Received: 21 August 2023 | Revised: 25 April 2024 | Accepted: 1 July 2024

DOI: 10.1002/qre.3619

RESEARCH ARTICLE

WILEY

Distribution-free multivariate process monitoring: A rank-energy statistic-based approach

Niladri Chakraborty[1] | **Maxim Finkelstein**[1,2]

[1]Department of Mathematical Statistics and Actuarial Science, University of the Free State, Bloemfontein, South Africa

[2]Department of Management Science, University of Strathclyde, Glasgow, UK

Correspondence
Niladri Chakraborty
Email: niladri.chakraborty30@gmail.com

Abstract

In this paper, a multivariate process monitoring scheme based on the rank-energy statistics is proposed which is suitable for high-dimensional applications such as sensorless drive diagnosis. The rank-energy statistic is based on multivariate ranks that is grounded on the measure transportation theory. Univariate ranks could be interpreted as a solution to an optimisation problem involving a given set of observations of size n and the set $\{1,\ 2,\ 3, .., \ n\}$. Recently, attaining greater robustness than spatial sign or depth-based ranks, multivariate ranks are proposed as solutions to such optimisation problem in multivariate settings (measure transportation problem). The proposed multivariate process monitoring scheme based on the rank-energy statistic, subsequently, attains greater robustness than existing nonparametric multivariate process monitoring methods based on spatial sign or depth-based ranks. The proposed method is also applicable to high-dimensional data unlike some of the existing nonparametric multivariate process monitoring methods. A rigorous simulation study demonstrates its effective shift detection ability and other important features. A practical application of the proposed method is demonstrated with the sensorless drive diagnosis case study.

KEYWORDS

measure transportation, multivariate process monitoring, multivariate ranks, rank-energy statistic, sensorless drive diagnosis

1 | INTRODUCTION

Since Hotelling's pioneering work introducing the Hotelling T^2-statistic,[1] multivariate statistical process control (SPC) in literature has attracted remarkable attentions of statisticians and industry practitioners.[2] With 'arrival' of industry 4.0, these multivariate SPC methodologies have been applied across a variety of sectors. For instance, in semiconductor manufacturing,[3] network monitoring,[4] image surveillance,[5,6] to name a few. Numerous papers focusing on parametric multivariate SPC for symmetric and skewed processes could be found in the literature. Comprehensive reviews on multivariate SPC are available in the monographs by Ge and Song[7] and Qiu.[2]

Parametric multivariate SPC methods, despite their utility, have certain limitations. For instance, traditional T^2-type monitoring schemes require the normality assumption for a multivariate process, a condition which is often violated in a real-world scenario. In addition, T^2-type monitoring schemes assume that the sample size, n is larger than the dimension, d and their performance tends to deteriorate as d increases.[8] The performance of parametric monitoring methods, either for Gaussian or non-Gaussian assumptions, deteriorates when these distributional assumptions are violated in practice. The impact of parameter estimation also proves to be a crucial factor in designing multivariate parametric control charts.[9,10] In addition, the estimation and monitoring of covariance matrix are also critical in multivariate SPC.[11] Qiu[12] mentioned that '*A direct conclusion of the multivariate normality assumption is that the regression relationship between any two subsets of the individual variables must be linear, which is rarely valid in practice*'.

In contemporary applications enriched with high-dimensional data, the adoption of machine learning approaches in SPC has become popular in recent years.[13,14] Although some machine learning techniques applied in multivariate SPC do not depend on the assumption of normality, the effectiveness of such methods is influenced by optimal parameter selection (e.g., the number of trees, window size, number of layers) and the quality of the training data.[15–17] Some machine learning approaches in multivariate SPC require density estimation or estimation of the correlation structure of the multivariate data.[18,19] Moreover, these methods may suffer from the lack of power when the data is highly correlated.[3] In that sense, strictly speaking, the machine learning approach-based SPC methods are not *exactly* distribution-free. Therefore, the multivariate nonparametric SPC tools that do not require information about the distribution of a multivariate process are critically important.

The first effort, to the best of our knowledge, to develop multivariate nonparametric SPC method was undertaken by Liu,[20] who proposed multivariate control charts based on data depth. Systematic study on nonparametric multivariate SPC then began with the seminal works of Qiu and Hawkins,[21,22] and Qiu.[23] As natural extension to the concepts of univariate nonparametric SPC methods, multivariate nonparametric SPC methods are developed using the ranking information of the multivariate process observations. Based on the notion of multivariate ranking, multivariate nonparametric SPC methods could be broadly classified into two categories; (i) Longitudinal ranking-based methods; (ii) Cross-component ranking-based methods.[12] As mentioned in Qiu,[12] longitudinal ranking refers to the ranking of multivariate observations at different time points and cross-component ranking refers to the ranking across components. Qiu and Hawkins[21,22] proposed cumulative sum (CUSUM) charts based on the antiranks of the components of the multivariate observations. Qiu[23] proposed a multivariate nonparametric control chat based on log-linear modelling.

