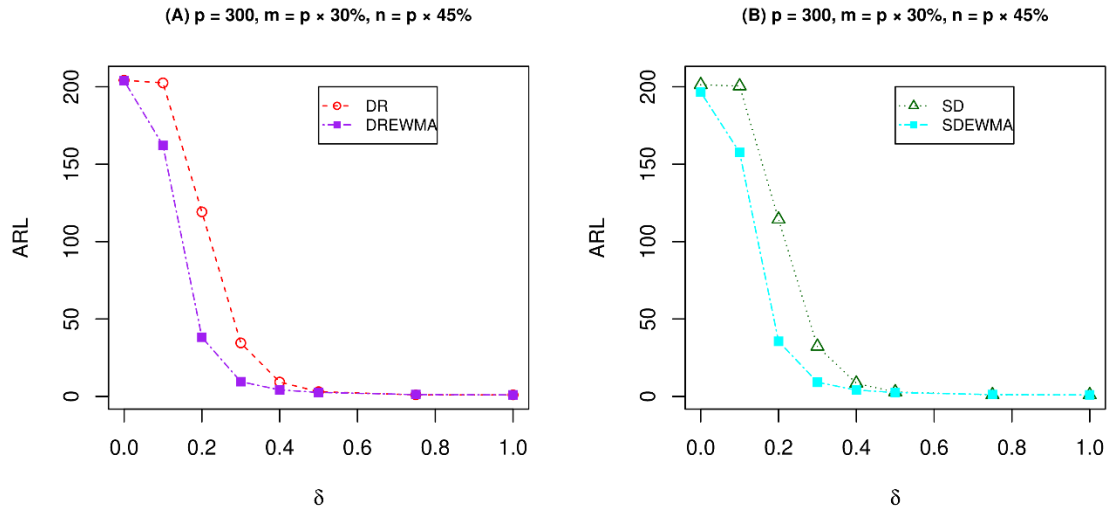


Figure 22: Comparison of (A) DR with DREWMA charts, and (B) SD with SDEWMA charts for MVG

For all the comparisons that are illustrated in Figure 19, Figure 20, Figure 21, and Figure 22, it is clear that the detection ability of the proposed memory charts under EWMA setup are far better than the memoryless charts.

4.4 Application of proposed Charts on Wind Turbine Bearings

Wind power has experienced substantial progress over the past decade and stands as the most auspicious source of renewable energy at present. The bearings – pitch bearing, main bearing, yaw bearing, gearbox bearing, and generator bearing – as shown in are subjected to constant variations in impact forces and loads within intricate wind and wave conditions, increasing the likelihood of mechanical failures [118]. The aforementioned bearings constitute essential components within large-scale wind turbines, serving a pivotal function in the transformation of the wind's kinetic energy into electrical power [112]. Damage to any of the bearings in wind turbines can be attributed to various factors such as high wind loads, imbalance, misalignment, inadequate lubrication of substances like grease, harsh environmental conditions, overloading, manufacturing defects, aging, excessive vibration, electrical fluctuations, and inadequate maintenance. Grease for wind turbines is a specialized lubricating compound formulated to meet the demanding requirements of wind turbine systems. It consists of three key constituents: a base oil, which serves as the primary lubricant and is chosen based on factors like temperature and load conditions; thickeners, which impart the semi-solid consistency to the grease and can include various types such as lithium and calcium; and additives, which enhance the grease's performance and protection characteristics through the inclusion of elements like antioxidants, anti-wear agents, and corrosion inhibitors. The primary function of grease is to maintain smooth, low-friction rotation, mitigate heat generation, and enhance the durability of essential wind turbine parts. Grease plays an integral role in preserving the reliability and longevity of critical components, contributing to the efficiency and performance of wind turbine systems. Grease-induced damage stands as a significant contributor to the degradation of

wind turbine bearings. This case study seeks to employ newly developed control charts to effectively identify irregularities in the main bearings of wind turbines attributed to grease-related issues.

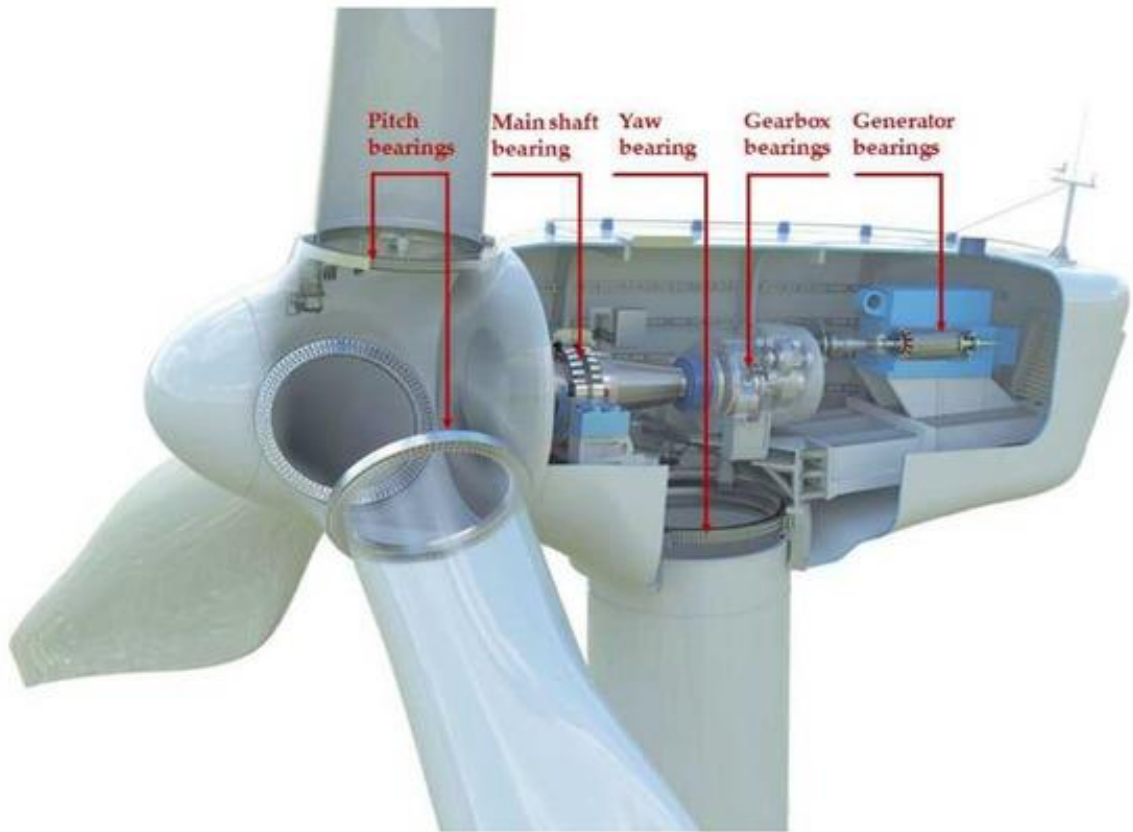


Figure 23: Wind turbine bearings [118].

In this section, provide practical insights into the real-world application of our proposed DREWMA, BSEWMA, and SDEWMA control charts for monitoring mean shifts in high-dimensional data involving two samples. To make it practical, a real dataset that measures grease damage over a 6-month period [115] has been chosen. This dataset, which serves as a great representation of high-dimensional data, can be easily accessed online through the

following

link:

<https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/J3IUVR>.

The dataset encompasses 25,920 rows designated for observations and comprises 123 columns representing distinctive features or dimensions. Each of these columns corresponds to an individual wind turbine number, collectively amounting to 123 wind turbines equipped with main bearings. Each row in the dataset records the measurement of true grease damage, taken at 10-minute intervals for all 123 turbines. As a result, this dataset encapsulates a comprehensive six-month record of observations for the entire array of 123 wind turbines, each featuring main bearings.

With the purpose to find out the changepoint from the total 25,920 row means, the following twelve models are fitted by employing “EnvCpt” R package version 1.1.3 [116]: (i) constant mean and variance, (ii) piecewise constant mean and variance, (iii) constant mean with AR (1) errors, (iv) constant mean with AR (2) errors, (v) piecewise constant mean with AR(1) errors, (vi) piecewise constant mean with AR(2) errors, (vii) linear trend over time, (viii) piecewise linear trend over time, (ix) linear trend over time with AR(1) errors, (x) linear trend over time with AR(2) errors, (xi) piecewise linear trend over time with AR(1) errors, and (xii) piecewise linear trend over time with AR(2) errors. These models provide that after the first 17,280 rows, that is, after the first 4 months of observations, the changepoint occurs. Hence the first four months’ data are considered as IC data. Then, considering a sample size of 60, a total of 288 samples are created from the IC data. Each sample is grouped into two sub-samples by considering equal sample size as $m = n = 30$. Considering $ARL_0 = 200$ and 1,000 iterations, the values of L_1 and L_2 for the proposed charts are obtained as followings: 2.705 and 3.053 for DREWMA, 2.69 and

3.056 for BSEWMA, and 3.42 and 1.67285 for SDEWMA. After that, the control limits of the proposed DREWMA, BSEWMA, and SDEWMA charts are found as (168.0295, 179.0841), (119.2460, 127.1227), and (22.8350, 35.1588), respectively. Then a total of 288 IC values based on the test statistics for each of the proposed charts are obtained from the high-dimensional data with two samples.

Out-of-control (OOC) data is generated to simulate a scenario where a mean shift occurred in the second sample while the first sample remained in a stable, in-control state. The OOC data for the first sample includes 8,640 observations from rows 8,641 – 17,280, which corresponds to observations made during the second and third months. The OOC data for the second sample also contains 8,640 observations, taken from rows indexed 17,281 to 25,920, covering observations from the final two months. Both samples have identical sample sizes, denoted as $m = n = 30$, resulting in a total sample size of 60.

The two-sample high-dimensional data obtained in such a way gives 288 OOC values, which serves as the basis for computing the test statistics for the DREWMA, BSEWMA, and SDEWMA charts. A total 576 test statistic values for these charts are presented in Figure 24, where half of the test statistics come from IC data, and the other half are from OOC data. The IC and OOC data are separated by a vertical line, as shown in Figure 24. In addition, the IC region is highlighted in pink color, while white-shaded area represents OOC region. Two horizontal green lines on each plot represent the UCL and LCL specific to the respective charts. When the values of the test statistics are in a state of control, they are denoted by blue points, while red points are employed to indicate out-of-control states. Two instances of false alarms are observed for SDEWMA chart, whereas, no false alarm is noticed for both the DREWMA and SDEWMA charts. The performance of the

DREWMA and BSEWMA charts are the same, as both charts detect 199 OOC values, while the SDEWMA chart detects 58 OOC values. Hence, compared to the SDEWMA chart, the DREWMA and BSEWMA charts are more efficient in studying the behavior of grease degradation in the bearing of wind turbines.

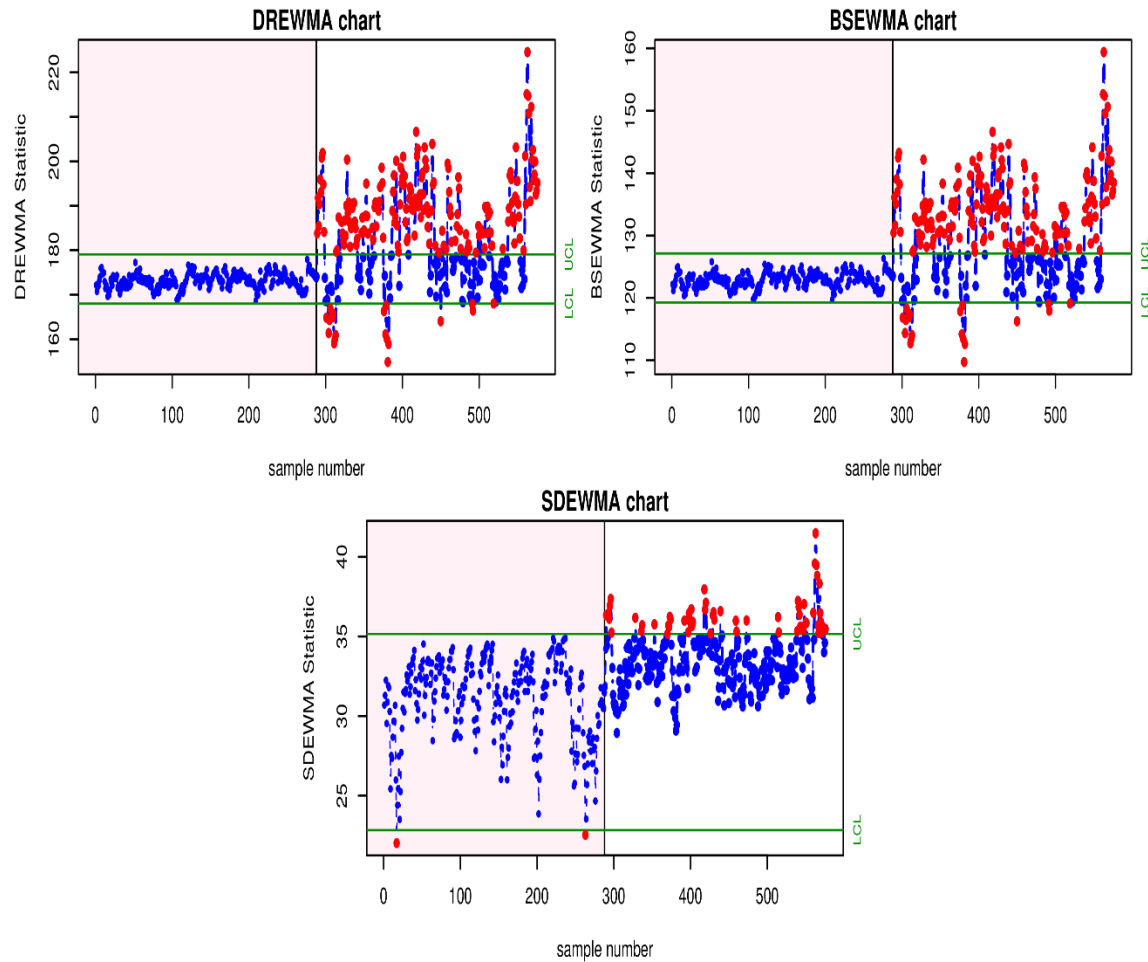


Figure 24: Graphical displays of plotting statistics for DREWMA, BSEWMA, and SDEWMA charts corresponding to their control limits.

CHAPTER 5

HIGH-DIMENSIONAL CONTROL CHARTS UNDER CUSUM ENVIRONMENT

The importance of using memory control charts over memoryless charts to monitor the shifts in high-dimensional process is significant to ensure that the historical information is incorporated by enabling the proper monitoring of the data behavior over time. Conventional memoryless chart under Shewhart setup is suitable for lower dimensions and higher shifts. When dimension increases and the target is to detect anomaly for small shifts, Exponentially Weighted Moving Average (EWMA) chart is more suitable. And Cumulative Sum (CUSUM) control chart performs even better than EWMA. Page [119] introduced the CUSUM chart for the first time in the literature. There are two ways to represent the CUSUM chart: (i) V-mask CUSUM, and (ii) tabular or algorithmic CUSUM. The second one is preferred. For the in-control stage of the process, let x_i follows normal distribution with mean μ_0 and standard deviation σ_x . Then the tabular or algorithmic CUSUM can be defined with the following two statistics: one-sided upper CUSUM, $C_i^+ = \max [0, (x_i - \mu_0) - K\sigma_x + C_{i-1}^+]$ and one-sided lower CUSUM, $C_i^- = \min [0, (x_i - \mu_0) + K\sigma_x + C_{i-1}^-]$, where the initial values are given by $C_0^+ = C_0^- = 0$, and K is known as the reference value usually taken as the one-half of the magnitude of shift we are interested to detect.

The Lower Control Limit (LCL) is given by: $LCL = -H = -h\sigma_x$, and the Upper Control Limit (UCL) is obtained by: $UCL = H = h\sigma_x$. The width of the control limit h can be obtained empirically based on the desired Average Run Length (ARL). For each sample, the two-sided CUSUM chart will plot both C_i^+ and C_i^- against the control limits. The CUSUM chart gives out-of-control signal, whenever either of the statistics falls outside the control limits.

In the literature, after applying dimension reduction techniques, the CUSUM chart was applied to monitor shifts in high-dimensional process. For example, to surveil the locations of high-dimensional data during the data gathering stage, reduced-dimension (RD) control charts using spatial scanning are proposed by [59]. CUSUM chart has also been frequently applied after the application of variable selection techniques that can reduce the size of the high-dimensional data. Variable selection methods aim to select a subset of important variables from a high-dimensional dataset with many variables. These methods are utilized to select relevant variables effectively, and then the CUSUM chart is implemented to monitor high-dimensional processes. For example, for mean monitoring, a variable-selection-based multivariate cumulative sum (VS-MCUSUM) chart was established by Abdella, et al. [95]; the chart was proved superior to the usual MCUSUM and Hotelling's T^2 charts.

Although the dimension reduction and variable selection methods are popular to deal with high-dimensional data, the problem of multicollinearity encourages the researchers to handle dimensional process without using either of the techniques. For example, to identify the shifts in online high-dimensional process, on the basis of TRAS (Top-r-based Adaptive Sampling) statistics, Liu, et al. [84] proposed a CUSUM chart. Zou, et al. [22] also used

the CUSUM scheme by synthesizing information from multiple data sources based on a goodness-of-fit (GOF) test. Li [3] proposed a two-stage CUSUM procedure that can maintain the Per-Comparison Error Rate (PCER). Furthermore, Moala, et al. [65] showed that the performance of MCUSUM control chart is slightly superior to that of the MEWMA chart for dealing with high-dimensional Spatial–Temporal Autoregressive Moving Average (STARMA) process.

These CUSUM based control charts predominantly focused on monitoring single sample high-dimensional data, leaving us the space to work on two-sample cases. Hence for two-sample high-dimensional data, new high-dimensional charts are proposed. The new charts are obtained by integrating Shewhart setup with CUSUM setup. The proposed charts are memory control charts that will be able to detect small shifts for two-sample high-dimensional data. The proposed memory control charts are named DRCUSUM, BSCUSUM, and SDCUSUM charts. As our main goal is to check if the high-dimensional control chart under CUSUM setup outperforms such chart under EWMA setup, we will evaluate and compare only DRCUSUM chart. Profile length study will be carried out to evaluate the proposed DRCUSUM chart. To validate the effectiveness of the proposed DRCUSUM chart for two-sample high-dimensional data, due to the lack of such research in literature for two-sample cases, our proposed chart will be compared with our previously proposed DR and DREWMA charts.

The remaining part of the chapter is as follows. In Section 5.1, the proposed two-sample high-dimensional charts are developed. Section 5.2 describes the designs required for the simulation study. In Section 5.3, the performance and the comparison of the proposed charts are presented.

5.1 Proposed Two-Sample High-dimensional Charts

We delve into the methodologies of the proposed charts in this section. Suppose, for a two-sample high-dimensional case, x_{ijk} , $i = 1, 2, \dots, m$ or $i = 1, 2, \dots, n$, $j = 1, 2, \dots, p$, and $k = 1, 2$ are the p -dimensional multivariate normal random vectors with mean vector $\mu = (\mu_1, \dots, \mu_p)'$ and unknown common positive variance-covariance matrix Σ . Here, sizes of first and second sample are denoted by m and n , respectively; p is the size of dimension and k is the sample number. The two-sample statistic T_1 [117] for the high-dimensional data can be defined as:

$$T_1 = \frac{\frac{mn}{m+n}(\bar{x}_1 - \bar{x}_2)' D_S^{-1}(\bar{x}_1 - \bar{x}_2) - \frac{(m+n)p}{m+n-2}}{\left[2 \left(\text{tr } R^2 - \frac{p^2}{m+n} \right) c_{p,m+n} \right]^{\frac{1}{2}}}, \quad (5.1)$$

where the diagonal matrix (D_S) of sample variances (S) is defined by

$$D_S = \text{diag} (s_{11}, \dots, s_{pp}), \quad (5.2)$$

The sample correlation matrix (R) is defined as follows:

$$R = D_S^{-\frac{1}{2}} S D_S^{-\frac{1}{2}}, \quad (5.3)$$

The adjustment coefficient $c_{p,m+n}$ can be chosen by

$$c_{p,m+n} = 1 + \frac{\text{tr } R^2}{p^{3/2}}. \quad (5.4)$$

Now the T_1 statistic obtained from two-sample high-dimensional data is integrated to two-sided CUSUM statistics as follows:

$$\begin{aligned} C_{ij(SDCUSUM)}^+ &= \max [0, (T_{1ij} - \mu_{0T_1}) - K \sigma_{T_1} + C_{ij-1(SDCUSUM)}^+] \\ C_{ij(SDCUSUM)}^- &= \min [0, (T_{1ij} - \mu_{0T_1}) + K \sigma_{T_1} + C_{ij-1(SDCUSUM)}^-] \end{aligned} \quad (5.5)$$

The lower control limit (LCL) and upper control limit (UCL) of the SDCUSUM chart for a given false alarm α , can be obtained from the following limits:

$$LCL_{(SDCUSUM)} = -H = -h \sigma_{T_1} \quad (5.6)$$

$$UCL_{(SDCUSUM)} = H = h \sigma_{T_1} \quad (5.7)$$

where, μ_{0T_1} and σ_{T_1} are the mean and standard deviation of the statistic T_1 . The value of width of control limit h can be determined empirically, and the reference value K is usually chosen as the half of the mean shift of the process. It is to be noted that the test statistic T_1 was proposed by Srivastava and Du [117]. Therefore, the control chart based on statistic T_1 and CUSUM is named as SDCUSUM chart.

The control chart on the basis of the statistic T_2 [106, 107] and two-sided CUSUM is named as DRCUSUM chart. The statistic T_2 can be expressed as follows:

$$T_2 = \frac{\frac{mn}{m+n} (\bar{x}_1 - \bar{x}_2)' (\bar{x}_1 - \bar{x}_2)}{\text{tr } S}. \quad (5.8)$$

The statistics for DRCUSUM chart are as follows:

$$\begin{aligned} C_{ij(DRCUSUM)}^+ &= \max [0, (T_{2ij} - \mu_{0T_2}) - K \sigma_{T_2} + C_{ij-1(DRCUSUM)}^+] \\ C_{ij(DRCUSUM)}^- &= \min [0, (T_{2ij} - \mu_{0T_2}) + K \sigma_{T_2} + C_{ij-1(DRCUSUM)}^-] \end{aligned} \quad (5.9)$$

The lower control limit (LCL) and upper control limit (UCL) of the DRCUSUM chart for a given false alarm α , can be obtained from the following limits:

$$LCL_{(DRCUSUM)} = -H = -h \sigma_{T_2} \quad (5.10)$$

$$UCL_{(DRCUSUM)} = H = h \sigma_{T_2} \quad (5.11)$$

where, μ_{0T_2} and σ_{T_2} are the mean and standard deviation of the statistic T_2 . The value of the limit width h can be determined empirically, and K is set to half of the difference between target mean and shifted mean.