Univariate sign and rank statistics were extended by several authors to d-dimensional Euclidean space ($d \geq 2$), for example, in Marden[24] and Randles[25] with an extensive overview given by Hettmansperger and McKean.[26] Subsequently, spatial sign or spatial rank-based SPC methods were proposed. For instance, Holland and Hawkins[27] proposed a directional rank test-based SPC scheme. Whilst their method does not require a large amount of reference data, its robustness is compromised.[5,27] Bae et al.[28] reviewed multivariate process monitoring methods based on the data-depth. Mukherjee and Marozzi[29] noted that the depth-based methods are not robust. Several researchers have noted that SPC methods based on data depth tend to be less effective, particularly in scenarios where extensive reference data are unavailable.[12,30]

Following Randles[25] and Hettmansperger and Randles,[31] several multivariate nonparametric SPC methods were proposed based on multivariate sign and rank statistics.[3,32–39] Note that, Randles[25] indicated that the multivariate sign test is robust for elliptical-direction class of distributions. Chen et al.[40] have proposed Wilcoxon statistic-based approach to monitor the data streams individually. Zhang et al.[41] have extended Chen et al.[40] by splitting the high-dimensional data into 2-dimensional ones, risking loss of information as noted by Zhang et al.[3] Mukherjee and Marozzi[29] introduced an interpoint Euclidean distance-based multivariate SPC method. Sometimes the distance-based methods may suffer from 'practitioners bias' as we explain later in this article.

Due to the lack of canonical ordering in the d-dimensional Euclidean space $\mathcal{R}^d$, for dimension $d \geq 2$, fundamental statistical notions like ranks and empirical quantiles do not extend canonically to $d \geq 2$.[42] Consequently, component-wise rank, spatial rank, spatial sign and depth-based ranks and the corresponding statistics do not possess the *exact* distribution-freeness unlike their one-dimensional counterparts.[42,43] Recently, Chernozhukov et al.[44] and Hallin et al.[42] have suggested a significant advancement by proposing the multivariate ranks based on the measure transportation theory. For $\boldsymbol{X} \in R^d$, the measure transportation theory is about finding an optimal map $V : R^d \to R^d$ that would minimise the Euclidean distance between $\boldsymbol{X}$ and $V(\boldsymbol{X})$, see Hallin[45] and Hallin and Mordant[46] for further details. More comprehensive discussion on this topic is provided in the subsequent section. In a recent article, Deb and Sen[43] proposed a distribution-free, multivariate two-sample test statistic, called the rank-energy (RE^2) statistic, based on the measure transportation theory. Whilst Deb and Sen[43] have laid the groundworks with foundational theorems, the power properties of the rank-energy test and other related issues need further study.

TABLE 1 Abbreviations and corresponding full forms.

Abbreviation	Full form
RE^2	Rank-energy statistic
iid	Independent and identically distributed
IC	In-control
OC	Out-of-control
FAR	False alarm rate
ARL	Average run length
HDS	High-dimensional Shewhart-type monitoring scheme
HDSOR	High-dimensional Shewhart-type monitoring scheme based on distance from the origin
HDSIP	High-dimensional Shewhart-type monitoring scheme based on inter-point distance
HDSGM	High-dimensional Shewhart-type monitoring scheme based distance from the generalised median
VSD	Variable speed drive
HVAC	Heating, ventilation, air conditioning system.

In SPC literature, all SPC methods could be typically categorised into two types: Phase I and Phase II methods. Phase I methods focus on monitoring offline data, during which no new process observations are collected. The stable dataset established after Phase I monitoring is referred to as the '*reference*' data. Conversely, Phase II methods involve online process monitoring that collects new observations sequentially throughout the monitoring process. The data gathered during this online monitoring is known as the '*test*' data. Jones–Farmer et al.[47] provided a comprehensive overview on Phase I SPC for both univariate and multivariate processes. Qiu[12] and Qiu[48] provided an extensive overview of Phase II SPC for univariate and multivariate processes. In this article, it is assumed that a Phase I *reference* data are available that could be used for Phase II monitoring.