The control chart on the basis of the statistic T_3 [108] and two-sided CUSUM is named as BSCUSUM chart. The statistic T_3 is given by:

$$T_3 = \frac{\frac{mn}{m+n} (\bar{x}_1 - \bar{x}_2)' (\bar{x}_1 - \bar{x}_2) - \text{tr } S}{\left[\frac{2(m+n)(m+n+1)}{(m+n-1)(m+n+2)} \left(\text{tr } S^2 - \frac{(\text{tr } S)^2}{m+n} \right) \right]^{\frac{1}{2}}} \quad (5.12)$$

The statistics for BSCUSUM chart are as follows:

$$\begin{aligned} C_{ij(BSCUSUM)}^+ &= \max [0, (T_{3ij} - \mu_{0T_3}) - K \sigma_{T_3} + C_{ij-1(BSCUSUM)}^+] \\ C_{ij(BSCUSUM)}^- &= \min [0, (T_{3ij} - \mu_{0T_3}) + K \sigma_{T_3} + C_{ij-1(BSCUSUM)}^-] \end{aligned} \quad (5.13)$$

The lower control limit (LCL) and upper control limit (UCL) of the BSCUSUM chart for a given false alarm α , can be obtained as follows:

$$LCL_{(BSCUSUM)} = -H = -h \sigma_{T_3} \quad (5.14)$$

$$UCL_{(BSCUSUM)} = H = h \sigma_{T_3} \quad (5.15)$$

where, μ_{0T_3} and σ_{T_3} are the mean and standard deviation of the statistic T_3 . The width of control limit h can be determined empirically, and the reference value K is usually chosen as half of the shifts in the process that we are interested to get early detection.

It is to be noted that a particular chart among any of the proposed three charts, named SDCUSUM, DRCUSUM, and BSCUSUM is known to be in-control (IC) stage or stable situation if both the statistics for that chart fall inside the control limits. Whereas the chart signals an out-of-control (OOC) point when either of its statistics falls outside of their control limits.

5.2 Design of Simulation Study

The simulation for this study is conducted in a three-step process. Using 100,000 iterations, the values of means and standard deviations of T_1 , T_2 , and T_3 statistics are obtained under different distributions and dependency structures in the initial step. Subsequently, in the second step, using predefined $ARL_0 = 200$ under different distributions and dependency structures, the values of h are found using 5,000 iterations for the pre-obtained means and standard deviations. Finally, considering the values of means, standard deviations, and h for different combinations of distributions and dependencies, the third step involves the study of ARL_1 based on 5,000 iterations. The following combinations of distributions and dependencies [120] are taken into account: (i) multivariate normal with unit variances and covariances equal to 0, denoted as MVN_0 ; (ii) multivariate normal with unit variances and covariances equal to 0.5, denoted as MVN_1 ; (iii) multivariate t with 5 degrees of freedom, denoted as MVT ; (iv) multivariate gamma with shape parameter 5 and scale parameter 1, denoted as MVG. Unit variances and covariances are assumed equal to 0.5 for both MVT and MVG distributions.

The size of high dimensions is set to $p = 100$. The sample size for the two samples is considered as $m = p \times 30\%$, $n = p \times 30\%$. A simulation study of 100,000 iterations has been performed to provide means and standard deviations and then h using 10000 iterations under the distribution cases (i) – (iv) for the aforementioned sample size for the proposed DRCUSUM chart; these values are reported in Table 10. Finally, the run length profile of the proposed chart based on $ARL_0 = 200$, shifts ($\delta = 0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.75, 1$), and distribution cases (i) – (iv) have been provided in Table 11.

Moreover, for testing the means of two-sample high-dimensional data, Bai and Saranadasa [108] showed that the T_3 and T_1 statistics have the similar performance. Similarly, we found in initial findings that the BSCUSUM chart has the same run length profile as the DRCUSUM chart. Therefore, for brevity, we have excluded the results of the BSCUSUM chart. The results of SDEWMA charts are also not considered, as the focus of the study is to explore whether the integration of CUSUM chart to either of the T_1 , T_2 and T_3 statistic outperforms EWMA chart. Hence, only the performance of the DRCUSUM chart is evaluated. The proposed chart is constructed based on $K = 0.5$ under the CUSUM environment. Since, in the literature there is no control chart available to detect shifts in the second sample of two-sample high-dimensional data, the comparison of the proposed chart has been made with our previously proposed DR and DREWMA chart; the comparison is provided in Table 9. The results are also graphically represented at the end.

5.3 Results and Discussions

Average Run Length (ARL) is the mean number of points required to identify an out-of-control (OOC) signal in a process. Whereas the term run length indicates the consecutive points observed before detecting the out-of-control (OOC) state of a process. The ARL for the two-sided CUSUM chart can be obtained using the following equation: $\frac{1}{ARL} = \frac{1}{ARL^+} + \frac{1}{ARL^-}$, where ARL^+ is the average run length of the one-sided upper CUSUM chart, and ARL^- denotes the average run length of the one-sided lower CUSUM chart [121]. When the process is in a stable state the ARL can be termed as ARL_0 , while the term ARL is denoted by ARL_1 if the process is not in a stable state. Commencing with a predefined value of ARL_0 , lower values of ARL_1 is an indication that a particular control chart can

detect shift early. To ensure earlier detection of shifts, it is desired to obtain a lower ARL_1 values. Hence the performance of the proposed charts is mainly evaluated based on the ARL values. However, alongside the values of ARL, other measures such as Standard Deviation of Run Length (SDRL), Q_1 (1st quartile or 25th percentile), Q_2 (2nd quartile or median or 50th percentile), and Q_3 (3rd quartile or 75th percentile) of run length are also included.

5.3.1 Performance of the Proposed Charts

This study assesses the performance of the proposed chart based on run length profile for high-dimensional data with the dimension $p = 100$.

- The mean and standard deviation, and h values for the proposed chart for $ARL_0 = 200$ under the distribution cases (i) – (iv) for the sample size of $m = n = p \times 30\%$ are provided in Table 10. For example, for $ARL_0 = 200$, with $p = 100$ ($m = 30, n = 30$), distribution = MVN_0 , the mean and standard deviation, and h value of DRCUSUM chart are 0.00020, 1.00005, and 4.24000, respectively. In the same way, with $p = 100$ ($m = 30, n = 30$), distribution = MVT , the mean and standard deviation, and h value of DRCUSUM chart are -0.00023, 0.72113, and 4.32000, respectively.
- Under the distribution structure MVN_0 , for $p = 100$ ($m = n = p \times 30\%$) DRCUSUM chart has shown better detection ability than the other distribution structures (shown in Table 11), as its values of ARL_1 for different shifts are least than that of the other distribution structures. For initial shift such as $\delta = 0.1$, the ARL_1 value of the distribution structure MVN_0 is slightly higher than the distribution

structures MVN_1 and MVG. However, in general for large shifts the DRCUSUM chart has the earliest detection ability under the distribution structure MVN_0 compared to the other distributions.

- For all shifts, under the distribution structure of MVG, with the dimension $p = 100$ ($m = p \times 30\%$, $n = p \times 30\%$), the DRCUSUM chart provides smaller ARL_1 values than the average run length values of the distribution structure of MVN_1 and MVT (cf. Table 11).
- The performance of DRCUSUM chart under the distribution structure MVN_1 is better than the distribution structure of MVT for all shifts with the dimension of $p = 100$ and sample size of $m = 30\%$, $n = 30\%$ (cf. Table 11).

Table 10: Mean, standard deviation, and h values for DRCUSUM chart with $m = n = p \times 30\%$

Distribution	$ARL_0 = 200$	$p = 100$
		DRCUSUM
MVN_0	μ_0	0.00020
	σ_0	1.00005
	h	4.24000
MVN_1	μ_0	0.00200
	σ_0	1.00057
	h	4.28000
MVT	μ_0	0.00023
	σ_0	0.72113
	h	4.32000
MVG	μ_0	-0.00058
	σ_0	0.94721
	h	4.28000

Table 11: Run length profile of DRCUSUM based on ARL (SDRL) with $m = n = p \times 30\%$

Dist.	δ	p = 100	Dist.	p = 100
		DRCUSUM		DRCUSUM
MVN ₀	0	201.5 (189.9) 63,147,276	MVT	199.7 (196.3) 59,137,267
	0.1	184.2 (176.4) 59,131,252		189.2 (187.1) 57,133,259
	0.2	126.4 (118.9) 41,91,172		159.5 (152.6) 49,113,226
	0.3	58.3 (54.6) 20,42,79		109.9 (106.1) 34,77,148
	0.4	24.3 (19.9) 10,18,32		66.7 (61.6) 22,48,92
	0.5	11.9 (8.1) 6,10,15		38.4 (34.2) 14,28,52
	0.75	4.1 (1.8) 3,4,5		10.7 (7.1) 6,9,14
	1	2.3 (0.7) 2,2,3		4.9 (2.4) 3,4,6
MVN ₁	0	201.5 (199.8) 62,141,276	MVG	199.7 (197.2) 60,139,275
	0.1	181.8 (177) 56,128,250		181.6 (176.8) 55,127,253
	0.2	140.9 (132.2) 46,103,195		134.2 (128.4) 43,95,183
	0.3	76.3 (71.2) 25,56,105		73.1 (68.5) 24,53,100
	0.4	37.6 (32.9) 14,28,50		34.2 (28.9) 13,25,46
	0.5	18.5 (14.3) 8,15,24		17.1 (13.1) 8,13,22
	0.75	5.6 (2.8) 4,5,7		5.3 (2.5) 3,5,7
	1	2.9 (1.1) 2,3,3		2.8 (1) 2,3,3

The first row of each cell shows ARL (SDRL), second row of each cell shows the values of 25th, 50th, and 75th percentiles, respectively. Bold numbers mean the smallest ARL for each shift in different dimension size.

5.3.2 Performance Comparison between DR, DREWMA, and DRCUSUM Charts

At this stage of the study, the focus is to check if the CUSUM integration to the DR chart outperforms the EWMA integration to the DR chart. Hence, the comparison of the proposed DRCUSUM chart will be compared with the previously proposed DREWMA and DR charts. The comparison is discussed below.

- In general, the performance of the memory-based CUSUM chart is better than both the other memory-based DREWMA chart and the memoryless DR chart for detecting the shifts in the mean of the second sample for all the distribution structures with the dimension size $p = 100$, and sample sizes $m = n = p \times 30\%$. The only exception occurs under the distribution structure MVN_0 for higher shift of $\delta = 1$, when the value of ARL_1 of DR chart is found to be smallest than that of DREWMA and DRCUSUM charts, which confirms that the DR chart as a Shewhart type chart may perform better for large shifts. In Table 12, the smallest values of ARL_1 are provided in bold font.
- Under the distribution structure MVN_1 , the performance of both the DRCUSUM and DREWMA charts are found to be better than the DR chart, as the ARL_1 values of the memory charts are found to be smaller than that of the memoryless charts (cf. Table 12), with dimension $p = 100$ ($m = n = p \times 30\%$) for all shifts. When compared to DRCUSUM with DREWMA chart, it is clearly seen that the performance of the DRCUSUM is superior to DREWMA chart.
- For all shifts ($\delta = 0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.75, 1$), with the dimension of $p = 100$ ($m = n = p \times 30\%$), and under distribution structure of MVT and MVG, the memory chart DREWMA outperforms the memoryless DR chart. Whereas the

proposed memory chart DRCUSUM of this study clearly shows better performance than both the previously proposed DREWMA and DR charts (cf. Table 12).

Table 12: Comparison between DR, DREWMA, and DRCUSUM charts for $m = n = p \times 30\%$

Dist.	δ	p = 100			Dist.	p = 100		
		DR	DREWMA	DRCUSUM		DR	DREWMA	DRCUSUM
MVN ₀	0	199.8 (196.1) 59,141,277	197.2 (194.3) 60,138,272	201.5 (189.9) 63,147,276	MVT	198.5 (197.6) 58,138,276	196 (195.1) 59,134,268	199.7 (196.3) 59,137,267
	0.1	202.7 (203.4) 58,140,280	196.9 (192.6) 58,139,273	184.2 (176.4) 59,131,252		195.9 (195.9) 55,136,273	193.8 (186.6) 60,138,266	189.2 (187.1) 57,133,259
	0.2	180.6 (180.4) 53,126,250	144.2 (139.1) 45,102,198	126.4 (118.9) 41,91,172		191.2 (189.7) 55,135,266	181.5 (179.6) 55,125,250	159.5 (152.6) 49,113,226
	0.3	136 (138.4) 40,93,185	64.7 (59.6) 22,46,89	58.3 (54.6) 20,42,79		185.5 (185.4) 52,127,259	139 (133.4) 43,98,193	109.9 (106.1) 34,77,148
	0.4	80.8 (79.5) 24,56,113	27 (22.6) 11,20,36	24.3 (19.9) 10,18,32		154.3 (151.5) 46,108,215	83.2 (78.1) 27,59,114	66.7 (61.6) 22,48,92
	0.5	42 (42.2) 13,29,58	12.7 (8.9) 6,10,17	11.9 (8.1) 6,10,15		119.2 (116.8) 35,83,167	45.8 (41.2) 17,34,62	38.4 (34.2) 14,28,52
	0.75	7.6 (7) 3,5,10	4.1 (1.8) 3,4,5	4.1 (1.8) 3,4,5		42 (41.2) 13,29,59	12 (8.3) 6,10,15	10.7 (7.1) 6,9,14
	1	2.1 (1.5) 1,2,3	2.2 (0.8) 2,2,3	2.3 (0.7) 2,2,3		12.8 (12.1) 4,9,17	5.2 (2.7) 3,5,6	4.9 (2.4) 3,4,6
MVN ₁	0	197.6 (197.3) 58,138,272	198.8 (197.3) 61,137,272	201.5 (199.8) 62,141,276	MVG	195 (196.9) 56,134,270	196.1 (189.7) 61,137,273	199.7 (197.2) 60,139,275
	0.1	199.1 (195.2) 59,138,280	202.3 (198.9) 61,142,275	181.8 (177) 56,128,250		198.5 (197.7) 57,138,278	200.4 (198.8) 60,138,278	181.6 (176.8) 55,127,253
	0.2	190.8 (190.2) 55,131,264	169.8 (165.7) 53,119,234	140.9 (132.2) 46,103,195		188.8 (191.5) 54,129,261	167.1 (164.6) 50,117,229	134.2 (128.4) 43,95,183
	0.3	168.7 (164.7) 48,119,236	97.4 (92) 31,70,134	76.3 (71.2) 25,56,105		163.6 (164.1) 47,114,225	91.8 (87.4) 29,65,126	73.1 (68.5) 24,53,100
	0.4	116.8 (118) 34,81,159	45.2 (40.1) 17,33,61	37.6 (32.9) 14,28,50		115.3 (115.2) 33,79,159	41.1 (36.7) 15,30,55	34.2 (28.9) 13,25,46
	0.5	72.8 (73.7) 21,50,99	21.7 (17) 10,17,29	18.5 (14.3) 8,15,24		68.8 (68.2) 21,48,95	19.5 (15.2) 9,15,26	17.1 (13.1) 8,13,22
	0.75	17.5 (17) 5,12,24	5.9 (3.1) 4,5,7	5.6 (2.8) 4,5,7		15.2 (14.9) 5,11,21	5.5 (2.8) 3,5,7	5.3 (2.5) 3,5,7
	1	4.6 (4.1) 2,3,6	3 (1.1) 2,3,4	2.9 (1.1) 2,3,3		3.9 (3.4) 2,3,5	2.8 (1) 2,3,3	2.8 (1) 2,3,3

The first row of each cell shows ARL (SDRL), second row of each cell shows the values of 25th, 50th, and 75th percentiles, respectively. Bold numbers mean the smallest ARL for each shift in different dimension size.

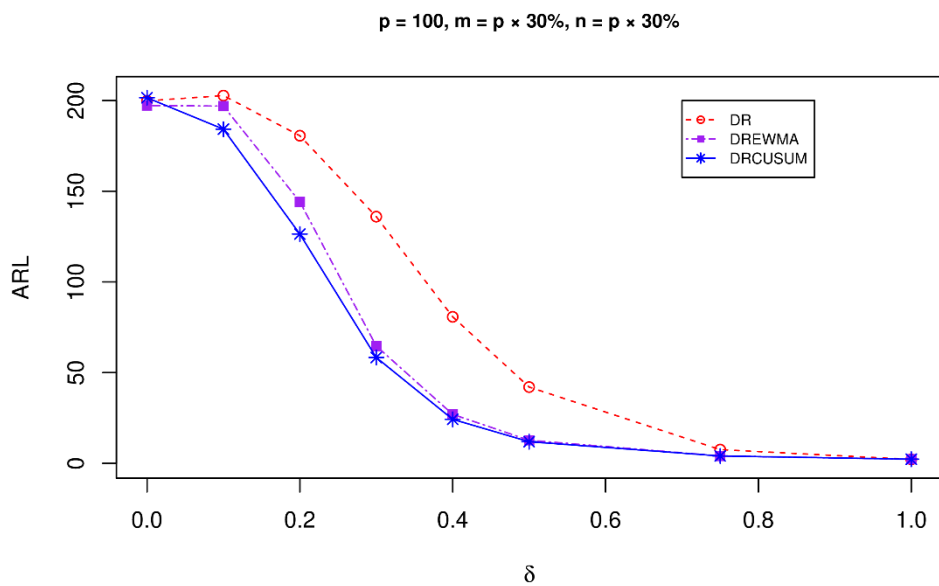


Figure 25: Comparison between DR, DREWMA, and DRCUSUM charts for MVN_0

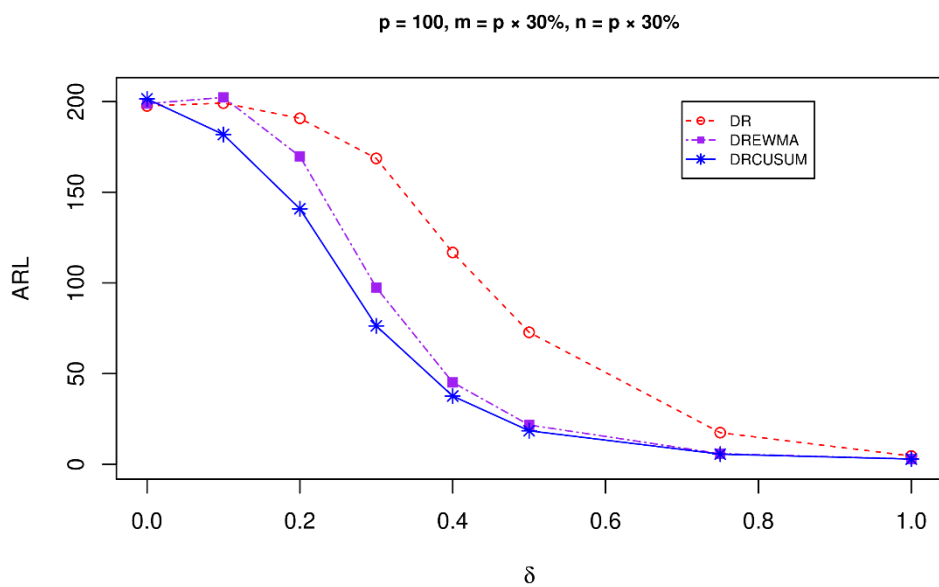


Figure 26: Comparison between DR, DREWMA, and DRCUSUM charts for MVN_1

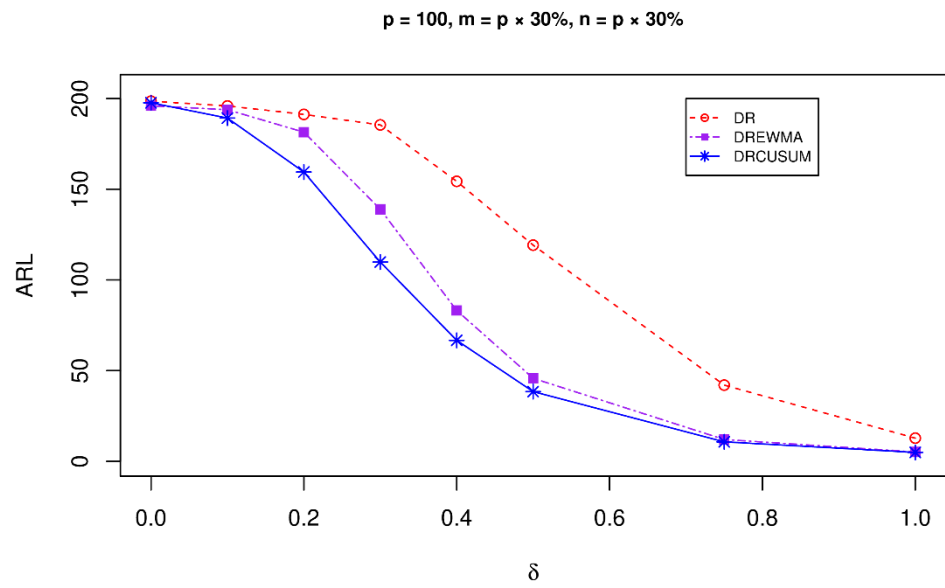


Figure 27: Comparison between DR, DREWMA, and DRCUSUM charts for MVT

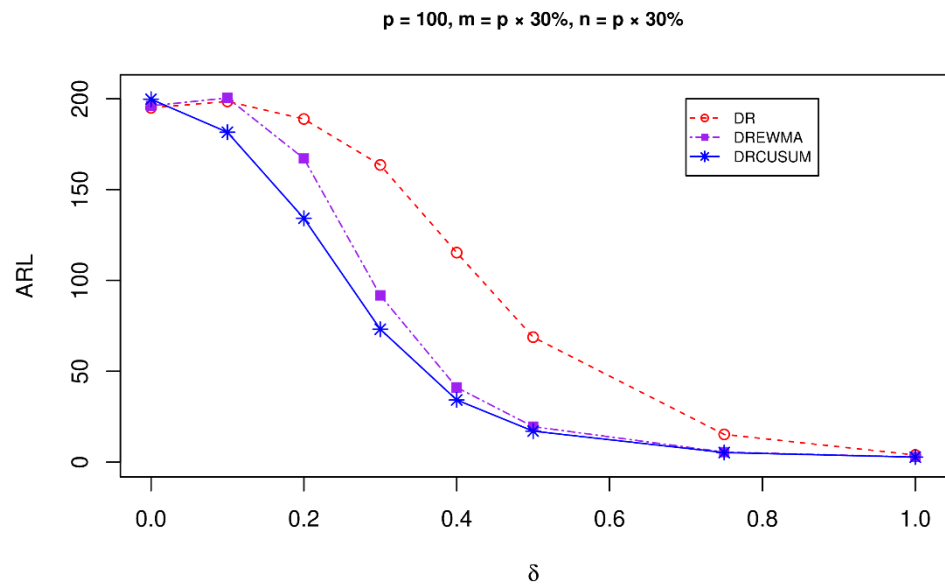


Figure 28: Comparison between DR, DREWMA, and DRCUSUM charts for MVG

The performance of the proposed memory charts are far better than the memoryless charts, which is clearly noticed from the above illustration in Figure 25, Figure 26, Figure 27, and Figure 28. In addition, if there is a comparison between the memory charts, DRCUSUM chart outperforms the DREWMA chart. When the comparison of performance is made between the DREWMA and DR charts for different combinations of distribution structures, the DREWMA chart is more efficient than the DR chart. Overall, for detecting shifts in the second sample of the two-sample high-dimensional data, the DRCUSUM chart outperforms both the DREWMA and DR charts regardless of the structures of the distribution.