It is well established in SPC literature that time-weighted control charts such as the EWMA and CUSUM charts are better suited in detecting persistent and smaller shifts, whilst the Shewhart charts excel in detecting transient and larger shifts.[12] As we do not need to estimate the weight parameters as in EWMA and CUSUM charts,[49,50] Shewhart charts are relatively easier to implement in practice. In this article, we explore the development of a multivariate Shewhart chart based on the RE^2 statistic. Time-weighted variations could be considered separately as topics for future research. The main contributions of this paper could be listed as follows:

(i) We propose a new multivariate, distribution-free Shewhart monitoring scheme for online process monitoring based on the rank-energy (RE^2) statistic;
(ii) We demonstrate that the proposed multivariate SPC method is robust whilst maintaining reasonable detection ability for the process shift, which is justified by rigorous simulation experiments;
(iii) Practical application and efficiency of the proposed methodology is demonstrated by considering the sensorless drive fault detection case study.

In Table 1, we provide the abbreviations and corresponding full forms for the ease of reading. The rest of the article is organised as follows: In Section 2, we discuss the rank-energy (RE^2) statistic in connection with the theory of measure transportation. In Section 3, process monitoring based on the proposed method is discussed. A detailed performance study has been carried out in Section 4. In Section 5, a real-life application of the proposed method is provided regarding sensorless drive fault detection. Finally, some concluding remarks are made in Section 6.

2 | PRELIMINARIES

The proposed multivariate SPC method is based on the rank-energy (RE^2) statistic[43] discussed in this section. The RE^2 statistic is based on the notion of measure transportation that we briefly discuss next. Hallin[45] has provided a simple formulation of Monge's problem (also known as the optimal transportation problem): Let $\mathcal{R}^d$ be the d-dimensional Euclidean space, and $\mathcal{F}$ be the family of all probability distributions defined on $\mathcal{R}^d$ and $\mathcal{P}_1, \mathcal{P}_2 \in \mathcal{F}$. For $\boldsymbol{X} \in \mathcal{R}^d$, $V : \mathcal{R}^d \to \mathcal{R}^d$, the

measure transportation problem is about finding an optimal transport map $\boldsymbol{V}$ that minimises.

$$\inf_{V} \int_{R^d} ||\boldsymbol{x} - \boldsymbol{V}(\boldsymbol{x})||^2 d\mathcal{P}_1 \text{ subject to } \boldsymbol{V}\#\,\mathcal{P}_1 = \mathcal{P}_2, \tag{1}$$

where $\boldsymbol{V}\#\,\mathcal{P}_1 = \mathcal{P}_2$ indicates that for $\boldsymbol{X} \sim \mathcal{P}_1$, $\boldsymbol{V}(\boldsymbol{X}) \sim \mathcal{P}_2$, for $\mathcal{P}_1, \mathcal{P}_2 \in \mathcal{F}$. The optimal solution $\boldsymbol{V}$ of the optimisation problem in Equation (1), if it exists, is referred to as the *optimal transport map*. Using this concept, Deb and Sen[43] defined an *empirical rank function*, also called the *empirical rank map*.

Let $\mathcal{D}_n^{\tilde{\boldsymbol{X}}} = \{\tilde{\boldsymbol{X}}_1, \tilde{\boldsymbol{X}}_2, \ldots, \tilde{\boldsymbol{X}}_n\}$ be a set of random vectors following distribution $\boldsymbol{F}_{\boldsymbol{n}}^{\boldsymbol{x}}$, where $\tilde{\boldsymbol{X}}_i \in \mathcal{R}^d$, $i = 1, 2, \ldots, n$. Also let $\mathcal{G}_n = \{\boldsymbol{g}_1, \boldsymbol{g}_2, \ldots, \boldsymbol{g}_n\}$ be the set of Halton sequence[51] of dimension d to be taken as the sample multivariate rank vectors. The empirical rank map[43] is defined as a function $\widehat{\mathcal{R}}_n^{\tilde{\boldsymbol{x}}} : \mathcal{D}_n^{\tilde{\boldsymbol{x}}} \to \mathcal{G}_n$ that is an optimal transport map from $\boldsymbol{F}_{\boldsymbol{n}}^{\boldsymbol{x}}$ to $\mathcal{U}_n$ (the empirical distribution on $\mathcal{G}_n$), that is,

$$\widehat{\mathcal{R}}_n^{\tilde{\boldsymbol{x}}} = argmin_V \int_{R^d} ||\tilde{\boldsymbol{x}} - V(\tilde{\boldsymbol{x}})||^2 d\boldsymbol{F}_{\boldsymbol{n}}^{\boldsymbol{x}} \text{ subject to } V\#\,\boldsymbol{F}_{\boldsymbol{n}}^{\boldsymbol{x}} = \mathcal{U}_n. \tag{2}$$

Note that Equation (2) is equivalent to the following optimisation problem:

$$\psi = argmin_\sigma \sum_{i=1}^{n} \left|\left|\tilde{\boldsymbol{x}}_{\boldsymbol{i}} - \boldsymbol{g}_{\sigma(i)}\right|\right|^2, \tag{3}$$

where σ is the set of all permutations of {1, 2, 3,…, n} and $\sigma(i) \in \sigma\ \forall i$. The empirical rank of $\tilde{\boldsymbol{x}}_{\boldsymbol{i}}$ is

$$\widehat{\mathcal{R}}_n^{\tilde{\boldsymbol{x}}}(\tilde{\boldsymbol{x}}_{\boldsymbol{i}}) = \boldsymbol{g}_{\hat{\sigma}(i)} \text{for } i = 1, 2, \ldots,\ n, \tag{4}$$

where $\{\hat{\sigma}(i) \in \sigma, i = 1, 2, \ldots, n\}$ minimises ψ in Equation (3). Computation of $\hat{\sigma}(i)$ in Equation (4) is not time-consuming. We have used the R package 'randtoolbox'[52] from the Comprehensive R Archive Network (CRAN) to generate Halton sequences. The optimisation task completes in 43 s with a computer equipped with a Core i7 processor, for a random sample of size 10 and dimension 20.

Based on the optimal rank map, Deb and Sen[43] have defined the RE^2 statistic for multivariate two sample test. Let us consider two d-dimensional datasets given by $\mathcal{D}_{n_1}^1 = \{\tilde{\boldsymbol{X}}_{11}, \tilde{\boldsymbol{X}}_{12}, \ldots, \tilde{\boldsymbol{X}}_{1\boldsymbol{n}_1}\} \sim^{iid} \boldsymbol{F}_1$, and $\mathcal{D}_{n_2}^2 = \{\tilde{\boldsymbol{X}}_{21}, \tilde{\boldsymbol{X}}_{22}, \ldots, \tilde{\boldsymbol{X}}_{2\boldsymbol{n}_2}\} \sim^{iid} \boldsymbol{F}_2$, respectively, where $\boldsymbol{F}_1$ and $\boldsymbol{F}_2$ are d-dimensional absolutely continuous distribution functions. We want to test the following hypothesis,

$$H_0 : \boldsymbol{F}_1 = \boldsymbol{F}_2 \text{ ag. } H_1 : \boldsymbol{F}_1 \neq \boldsymbol{F}_2. \tag{5}$$

The joint empirical distribution function of $\mathcal{D}_{n_1}^1$ and $\mathcal{D}_{n_2}^2$ is given by $\boldsymbol{F}_{\boldsymbol{n}_1,\boldsymbol{n}_2}^{\tilde{\boldsymbol{x}}_1,\tilde{\boldsymbol{x}}_2}$. Let $\mathcal{G}_{n_1+n_2}$ be the Halton sequence of size $(n_1 + n_2)$ and dimension d with empirical distribution $\mathcal{U}_{n_1+n_2}$. The joint empirical rank map $\widehat{\mathcal{R}}_{n_1,n_2}^{D^1D^2}(.)$ is defined as an optimal transport map from $\boldsymbol{F}_{\boldsymbol{n}_1,\boldsymbol{n}_2}^{\tilde{\boldsymbol{x}}_1,\tilde{\boldsymbol{x}}_2}$ to $\mathcal{U}_{n_1+n_2}$, following a similar definition as the empirical rank map in Equation (4). The rank-energy statistic[43] is defined as follows:

$$\begin{aligned} RE_{n_1,n_2}^2 = {} & \frac{2}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \left|\left| \widehat{\mathcal{R}}_{n_1,n_2}^{D^1D^2}(\tilde{\boldsymbol{x}}_{1i}) - \widehat{\mathcal{R}}_{n_1,n_2}^{D^1D^2}(\tilde{\boldsymbol{x}}_{2j}) \right|\right| \\ & - \frac{1}{n_1^2} \sum_{i,j=1}^{n_1} \left|\left| \widehat{\mathcal{R}}_{n_1,n_2}^{D^1D^2}(\tilde{\boldsymbol{x}}_{1i}) - \widehat{\mathcal{R}}_{n_1,n_2}^{D^1D^2}(\tilde{\boldsymbol{x}}_{1j}) \right|\right| \\ & - \frac{1}{n_2^2} \sum_{i,j=1}^{n_2} \left|\left| \widehat{\mathcal{R}}_{n_1,n_2}^{D^1D^2}(\tilde{\boldsymbol{x}}_{2i}) - \widehat{\mathcal{R}}_{n_1,n_2}^{D^1D^2}(\tilde{\boldsymbol{x}}_{2j}) \right|\right|. \end{aligned} \tag{6}$$