CHAPTER 6

SUMMARY, CONCLUSIONS AND RECOMMENDATIONS

The term “high-dimensional data” refers to a dataset comprising many variables or features, usually more than the number of observations. This study provides a detailed overview of high-dimensional monitoring methods using a sample of 82 relevant articles published between 2004 and 2023. Three categories can be found in the literature on high-dimensional control charts: control charts based on the dimension reduction approach, variable selection method, and high-dimensional techniques. In addition, the literature on each approach is categorized by control chart structures such as memory-less (Hotelling T^2 chart), memory-based (MEWMA and MCUSUM charts), and others. Real-life examples of high-dimensional datasets utilized in various fields, such as industry, medical science, chemical engineering, and image processing, are also presented. This study found that research on high-dimensional control charts increased significantly after 2016, and most studies aimed to develop high-dimensional control charts using high-dimensional techniques. Furthermore, most studies introduced memory-based (MEWMA and MCUSUM) charts and other structures to monitor high-dimensional data.

6.1 Concluding Remarks Based on Shewahrt Setup

High-dimensional data are those datasets where the number of features is more than the number of observations. As a result of progress in sensor technologies, automation, and the ability to gather data, high-dimensional data is becoming more widespread in modern manufacturing processes. To study the behavior of one-sample high-dimensional data, a good number of works have been carried out in the literature of SPC. The analysis is more challenging when the manufacturing processes become gradually complex and data-intensive, resulting in two-sample high-dimensional data in some cases. No research has been conducted till now to study the behavior of two-sample high-dimensional data using control charts. This study discusses applications of control charts on both two-sample low-dimensional and high-dimensional data. For low-dimensional data ($p = 10, 30, 50$), equal sample size ($m = 30, n = 30$) has been considered for both the samples. For high-dimensional data ($p = 100, 300, 500$), two combinations of sample sizes ($m = 30\%, n = 30\%$; $m = 30\%, n = 45\%$) are considered. For both two-sample low-dimensional and high-dimensional data, a simulation of 1000000 is used to find out the control limits of the charts under four combinations of distributions – MVN_0 , MVN_1 , MVT, and MVG based on pre-specified ARL_0 such as 168, 200, 370, and 500. However, only $ARL_0 = 200$ is used to obtain the run length profile of all the charts with the simulation of 10000. Although several measurements such as ARL, SDRL, Q_1 (1st quartile), Q_2 (2nd quartile or median), Q_3 (3rd quartile) are used as measure of performance evaluation, only ARL is used to compare the performance of the charts. Conventional Hotelling's T^2 chart performed well only for MVN_0 distribution with $p = 10$. For all the cases of low-dimensional data, our proposed DR and SD charts show better detection ability. To

compare the performance between our proposed charts under the framework of low-dimension, it is noticed that SD chart performed better when $p = 50$ under MVT and MVG distributions; DR chart performs better in other scenarios of low-dimensional data. It is found that two-sample Hotelling's T^2 chart can not be implemented for higher dimensions, e.g., $p > 60$. Hence, for higher dimensions, e.g., $p = 100$ and 300 , our proposed charts overcome the limitation of Hotelling's T^2 chart. Under the high-dimensional setup, DR chart is quick in detecting anomaly compared to SD chart for MVN_0 and MVN_1 distributions, while SD chart is found superior in performance for MVT and MVG distributions. It is to be mentioned that for another proposed chart, BS chart, only the control limits are provided. BS chart is kept out of the run length profile performance comparison as it provides the same performance as the DR chart. A case study has also been carried out for two-sample high-dimensional data, where it has been found that both the DR and BS charts perform the same to detect shifts in means of the second sample.

6.2 Concluding Remarks Based on EWMA Setup

In the area of modern manufacturing, high-dimensional data, characterized by having many features relative to observations, is on the rise due to advances in technology and data collection. Researchers have delved into the analysis of one-sample high-dimensional data, particularly in the domain of SPC. Nevertheless, as manufacturing processes become more intricate and data-intensive, two-sample high-dimensional data situations are emerging, presenting new challenges. Surprisingly, there is no research on the behavior of two-sample high-dimensional data using control charts yet. This gap is significant, given the increasing complexity of modern manufacturing. Understanding and effectively monitoring such types of data is vital, making it a promising area for further exploration. In this study, three

new memory-based control charts – DREWMA, BSEWMA, and SDEWMA are proposed. These charts are also compared with our previously proposed memoryless charts, such as DR, BS, and SD charts. In the current study, two types of high-dimensional data with $p = 100$ and 300 are considered with two different sample sizes ($m = p \times 30\%$, $n = p \times 30\%$; $m = p \times 30\%$, $n = p \times 45\%$) for each of the dimensions. The simulation study involved three sequential steps. First, the mean and standard deviation values of DR, BS, and SD charts are obtained across multiple distributions (MVN_0 , MVN_1 , MVT, and MVG) and dependency structures (covariance = 0 or 0.5; variance = 1), utilizing 100,000 iterations. In the second phase, L_1 and L_2 values for the proposed DREWMA and SDEWMA charts are calculated based on the pre-established means and standard deviations, along with a predefined $ARL_0 = 200$, using 10,000 iterations across various distribution patterns and dependency structures, as mentioned above. Since both the DREWMA and BSEWMA charts perform the same that is also proved with a real-life application, for brevity, the BSEWMA chart is not used in the simulation study. The final step focused on investigating ARL_1 behavior for the proposed charts with different shifts such as $\delta = 0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.75$, and 1 in the mean of the second sample of the two-sample high-dimensional data. It has been noticed that, generally, the detection ability of the DREWMA chart is better than the SDEWMA chart when the data comes from multivariate normal distribution regardless of dependency structure. Whereas the SDEWMA provides more satisfactory performance than DREWMA when the data follows multivariate t or gamma distribution with dependency structure. This study also compares the proposed memory-based charts, such as DREWMA and SDEWMA charts, with previously established memoryless DR and SD charts. It is found that the memory-based

charts perform with superiority over the memoryless control charts. In the end, all three memory charts – DREWMA, BSEWMA, and SDEWMA are applied to real-life data of grease damage on the main bearings of wind turbines. The real-life application shows that DREWMA and BSEWMA charts perform the same. Compared to the SDEWMA chart, both DREWMA and BSEWMA charts show superior performance in detecting shifts in the second sample of grease damage on wind turbines' main bearings.

6.3 Concluding Remarks Based on CUSUM Setup

The application of memory-based control charts to deal with the high-dimensional data, especially for two-sample high-dimensional data can play significant role by addressing the challenges associated with such data to ensure quality control across different domains including the field of manufacturing. With the aim to construct robust high-dimensional control chart that can handle two-sample high-dimensional data efficiently, previously we proposed three Shewhart type control charts – DR, BS, and SD charts. Due to the limitation of these memoryless charts for small shifts, we also proposed three more memory charts named DREWMA, BSEWMA, and SDEWMA. These memory charts were proved better than the memoryless charts in general for different combinations of distribution structures, dimension sizes, and sample sizes. In this study, with the focus to construct further better memory-based control chart, a new control chart under the CUSUM environment is proposed and named DRCUSUM chart. At first, using 100,000 iterations the values of mean, standard deviation, and the limit length h are obtained for the proposed chart under different distribution structures such as MVN_0 , MVN_1 , MVT, and MVG. The size of dimension for the high-dimensional data is set to $p = 100$ with the sample size for two-samples as $m = p \times 30\%$, $n = p \times 30\%$. The reference value for the DRCUSUM is set

to $K = 0.5$. After having the values of mean, standard deviation, and limit length h , using $ARL_0 = 200$ and simulation of size 5,000, the run length profile of the proposed DRCUSUM chart is studied. The chart performs the best under the distribution structure of MVN_0 . The proposed DRCUSUM chart is later compared with the previously proposed memory-based DREWMA and memoryless DR charts for different shifts ($\delta = 0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.75$, and 1) with $p = 100$ ($m = n = p \times 30\%$) under four different distribution structures. It is proved that though the memory chart under the EWMA setup (DREWMA chart) performs better than the chart under the Shewhart setup (DR chart), the performance of the memory chart under the CUSUM environment (DRCUSUM chart) is the best in detecting small shifts in the second sample of the two-sample high-dimensional data.

6.4 Future Research Directions

After a comprehensive review of the studies on high-dimensional monitoring methods and analyzing the efficiency of the proposed control charts, several future research directions were observed, with some significant ones being listed below:

- **Monitoring the covariance matrix of high-dimensional data:** In the literature, very few studies are designed to monitor the covariance matrix of high-dimensional data [7, 8, 35, 50]. However, the monitoring of the mean vector of high-dimensional data will be seriously affected if the covariance matrix of the high-dimensional data is unstable. Hence, more effective high-dimensional data-based charts are proposed for monitoring covariance matrices. In addition, one may offer combined structures to monitor the mean and covariance matrices simultaneously.

- **Joint monitoring of mean vector and covariance matrix of high-dimensional data:**
As discussed above, monitoring the mean vector of high-dimensional data requires stability in the covariance matrix. Instead of proposing combined structures, joint monitoring methods based on a single plotting statistic of high-dimensional data's mean vector and covariance matrix will be another possible opportunity to look at simultaneous shifts in means and variances in high-dimensional data.
- **Phase I analysis of the high-dimensional control charts:** In practice, most processes do not have known parameters (Case-U); hence, an effective Phase I analysis is typically required to determine the parameters of the monitoring mechanism. The analysis of high-dimensional control chart studies in this thesis show that most were designed under the assumption of known parameters (Case-U). Hence, exploring the high-dimensional control chart under Case-U will be an excellent contribution to the current literature.
- **Effect of estimation errors on high-dimensional control charts:** In the current era of Industry 4.0, efficient sensors are used to record data; however, sometimes natural estimation errors are recorded due to the sensor's lifetime, settlement, connectivity, synchronization, and human mishandling. Hence, this issue has not been studied and is a possible research topic to be highlighted and addressed using a the high-dimensional control chart based on estimation error models.
- **Effect of violation of independence and normality assumptions:** A good number of works has been carried out to monitor the shifts in high-dimensional data when the assumptions of independence and normality assumptions are not violated. However, only a few research is accomplished when the independence assumption is not satisfied.