To have an intuitive understanding of why the empirical rank map in Equation (4) obtained from the optimisation problem in Equation (3), and the corresponding statistics defined in Equation (6), thereafter, are distribution-free, we

consider an analogy within a one-dimensional setting. As discussed in Deb and Sen,[43] in one-dimensional setting, let $X_1, X_2, \ldots, X_n \in \mathcal{R}$ be *iid* random samples from some univariate distribution F and the ranks of these observations could be interpreted as a solution of the following optimisation problem given by

$$\hat{\sigma} = argmin_{\sigma} \sum_{i=1}^{n} \left| x_i - \frac{\sigma(i)}{n} \right|^2 \tag{7}$$

where σ is the set of all permutations of {1,2,3,…, n} and $\sigma(i) \in \sigma \ \forall i$. This interpretation of univariate ranking could be extended to multivariate settings as an optimal transport mapping problem in Equation (2).[43] Intuitively, the permutation $\hat{\sigma}$ that provides a solution of the optimisation problem in Equation (7) depends only on the observed values $x_1, x_2, \ldots, x_n$ irrespective of the distribution F.

Note that the sum of squares in Equation (7) represents the Euclidean distance between the observed univariate $\boldsymbol{X}$-sample and the set $\{\frac{\sigma(1)}{n}, \frac{\sigma(2)}{n}, \ldots, \frac{\sigma(n)}{n}\}$. This concept could be readily extended to find the Euclidean distance between random vectors $\mathcal{D}_n^{\tilde{\boldsymbol{x}}}$ and the set of Halton sequence $\mathcal{G}_n$,[51] and the solution of the optimal transport mapping problem should be free from the distribution of the random vectors $\mathcal{D}_n^{\tilde{\boldsymbol{x}}}$, intuitively. Indeed, Deb and Sen[43] proved that for the absolutely continuous $\boldsymbol{F}_1$ and $\boldsymbol{F}_2$, the rank-energy statistic $RE^2_{n_1,n_2}$ is *exactly distribution-free* and invariant under affine transformations. The code to calculate the RE^2 statistic can be found on the GitHub page of the first author of Deb and Sen.[43] We will use the rank energy statistic defined in Equation (6) for the monitoring plan to be developed in what follows.

3 | MONITORING FRAMEWORK

In order to use the $RE^2_{n_1,n_2}$ statistic in Equation (6) in an online monitoring scheme, it is assumed that a stable reference data is available against which the $RE^2_{n_1,n_2}$ statistic is obtained sequentially. It is recommended to confirm the stability of the hardware/equipment from a well-established system, to ensure the stability of the reference data.

Let us consider a reference sample $\mathcal{D}_{\boldsymbol{m}}^{\tilde{\boldsymbol{X}}} = \{\tilde{\boldsymbol{X}}_1, \tilde{\boldsymbol{X}}_2, \ldots, \tilde{\boldsymbol{X}}_{\boldsymbol{m}}\} \sim^{iid} \boldsymbol{F}_1$ of size m. During the online monitoring, subgroups $\mathcal{D}_{\boldsymbol{n,t}}^{\tilde{\boldsymbol{Y}}} = \{\tilde{\boldsymbol{Y}}_{1t}, \tilde{\boldsymbol{Y}}_{2t}, \ldots, \tilde{\boldsymbol{Y}}_{nt}\} \sim^{iid} \boldsymbol{F}_2$ of size n are obtained at time instances $t = 1, 2, 3, \ldots$ The online monitoring scheme therefore can be viewed as an online hypothesis testing problem given by

$$\begin{aligned} &H_0: \boldsymbol{F}_1 = \boldsymbol{F}_2 \\ &\text{ag. } H_1: \boldsymbol{F}_1 \neq \boldsymbol{F}_2, \end{aligned} \tag{8}$$

For each $t = 1, 2, 3, \ldots$, the rank energy statistic as defined in Equation (6), is calculated on $\mathcal{D}_{\boldsymbol{m}}^{\tilde{\boldsymbol{X}}}$ and $\mathcal{D}_{\boldsymbol{n,t}}^{\tilde{\boldsymbol{Y}}}$ and denoted by $RE^2_{m,n,t}$. When $RE^2_{m,n,t} \geq C$, for some *decision limit* $C > 0$, the process is deemed out-of-control (OC) and process monitoring is halted. Else the process is considered in-control (IC). The event $[RE^2_{m,n,t} \geq C]$ is called a *signalling* event. When a signal is detected, the system undergoes inspection. If the signal results from a system anomaly, the system will be recalibrated, and online monitoring will continue. If the signal is due to a random cause of variation, it is regarded as a *false alarm*, and no remedial action is required, and the online process monitoring continues.

Choice of the decision limit C is important in online process monitoring. The number of subgroups $\mathcal{D}_{\boldsymbol{n,t}}^{\tilde{\boldsymbol{Y}}}$ obtained till $RE^2_{m,n,t} \geq C$ for some t, is called the *run length*. A standard performance measure in online process monitoring is the *average run length* (*ARL*).[53] To decide the decision limits C for different sample sizes m, n and different dimension d, a nominal *ARL* value of $ARL_0 \approx 370$ is adopted when the process is IC. An efficient monitoring scheme should ideally detect faults early in Phase II. As a result, when the process is OC, the corresponding *ARL*, denoted by ARL_1, should be less than ARL_0. An analytical expression for the *ARL* is not available due to the density of the $RE^2_{m,n,t}$ statistic lacking a closed-form expression. We estimate the *ARL* under $H = H_0$ and H_1 by numerical integration via the Monte Carlo simulation, that is,

$$ARL = \int_0^{\infty} \frac{1}{P_H\left[RE^2_{m,n} > C | \mathcal{D}_{\boldsymbol{m}}^{\tilde{\boldsymbol{X}}}\right]} d\boldsymbol{F}_1, \tag{9}$$

where $P_H[RE^2_{m,n} > C | \mathcal{D}_{\boldsymbol{m}}^{\tilde{\boldsymbol{X}}}]$ is the conditional power estimated under $H = H_0$ and H_1. For a given decision limit C, the ARL_0 estimation involves averaging 100,000 values of corresponding conditional ARL_0, given specific values of m, n, d.

The decision limit is then iteratively adjusted to ensure that the obtained $ARL_0 \approx 370$. To estimate the ARL_1 for C and specific values of m, n and d, we estimate the conditional power $P_{\boldsymbol{H}}[RE^2_{m,n} > C|\boldsymbol{D}^{\tilde{\boldsymbol{X}}}_{\boldsymbol{m}}]$ for a distribution $\boldsymbol{F}_2$ different from $\boldsymbol{F}_1$, and then averaging over 100,000 values of conditional ARL_1. Other nominal choices for the ARL_0 could be also considered. The algorithms to determine the decision limits and the corresponding ARL_0 and ARL_1 are detailed in the Appendix A1 and A2.

4 | PERFORMANCE EVALUATION

The monitoring framework, based on the RE^2 statistic,[43] is quite general in the sense that it can be applied to monitoring any multivariate process, including the high-dimensional processes. This will be discussed later. As the spatial sign or depth-based monitoring methods are not distribution-free (as discussed in Section 1), the proposed monitoring scheme is evaluated and compared with a recent nonparametric multivariate monitoring scheme that do not rely on spatial sign or depth-based statistics. Given that the density of the rank-energy statistic defined by Equation (6) does not have an analytic form, the proposed monitoring scheme's efficacy is examined numerically via the Monte–Carlo simulation.

4.1 | Robustness study

Let us consider a reference data of size $m = 10, 20$ and test data of size $n = 5$ from a multivariate normal distribution of dimension $d = 20$ with mean $\boldsymbol{\mu} = 0_{d\times 1}$ and covariance matrix $\boldsymbol{\Sigma}_{d\times d} = (\sigma_{ij})_{d\times d}$ such that $\sigma_{ij} = \sigma^2\,(\frac{\min(i,j)}{\max(i,j)})$, where $\sigma = 1.5$. To numerically evaluate the robustness of the proposed monitoring scheme, we estimate the ARL_0 for various multivariate distributions, using the decision limit obtained for the aforementioned multivariate normal distribution. As symmetric distributions, we consider multivariate normal distributions with different covariance matrices, and as skewed distributions, we consider multivariate exponential distributions 'connected' by copulas. Copula modelling is widely used in practice in modelling the dependency structure in multivariate data.[54,55] We consider the Gaussian copula as an elliptical copula and the Clayton copula as an Archimedean copula as they are popular in literature for their flexibility and simple form.[54,55] Other copulas could be considered as well. However, for the purpose of the performance study, we restrict ourselves to the copulas mentioned above. The covariance structure for the Gaussian copula is taken as $\boldsymbol{\Sigma}_{d\times d} = (\rho_{ij})_{d\times d}$ where $\rho_{ij} = (\frac{\min(i,j)}{\max(i,j)})$. Clayton copula parameter is taken as $\xi = 2$. A larger ξ implies a stronger dependence among the end points. A brief discussion on copulas is provided in the Appendix A3. For more details on copula modelling, one may refer to the monographs by.[54–56] Thus, the distributions considered are:

(i) Multivariate normal distribution with mean $\boldsymbol{\mu} = 0_{d\times 1}$ and covariance matrix $\boldsymbol{\Sigma}_{d\times d} = (\sigma_{ij})_{d\times d}$ such that $\sigma_{ij} = \sigma^2\,(\frac{\min(i,j)}{\max(i,j)})$, where $\sigma = 2, 2.5, 3$;
(ii) d-dimensional exponential distribution with marginal distributions as Exponential (1) distribution connected by Gaussian copula and Clayton copula.