For example, Moala, et al. [65] used STARMA process to deal with such issue. Similarly, when normality assumption is violated, only some studies tried to develop proper control charts for high-dimensional data. Hence an interesting research direction would be analyzing the changes in the detection ability of high-dimensional control charts when the abovementioned assumptions are not satisfied.

- **Problems with sensitivity, robustness, and changepoint detection:** The design of high-dimensional control charts should be moderately sensitive. Excessive sensitivity can detect shifts in the process caused by natural or minor variation that will produce false alarms, which is not expected. Whereas lack of sensitivity may lead to late detection of shifts in the process that will result in ineffective control chart. Hence, high-dimensional control charts with moderate sensitivity should be considered in future studies. In addition, changepoint or outlier detection for high-dimensional data is not simple like the low-dimensional cases. Therefore, constructing robust high-dimensional control charts that will effectively detect shifts with moderate sensitivity to extreme values can be a possible direction for future researchers in this area of research.
- **Mean or variance problem of two or more populations:** The comprehensive review found that most of the high-dimensional control charts are designed for detecting a change in the mean vector or covariance matrix of the high-dimensional data of a single population/group. However, sometimes researchers are interested in real-time detection of change in the mean vectors of two or more groups or variance matrices of two or more groups. For example, biologists are mostly interested in simultaneously examining mean changes in the genes of male and female patient groups [122-124]. To

support such cases, extending the high-dimensional control charts for real-time detection of change in the mean vectors of two or more groups or variance matrices of two or more groups will be an exceptional addition to the literature.

- **Count-based high-dimensional control charts:** Most high-dimensional control charts are variable-type charts developed for continuous-type high-dimensional data, such as the multivariate normal distribution, multivariate Gamma distribution, multivariate Cauchy distribution, and multivariate t-distribution. However, many processes generate count-based or discrete-type high-dimensional data that follow a multivariate Poisson distribution, multivariate binomial distribution, multivariate negative binomial distribution, and multivariate Conway-Maxwell Poisson distribution. An example is the number of delamination in the (PCB) packaging process [18]. Hence, extending high-dimensional control charts to count-based or discrete-type high-dimensional data is a possible future research direction.
- **Regression model based high-dimensional control charts:** Generally, some covariates are recorded along with the main high-dimensional study variables. Therefore, linear or nonlinear profile methods are usually preferred to monitor data by maintaining the relationships among variables. In the current review, a few studies [21, 49, 71] have focused on linear and nonlinear profile monitoring under a high-dimensional setup. Hence, extending the existing literature to linear and nonlinear high-dimensional profile monitoring can be another direction to enhance the literature.
- **Machine learning based high-dimensional control charts:** With the fourth revolution in the 21st century, the use of machine learning techniques from artificial intelligence has become a demanding area of research for monitoring high-dimensional

data. This study has noticed a few of the research [13, 53, 61, 63, 67, 79] applied machine learning tools to monitor shifts in the process. More studies are needed to make a solid bridge between machine learning techniques and already proposed methods to address high-dimensional monitoring.

- **Computational libraries or packages for high-dimensional control charts:** A total of 82 high-dimensional control chart studies are discussed in the current review. However, the proposed methods are not available in any computational software. Adding these methods in the form of packages in libraries in R-language and Python may decrease the complexity and allow practitioners to implement them easily and frequently.

References

- [1] C. Giraud, *Introduction to high-dimensional statistics*. CRC Press, 2021.
- [2] S. Kim, M. K. Jeong, and E. A. Elsayed, "A penalized likelihood-based quality monitoring via L2-norm regularization for high-dimensional processes," *Journal of Quality Technology*, vol. 52, no. 3, pp. 265-280, 2020/07/02 2020.
- [3] J. Li, "A two-stage online monitoring procedure for high-dimensional data streams," *Journal of Quality Technology*, vol. 51, no. 4, pp. 392-406, 2019/10/02 2019.
- [4] W. G. Martinez, M. L. Weese, and L. A. Jones-Farmer, "A one-class peeling method for multivariate outlier detection with applications in phase I SPC," vol. 36, no. 4, pp. 1272-1295, 2020.
- [5] Z. Wang, Y. Li, and X. Zhou, "A Statistical Control Chart for Monitoring High-dimensional Poisson Data Streams," vol. 33, no. 2, pp. 307-321, 2017.
- [6] M. R. Ahmad and S. E. Ahmed, "On the distribution of the T2 statistic, used in statistical process monitoring, for high-dimensional data," *Statistics & Probability Letters*, vol. 168, p. 108919, 2021/01/01/ 2021.
- [7] J. Kim, G. M. Abdella, S. Kim, K. N. Al-Khalifa, and A. M. Hamouda, "Control charts for variability monitoring in high-dimensional processes," *Computers & Industrial Engineering*, vol. 130, pp. 309-316, 2019/04/01/ 2019.
- [8] G. M. Abdella, M. R. Maleki, S. Kim, K. N. Al-Khalifa, and A. M. S. Hamouda, "Phase-I monitoring of high-dimensional covariance matrix using an adaptive thresholding LASSO rule," *Computers & Industrial Engineering*, vol. 144, p. 106465, 2020/06/01/ 2020.
- [9] Y. Li, D. Pei, and Z. Wu, "A multivariate non-parametric control chart based on run test," *Computers & Industrial Engineering*, vol. 149, p. 106839, 2020/11/01/ 2020.
- [10] D. Sun, G. Lu, H. Zhou, and Y. Yan, "Condition Monitoring of Combustion Processes Through Flame Imaging and Kernel Principal Component Analysis," *Combustion Science and Technology*, vol. 185, no. 9, pp. 1400-1413, 2013/09/02 2013.
- [11] C. Kan and H. Yang, "Network models for monitoring high-dimensional image profiles," in *2015 IEEE International Conference on Automation Science and Engineering (CASE)*, 2015, pp. 1078-1083.
- [12] I. Ullah, M. D. M. Pawley, A. N. H. Smith, and B. Jones, "Improving the detection of unusual observations in high-dimensional settings," vol. 59, no. 4, pp. 449-462, 2017.
- [13] J. Yu and C. Zhang, "Manifold regularized stacked autoencoders-based feature learning for fault detection in industrial processes," *Journal of Process Control*, vol. 92, pp. 119-136, 2020/08/01/ 2020.
- [14] A. Ahmadi-Javid and M. Ebadi, "A two-step method for monitoring normally distributed multi-stream processes in high dimensions," *Quality Engineering*, vol. 33, no. 1, pp. 143-155, 2021/01/02 2021.

- [15] K. Hu and J. Yuan, "Multivariate statistical process control based on multiway locality preserving projections," *Journal of Process Control*, vol. 18, no. 7, pp. 797-807, 2008/08/01/ 2008.
- [16] K. Hu and J. Yuan, "Batch process monitoring with tensor factorization," *Journal of Process Control*, vol. 19, no. 2, pp. 288-296, 2009/02/01/ 2009.
- [17] R. C. Wang, M. Baldea, and T. F. Edgar, "Data Visualization and Visualization-Based Fault Detection for Chemical Processes," *Processes* vol. 5, no. 3, p. 45, 2017.
- [18] Z. Wang, Y. Li, and X. Zhou, "A Statistical Control Chart for Monitoring High-dimensional Poisson Data Streams," *Quality and Reliability Engineering International*, vol. 33, no. 2, pp. 307-321, 2017.
- [19] G. M. Abdella, K. N. Al-Khalifa, S. Kim, M. K. Jeong, E. A. Elsayed, and A. M. Hamouda, "Variable Selection-based Multivariate Cumulative Sum Control Chart," *Quality and Reliability Engineering International*, vol. 33, no. 3, pp. 565-578, 2017.
- [20] F. A. P. Peres, T. N. Peres, F. S. Fogliatto, and M. J. Anzanello, "Fault detection in batch processes through variable selection integrated to multiway principal component analysis," *Journal of Process Control*, vol. 80, pp. 223-234, 2019/08/01/ 2019.
- [21] C. Zou, S.-T. Tseng, and Z. Wang, "Outlier detection in general profiles using penalized regression method," *IIE Transactions*, vol. 46, no. 2, pp. 106-117, 2014/02/01 2014.
- [22] C. Zou, Z. Wang, X. Zi, and W. Jiang, "An Efficient Online Monitoring Method for High-Dimensional Data Streams," *Technometrics*, vol. 57, no. 3, pp. 374-387, 2015/07/03 2015.
- [23] L. Shu and J. Fan, "A distribution-free control chart for monitoring high-dimensional processes based on interpoint distances," *Naval Research Logistics*, vol. 65, no. 4, pp. 317-330, 2018.
- [24] G. M. Abdella *et al.*, "An adaptive thresholding-based process variability monitoring," *Journal of Quality Technology*, vol. 51, no. 3, pp. 242-256, 2019/07/03 2019.
- [25] W. Li, D. Xiang, F. Tsung, and X. Pu, "A Diagnostic Procedure for High-Dimensional Data Streams via Missed Discovery Rate Control," *Technometrics*, vol. 62, no. 1, pp. 84-100, 2020/01/02 2020.
- [26] C. Zhang, N. Chen, and J. Wu, "Spatial rank-based high-dimensional monitoring through random projection," *Journal of Quality Technology*, vol. 52, no. 2, pp. 111-127, 2020/04/02 2020.
- [27] A. Mukherjee and M. Marozzi, "Nonparametric Phase-II control charts for monitoring high-dimensional processes with unknown parameters," *Journal of Quality Technology*, vol. 54, no. 1, pp. 44-64, 2022/01/01 2022.
- [28] D. Xiang, P. Qiu, D. Wang, and W. Li, "Reliable Post-Signal Fault Diagnosis for Correlated High-Dimensional Data Streams," *Technometrics*, vol. 64, no. 3, pp. 323-334, 2022/07/03 2022.
- [29] H. Ye, Z. Zheng, J.-R. C. Cheng, B. Hable, and K. Liu, "Online monitoring of high-dimensional asynchronous and heterogeneous data streams for shifts in