The estimated ARL_0 values corresponding to $RE^2_{10,5}$ and $RE^2_{20,5}$ statistics are presented in Table 2. The critical limits are rounded off to 3 decimal places. It can be noted from Table 2 that the ARL_0 performance of the proposed monitoring scheme is quite robust across different distributions. Note that, in the robustness study, we considered $m = 10$ and $d = 20$ in Table 2. This suggests that the proposed monitoring scheme is suitable in 'high dimension, low sample size' settings.

4.2 | Anomaly detection

In this section, we conduct a numerical study to evaluate the anomaly detection capability of the proposed monitoring scheme. Table 3 provides decision limits for $m = 10, 15, 20, 25, 30, 40, 50, 100$; $n = 3, 5$ and $d = 2(1)10, 20, 30, 40, 50, 100$ such that $ARL_0 \approx 370$. As in Table 2, the decision limits C increases with dimension d. In what follows, we discuss the performance of the proposed method under various process shifts.

TABLE 2 ARL_0 for $RE^2_{m,n}$ for d-dimensional multivariate normal and exponential distributions, for $m = 10, 20$, $n = 5$, $d = 20$.

	$RE^2_{10,5}$ **Critical limit = 2.810**	$RE^2_{20,5}$ **Critical limit = 2.874**
Multivariate normal distribution		
$\sigma = 1.5$	369.00	367.07
$\sigma = 2$	365.97	370.21
$\sigma = 2.5$	372.36	369.64
$\sigma = 3.0$	376.86	370.25
Multivariate exponential distribution		
Gaussian copula	368.27	368.71
Clayton copula	373.01	369.83

TABLE 3 Decision limits C for different (m, n, d).

	$m = 10$	$m = 15$	$m = 20$	$m = 25$	$m = 30$	$m = 40$	$m = 50$	$m = 100$
$n = 3$								
d	C							
2	1.287	1.472	1.525	1.496	1.525	1.597	1.592	1.654
3	2.006	1.967	1.747	1.690	1.590	1.682	1.663	1.755
4	1.857	1.744	1.680	1.612	1.599	1.816	1.772	1.800
5	1.810	1.775	1.713	1.670	1.655	1.738	1.843	1.826
6	1.850	1.916	1.866	1.761	1.766	1.880	1.871	1.885
7	1.744	1.949	1.894	1.866	1.854	1.937	1.945	1.952
8	1.930	2.105	1.992	1.986	2.024	2.047	2.042	2.029
9	2.131	2.086	2.068	2.108	2.130	2.176	2.106	2.162
10	2.184	2.488	2.344	2.265	2.186	2.290	2.173	2.179
20	2.936	2.746	2.871	2.871	2.843	2.848	2.800	2.722
30	3.285	3.419	3.374	3.504	3.287	3.279	3.265	3.131
40	3.546	3.858	3.752	3.829	3.731	3.675	3.632	3.471
50	3.792	4.083	3.984	4.112	3.918	3.988	3.965	3.852
100	5.373	5.449	5.417	5.637	5.511	5.564	5.469	5.362
$n = 5$								
2	1.712	1.753	1.733	1.725	1.754	1.748	1.750	1.703
3	1.763	1.884	1.802	1.765	1.728	1.762	1.790	1.794
4	1.755	1.792	1.776	1.743	1.885	1.848	1.897	1.906
5	1.852	1.900	1.837	1.783	1.846	1.925	1.917	1.945
6	1.785	1.942	1.912	1.887	1.964	2.026	2.008	2.013
7	1.846	2.027	2.001	1.934	2.028	2.002	2.020	2.040
8	1.951	2.204	2.160	2.070	2.155	2.116	2.116	2.176
9	2.057	2.189	2.269	2.195	2.228	2.227	2.165	2.200
10	2.182	2.365	2.367	2.242	2.317	2.284	2.207	2.260
20	2.812	2.808	2.897	2.896	2.847	2.780	2.748	2.721
30	3.123	3.355	3.297	3.306	3.344	3.284	3.249	3.114
40	3.623	3.732	3.672	3.841	3.763	3.691	3.671	3.548
50	3.929	4.038	3.944	4.096	4.051	3.975	3.898	3.822
100	5.341	5.368	5.318	5.405	5.344	5.408	5.438	5.270