- location and scale," *International Journal of Production Research*, pp. 1-17, 2023.
- [30] J. Yu, "Local and global principal component analysis for process monitoring," *Journal of Process Control*, vol. 22, no. 7, pp. 1358-1373, 2012/08/01/ 2012.
 - [31] R. T. Samuel and Y. Cao, "Fault detection in a multivariate process based on kernel PCA and kernel density estimation," in *2014 20th International Conference on Automation and Computing*, 2014, pp. 146-151.
 - [32] S. Kim and M. Turkoz, "Bayesian sequential update for monitoring and control of high-dimensional processes," *Annals of Operations Research*, vol. 317, no. 2, pp. 693-715, 2022/10/01 2022.
 - [33] A. Alaeddini, A. Motasemi, and S. H. A. Faruqi, "A spatiotemporal outlier detection method based on partial least squares discriminant analysis and area Delaunay triangulation for image-based process monitoring," *IIE Transactions*, vol. 50, no. 2, pp. 74-87, 2018/02/01 2018.
 - [34] C. Zou, W. Jiang, and F. Tsung, "A LASSO-Based Diagnostic Framework for Multivariate Statistical Process Control," *Technometrics*, vol. 53, no. 3, pp. 297-309, 2011/08/01 2011.
 - [35] Z. Li and F. Tsung, "A control scheme for monitoring process covariance matrices with more variables than observations," vol. 35, no. 1, pp. 351-367, 2019.
 - [36] J. Merlo, A. E. Cordero-Franco, and V. G. Tercero-Gómez, "Nonparametric multivariate processes monitoring with guaranteed in-control performance for changes in location," *Computers & Industrial Engineering*, vol. 166, p. 107940, 2022/04/01/ 2022.
 - [37] J. Kim, M. K. Jeong, E. A. Elsayed, K. N. Al-Khalifa, and A. M. S. Hamouda, "An adaptive step-down procedure for fault variable identification," *International Journal of Production Research*, vol. 54, no. 11, pp. 3187-3200, 2016/06/02 2016.
 - [38] M. Turkoz *et al.*, "Bayesian framework for fault variable identification," *Journal of Quality Technology*, vol. 51, no. 4, pp. 375-391, 2019/10/02 2019.
 - [39] Y. Wang, Y. Mei, and K. Paynabar, "Thresholded Multivariate Principal Component Analysis for Phase I Multichannel Profile Monitoring," *Technometrics*, vol. 60, no. 3, pp. 360-372, 2018/07/03 2018.
 - [40] W. Zhao, Z. Wang, and C. Wu, "Adaptive multivariate EWMA charts for monitoring sparse mean shifts based on parameter optimization design," *Journal of Statistical Computation and Simulation*, vol. 91, no. 13, pp. 2670-2683, 2021/09/02 2021.
 - [41] J.-B. Yu, "Bearing performance degradation assessment using locality preserving projections," *Expert Systems with Applications*, vol. 38, no. 6, pp. 7440-7450, 2011/06/01/ 2011.
 - [42] D. Qi, Z. Li, and Z. Wang, "On-line monitoring data quality of high-dimensional data streams," *Journal of Statistical Computation and Simulation*, vol. 86, no. 11, pp. 2204-2216, 2016/07/23 2016.
 - [43] S. Huang, Z. Kong, and W. Huang, "High-dimensional process monitoring and change point detection using embedding distributions in reproducing kernel Hilbert space," *IIE Transactions*, vol. 46, no. 10, pp. 999-1016, 2014/10/03 2014.

- [44] H. Yan, K. Paynabar, and J. Shi, "Real-Time Monitoring of High-Dimensional Functional Data Streams via Spatio-Temporal Smooth Sparse Decomposition," *Technometrics*, vol. 60, no. 2, pp. 181-197, 2018/04/03 2018.
- [45] A. M. E. Gómez, D. Li, and K. Paynabar, "An Adaptive Sampling Strategy for Online Monitoring and Diagnosis of High-Dimensional Streaming Data," *Technometrics*, vol. 64, no. 2, pp. 253-269, 2022/04/03 2022.
- [46] S. Kim, M. Turkoz, and J. W. Baek, "High-Speed Monitoring of Multidimensional Processes Using Bayesian Updates," *IEEE Access*, vol. 10, pp. 97450-97464, 2022.
- [47] L. Du and C. Zou, "On-line control of false discovery rates for multiple datastreams," *Journal of Statistical Planning and Inference*, vol. 194, pp. 1-14, 2018/03/01/ 2018.
- [48] C. Zou, X. Ning, and F. Tsung, "LASSO-based multivariate linear profile monitoring," *Annals of Operations Research*, vol. 192, no. 1, pp. 3-19, 2012/01/01 2012.
- [49] J.-M. Kim, N. Wang, Y. Liu, and K. Park, "Residual Control Chart for Binary Response with Multicollinearity Covariates by Neural Network Model," *Symmetry*, vol. 12, no. 3, p. 381, 2020.
- [50] J. Fan, L. Shu, A. Yang, and Y. Li, "Phase I analysis of high-dimensional covariance matrices based on sparse leading eigenvalues," *Journal of Quality Technology*, vol. 53, no. 4, pp. 333-346, 2021/08/08 2021.
- [51] E. Maboudou-Tchao, C. W. Harrison, and S. Sen, "A comparison study of penalized likelihood via regularization and support vector-based control charts," *Quality Technology & Quantitative Management*, vol. 20, no. 2, pp. 147-167, 2023/03/04 2023.
- [52] C. Kan and H. Yang, "Dynamic network monitoring and control of in situ image profiles from ultraprecision machining and biomanufacturing processes," *Quality and Reliability Engineering International*, vol. 33, no. 8, pp. 2003-2022, 2017.
- [53] E. M. Maboudou-Tchao, "High-dimensional data monitoring using support machines," *Communications in Statistics - Simulation and Computation*, vol. 50, no. 7, pp. 1927-1942, 2021/07/03 2021.
- [54] I. Ullah, M. D. M. Pawley, A. N. H. Smith, and B. Jones, "Improving the detection of unusual observations in high-dimensional settings," *Australian & New Zealand Journal of Statistics*, vol. 59, no. 4, pp. 449-462, 2017.
- [55] D. Yan, S. Zhang, and U. Jung, "A variable-selection control chart via penalized likelihood and Gaussian mixture model for multimodal and high-dimensional processes," *Quality and Reliability Engineering International*, vol. 35, no. 4, pp. 1263-1275, 2019.
- [56] K. Sangahn, "Variable selection-based SPC procedures for high-dimensional multistage processes," *Journal of Systems Engineering and Electronics*, vol. 30, no. 1, pp. 144-153, 2019.
- [57] J. Yu, "A nonlinear kernel Gaussian mixture model based inferential monitoring approach for fault detection and diagnosis of chemical processes," *Chemical Engineering Science*, vol. 68, no. 1, pp. 506-519, 2012/01/22/ 2012.

- [58] G. J. Odom, K. B. Newhart, T. Y. Cath, and A. S. Hering, "Multistate multivariate statistical process control," *Applied stochastic models in business and industry*, vol. 34, no. 6, pp. 880-892, 2018.
- [59] J. Chen, C. Park, S.-H. Kim, and Y. Xie, "To reduce or not to reduce: a study on spatio-temporal surveillance," *Environmental and Ecological Statistics*, vol. 26, no. 3, pp. 217-238, 2019/09/01 2019.
- [60] Z. Fang, W. Li, X. Liu, X. Pu, and D. Xiang, "Online monitoring of high-dimensional binary data streams with application to extreme weather surveillance," *Journal of Applied Statistics*, vol. 49, no. 16, pp. 4122-4136, 2022/12/10 2022.
- [61] S. Lee, M. Kwak, K.-L. Tsui, and S. B. Kim, "Process monitoring using variational autoencoder for high-dimensional nonlinear processes," *Engineering Applications of Artificial Intelligence*, vol. 83, pp. 13-27, 2019/08/01/ 2019.
- [62] S. Jang, S. H. Park, and J.-G. Baek, "Real-time contrasts control chart using random forests with weighted voting," *Expert Systems with Applications*, vol. 71, pp. 358-369, 2017/04/01/ 2017.
- [63] S. He, W. Jiang, and H. Deng, "A distance-based control chart for monitoring multivariate processes using support vector machines," *Annals of Operations Research*, vol. 263, no. 1, pp. 191-207, 2018/04/01 2018.
- [64] J. Guo, H. Yan, and C. Zhang, "A Bayesian Partially Observable Online Change Detection Approach with Thompson Sampling," *Technometrics*, pp. 1-13, 2022.
- [65] A. B. Moala, L. L. Ho, and R. C. Quinino, "Multivariate control charts to monitor the monthly frequency of vehicle robberies in São Paulo city," *Spatial Statistics*, vol. 29, pp. 49-65, 2019/03/01/ 2019.
- [66] L. Feng, H. Ren, and C. Zou, "A setwise EWMA scheme for monitoring high-dimensional datastreams," *Random Matrices: Theory and Applications*, vol. 09, no. 02, p. 2050004, 2020.
- [67] W. G. Martinez, M. L. Weese, and L. A. Jones-Farmer, "A one-class peeling method for multivariate outlier detection with applications in phase I SPC," *Quality and Reliability Engineering International*, vol. 36, no. 4, pp. 1272-1295, 2020.
- [68] M. Shaohui, G. Tuerhong, M. Wushouer, and T. Yibulayin, "PCA mix-based Hotelling's T^2 multivariate control charts for intrusion detection system," *IET Information Security*, vol. 16, no. 3, pp. 161-177, 2022.
- [69] K. Wang and W. Jiang, "High-Dimensional Process Monitoring and Fault Isolation via Variable Selection," *Journal of Quality Technology*, vol. 41, no. 3, pp. 247-258, 2009/07/01 2009.
- [70] H. Zhang and S. Albin, "Detecting outliers in complex profiles using a χ^2 control chart method," *IIE Transactions*, vol. 41, no. 4, pp. 335-345, 2009/02/23 2009.
- [71] S. Chen and H. B. Nembhard, "A high-dimensional control chart for profile monitoring," *Quality and Reliability Engineering International*, vol. 27, no. 4, pp. 451-464, 2011.
- [72] N. D. Tracy, J. C. Young, and R. L. Mason, "Multivariate control charts for individual observations," *Journal of Quality Technology*, vol. 24, no. 2, pp. 88-95, 1992.

- [73] S. Bersimis, S. Psarakis, and J. Panaretos, "Multivariate statistical process control charts: an overview," *Quality and Reliability engineering international*, vol. 23, no. 5, pp. 517-543, 2007.
- [74] C. A. Lowry, W. H. Woodall, C. W. Champ, and S. E. Rigdon, "A multivariate exponentially weighted moving average control chart," *Technometrics*, vol. 34, no. 1, pp. 46-53, 1992.
- [75] R. B. J. T. Crosier, "Multivariate generalizations of cumulative sum quality-control schemes," *Technometrics*, vol. 30, no. 3, pp. 291-303, 1988.
- [76] J.-M. Lee, C. Yoo, S. W. Choi, P. A. Vanrolleghem, and I.-B. Lee, "Nonlinear process monitoring using kernel principal component analysis," *Chemical Engineering Science*, vol. 59, no. 1, pp. 223-234, 2004/01/01/ 2004.
- [77] E. Skubalska-Rafajłowicz, "Random projections and Hotelling's T2 statistics for change detection in high-dimensional data streams," *International Journal of Applied Mathematics and Computer Science*, vol. Vol. 23, no. 2, pp. 447-461, 2013.
- [78] K. Fujiwara and M. Kano, "Development of correlation-based process characteristics visualization method and its application to fault detection," in *2017 13th IEEE International Conference on Control & Automation (ICCA)*, 2017, pp. 940-945.
- [79] J. Yu and X. Liu, "One-dimensional residual convolutional auto-encoder for fault detection in complex industrial processes," *International Journal of Production Research*, vol. 60, no. 18, pp. 5655-5674, 2022/09/17 2022.
- [80] X. Ning and F. Tsung, "A density-based statistical process control scheme for high-dimensional and mixed-type observations," *IIE Transactions*, vol. 44, no. 4, pp. 301-311, 2012/04/01 2012.
- [81] S. J. Bae, G. Do, and P. Kvam, "On data depth and the application of nonparametric multivariate statistical process control charts," *Applied Stochastic Models in Business and Industry*, vol. 32, no. 5, pp. 660-676, 2016.
- [82] C. Zhang, H. Yan, S. Lee, and J. Shi, "Multiple profiles sensor-based monitoring and anomaly detection," *Journal of Quality Technology*, vol. 50, no. 4, pp. 344-362, 2018/10/02 2018.
- [83] C. Zhang, H. Yan, S. Lee, and J. Shi, "Weakly correlated profile monitoring based on sparse multi-channel functional principal component analysis," *IIE Transactions*, vol. 50, no. 10, pp. 878-891, 2018/10/03 2018.
- [84] K. Liu, Y. Mei, and J. Shi, "An Adaptive Sampling Strategy for Online High-Dimensional Process Monitoring," *Technometrics*, vol. 57, no. 3, pp. 305-319, 2015/07/03 2015.
- [85] S. Ahmad, N. Enshaei, F. Naderkhani, and A. Awasthi, "Integrated Deep Learning and Statistical Process Control for Online Monitoring of Manufacturing Processes," in *2020 IEEE International Conference on Prognostics and Health Management (ICPHM)*, 2020, pp. 1-6.
- [86] J. Li, "Efficient global monitoring statistics for high-dimensional data," *Quality and Reliability Engineering International*, vol. 36, no. 1, pp. 18-32, 2020.
- [87] Y. Li, Y. Liu, C. Zou, and W. Jiang, "A self-starting control chart for high-dimensional short-run processes," *International Journal of Production Research*, vol. 52, no. 2, pp. 445-461, 2014/01/17 2014.

- [88] M. Turkoz, S. Kim, Y.-S. Jeong, K. N. Al-Khalifa, and A. M. Hamouda, "Distribution-Free Adaptive Step-Down Procedure for Fault Identification," *Quality and Reliability Engineering International*, vol. 32, no. 8, pp. 2701-2716, 2016.
- [89] X. Guo, Z. He, and H. Chen, "A Real-Time Contrasts Method for Monitoring Image Data," in *2019 IEEE 6th International Conference on Industrial Engineering and Applications (ICIEA)*, 2019, pp. 354-361.
- [90] Y. Mei, "Efficient scalable schemes for monitoring a large number of data streams," *Biometrika*, vol. 97, no. 2, pp. 419-433, 2010.
- [91] P. Qiu and D. Hawkins, "A nonparametric multivariate cumulative sum procedure for detecting shifts in all directions," *Journal of the Royal Statistical Society: Series D (The Statistician)*, vol. 52, no. 2, pp. 151-164, 2003.
- [92] A. Wang, X. Xian, F. Tsung, and K. Liu, "A spatial-adaptive sampling procedure for online monitoring of big data streams," *Journal of Quality Technology*, vol. 50, no. 4, pp. 329-343, 2018.
- [93] R. L. Mason, N. D. Tracy, and J. C. Young, "Decomposition of T^2 for multivariate control chart interpretation," *Journal of Quality Technology*, vol. 27, no. 2, pp. 99-108, 1995.
- [94] J. H. Sullivan, Z. G. Stoumbos, R. L. Mason, and J. C. Young, "Step-down analysis for changes in the covariance matrix and other parameters," *Journal of Quality Technology*, vol. 39, no. 1, pp. 66-84, 2007.
- [95] G. M. Abdella, K. N. Al-Khalifa, S. Kim, M. K. Jeong, E. A. Elsayed, and A. M. Hamouda, "Variable Selection-based Multivariate Cumulative Sum Control Chart," vol. 33, no. 3, pp. 565-578, 2017.
- [96] N. S. Meneces, S. A. Olivera, C. D. Saccone, and J. Tessore, "Statistical Control of Multiple-Stream Processes: A Shewhart Control Chart for Each Stream," *Quality Engineering*, vol. 20, no. 2, pp. 185-194, 2008/03/18 2008.
- [97] D. M. Hawkins and E. M. Maboudou-Tchao, "Self-Starting Multivariate Exponentially Weighted Moving Average Control Charting," *Technometrics*, vol. 49, no. 2, pp. 199-209, 2007/05/01 2007.
- [98] C. Zou, Z. Wang, and F. Tsung, "A spatial rank-based multivariate EWMA control chart," *Naval Research Logistics*, vol. 59, no. 2, pp. 91-110, 2012.
- [99] N. Chen, X. Zi, and C. Zou, "A Distribution-Free Multivariate Control Chart," *Technometrics*, vol. 58, no. 4, pp. 448-459, 2016/10/01 2016.
- [100] W. Jiang, K. Wang, and F. Tsung, "A Variable-Selection-Based Multivariate EWMA Chart for Process Monitoring and Diagnosis," *Journal of Quality Technology*, vol. 44, no. 3, pp. 209-230, 2012/07/01 2012.
- [101] Y. Li and F. Tsung, "False Discovery Rate-Adjusted Charting Schemes for Multistage Process Monitoring and Fault Identification," *Technometrics*, vol. 51, no. 2, pp. 186-205, 2009/05/01 2009.
- [102] Y. Li and F. Tsung, "Multiple Attribute Control Charts with False Discovery Rate Control," *Quality and Reliability Engineering International*, vol. 28, no. 8, pp. 857-871, 2012.
- [103] K. Zamba and D. M. Hawkins, "A multivariate change-point model for statistical process control," *Technometrics*, vol. 48, no. 4, pp. 539-549, 2006.

- [104] B. Li, K. Wang, and A. B. Yeh, "Monitoring the covariance matrix via penalized likelihood estimation," *IIE Transactions*, vol. 45, no. 2, pp. 132-146, 2013.
- [105] M. S. Srivastava and M. Du, "A test for the mean vector with fewer observations than the dimension," *Journal of Multivariate Analysis*, vol. 99, no. 3, pp. 386-402, 2008.
- [106] A. P. Dempster, "A high dimensional two sample significance test," *The Annals of Mathematical Statistics*, pp. 995-1010, 1958.
- [107] A. P. Dempster, "A significance test for the separation of two highly multivariate small samples," *Biometrics*, vol. 16, no. 1, pp. 41-50, 1960.
- [108] Z. Bai and H. Saranadasa, "Effect of high dimension: by an example of a two sample problem," *Statistica Sinica*, pp. 311-329, 1996.
- [109] K. Lee, L. Lin, and K. You, "Statistical Hypothesis Testing Toolbox," R package version 0.1.8 ed: CRAN, 2022.
- [110] Z. G. Stoumbos and J. H. J. J. o. Q. T. Sullivan, "Robustness to non-normality of the multivariate EWMA control chart," vol. 34, no. 3, pp. 260-276, 2002.
- [111] R. S. Amano, "Review of wind turbine research in 21st century," *Journal of Energy Resources Technology*, vol. 139, no. 5, p. 050801, 2017.
- [112] Z. Liu and L. Zhang, "A review of failure modes, condition monitoring and fault diagnosis methods for large-scale wind turbine bearings," *Measurement*, vol. 149, p. 107002, 2020/01/01/ 2020.
- [113] T. Fierro, "Extending Bearing Life in Wind Turbine Mainshafts," *Power Engineering*, vol. 120, 2016.
- [114] F. Oyague, "Gearbox modeling and load simulation of a baseline 750-kW wind turbine using state-of-the-art simulation codes," National Renewable Energy Lab.(NREL), Golden, CO (United States)2009.
- [115] Y. Yucesan, "Replication Data for: A Hybrid Modeling for Wind Turbine Main Bearing Fatigue with Uncertainty in Grease Observations," V1 ed: Harvard Dataverse, 2020.
- [116] R. Killick, C. Beaulieu, S. Taylor, and H. Hullait, "EnvCpt: Detection of Structural Changes in Climate and Environment Time Series," R package version 1.1.3 ed: CRAN, 2021.
- [117] M. S. Srivastava and M. J. J. o. M. A. Du, "A test for the mean vector with fewer observations than the dimension," vol. 99, no. 3, pp. 386-402, 2008.
- [118] S. Yang *et al.*, "A 2DCNN-RF Model for Offshore Wind Turbine High-Speed Bearing-Fault Diagnosis under Noisy Environment," vol. 15, no. 9, p. 3340, 2022.
- [119] E. S. Page, "CONTINUOUS INSPECTION SCHEMES," *Biometrika*, vol. 41, no. 1-2, pp. 100-115, 1954.
- [120] Z. G. Stoumbos and J. H. Sullivan, "Robustness to non-normality of the multivariate EWMA control chart," *Journal of Quality Technology*, vol. 34, no. 3, pp. 260-276, 2002.
- [121] D. C. Montgomery, *Introduction to statistical quality control*, Sixth ed. John Wiley & Sons, 2009.
- [122] S. X. Chen and Y.-L. Qin, "A two-sample test for high-dimensional data with applications to gene-set testing," *The Annals of Statistics*, vol. 38, no. 2, pp. 808-835, 28, 2010.

- [123] M. Thulin, "A high-dimensional two-sample test for the mean using random subspaces," *Computational Statistics & Data Analysis*, vol. 74, pp. 26-38, 2014/06/01/ 2014.
- [124] H. Zhang and H. Wang, "A more powerful test of equality of high-dimensional two-sample means," *Computational Statistics & Data Analysis*, vol. 164, p. 107318, 2021/12/01/ 2021.

Vitae

Name :Fuhad Ahmed

Nationality :Bangladeshi

Date of Birth :3/7/1986

Email :fuhad.ahmed055@gmail.com

Address :Panjipuri, Beanibazar, P.O. Box 3175, Sylhet, Bangladesh

Academic Background :Master of Science in Applied Statistics, KFUPM, KSA

- Master of Science in Statistics, SUST, Bangladesh
- Bachelor of Science in Statistics, SUST, Bangladesh

Professional Experience :

- Senior Lecturer of Statistics (Study Leave), Port City International University, Chattogram, Bangladesh (November 2018 to Present)
- Lecturer of Statistics, Port City International University, Chattogram, Bangladesh (September 2015 to October 2018)
- Assistant Teacher of Statistics, The Sylhet Khajanchibari International School & College (SKISC), Bangladesh (January 2014 to September 2015)

Research Papers:

- ❖ F. Ahmed, T. Mahmood, M. Riaz, and N. Abbas, “The Current and Past Trends of High-Dimensional Monitoring Methods”, 2023. (Submitted)
- ❖ T. Mahmood, F. Ahmed, M. Riaz, and N. Abbas, “High-Dimensional Control Charts with Application to Surveillance of Grease Damage in Bearings of Wind Turbines”, *Production & Manufacturing Research*, 2023. (To be re-submitted)
- ❖ F. Ahmed, T. Mahmood, M. Riaz, and N. Abbas, “Real-time Surveillance of Two-Sample High-Dimensional Processes”, 2023. (Submitted)
- ❖ F. Ahmed, T. Mahmood, M. Riaz, and N. Abbas, “High-Dimensional Control Charts Under CUSUM Setup”, 2023. (To be submitted)
- ❖ J. Keya, F. Ahmed, and S. A. Karim, “NEURAL NETWORK APPROACH IN FINDING THE FACTORS BEHIND MATH PHOBIA”, *Port City International University Journal (ISSN: 2311-3146)*, vol. 8, no. (1&2), pp. 45-56, 2021.
- ❖ F. Ahmed and J. Keya, “Time Series Analysis for Predicting the Exchange Rate of USD to BDT”, *International Journal of Academic Research*, vol. 1, no. 2, pp. 282-294, 2019.
- ❖ F. Ahmed, “Application of ARIMA Models in Forecasting Monthly Total Rainfall of Rangamati, Bangladesh”, *Asian Journal of Applied Science and Engineering*, vol. 6, no. 2, pp. 95-102, 2017.