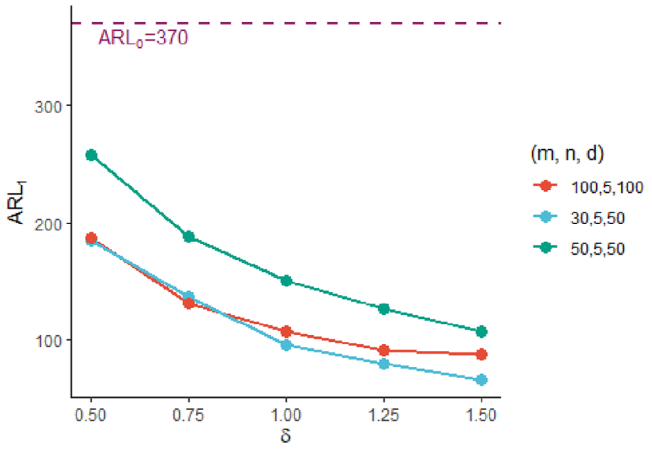

FIGURE 1 ARL_1 for different shift δ in location vector of high-dimensional Gaussian process.

4.2.1 | Location shift in multivariate Gaussian process

In the manufacturing industry, Gaussian processes are widely used in quality control, predictive maintenance, optimisation among other applications.[57,58] To justify the utility of the proposed method in detecting location shift, we consider $d = 50, 100$-dimensional Gaussian process. At time instances t, $\{\tilde{\boldsymbol{x}}_{it},\ \boldsymbol{i} = 1(1)\boldsymbol{n}\}$ are assumed to be temporally independent and follow a shifted multivariate normal distribution with mean

$$\boldsymbol{\mu}_1 = \boldsymbol{\mu}_0 + \boldsymbol{\delta}, \text{where } \boldsymbol{\delta} = \delta 1_{d\times 1}, \tag{10}$$

and the covariance matrix

$$\boldsymbol{\Sigma}_0 = \left((\sigma_{ij})\right)_{d\times d}, \tag{11}$$

where $\sigma_{ij} = \sigma^2$ for $i = j$, and $\sigma_{ij} = \rho\sigma^2$ for $i \neq j$.

We consider $\rho = 0.5$, $\sigma = 1$ and $\boldsymbol{\mu}_0 = (0)_{d\times 1}$. For $\delta = 0.5, 0.75, 1.0, 1.25, 1.5$, ARL_1 values are plotted for $(m, n, d) =$ (30,5,50), (50,5,50), (100,5,100) in Figure 1. Note that, in Figure 1, for a high-dimensional Gaussian process with $d = 50$, 100 and $d \geq m$, ARL_1 values decrease with increasing location shift δ. For $m = 30$, shift detection is quicker than $m = 50$, 100 for larger shift. This shows that the proposed monitoring scheme is useful in monitoring high-dimensional processes for location shift.

4.2.2 | Shift in the Gaussian process variance

Let us consider a multidimensional Gaussian process with dimension $d = 3, 30, 100$. For all time instances t, $\{\tilde{\boldsymbol{x}}_{it},\ \boldsymbol{i} = 1(1)\boldsymbol{n}\}$ independently follow a multivariate normal distribution with mean $\boldsymbol{\mu}_0$, and the covariance matrix

$$\boldsymbol{\Sigma}_1 = \left(\sigma_{1ij}\right)_{d\times d}, \tag{12}$$

where $\sigma_{1ij} = \Delta\sigma^2$ for $i = j$, and $\sigma_{ij} = \Delta\rho\sigma^2$ for $i \neq j$.

Let $\rho = 0.5$, $\sigma = 1$ and $\Delta = 0.5, 1, 1.5, 2$. For the process mean, we consider $\boldsymbol{\mu}_0 = (0)_{d\times 1}$. For $m = 50, 100$ and $n = 5$, ARL_1 values are plotted in Figure 2 against different Δ, for different d. Note that, as in Figure 1, the proposed method is capable in detecting variance shift for a high-dimensional Gaussian process when $d \geq m$. However, when dimension d is much higher than the reference sample size m, ARL_1 values also increase causing delay in detecting shift. Also, for increasing variance shift, the ARL_1 values do not decrease as fast as in the case of location shift. Hence, to detect a variance shift in a high-dimensional data with $d \geq 100$, it is recommended to not have a small reference sample.

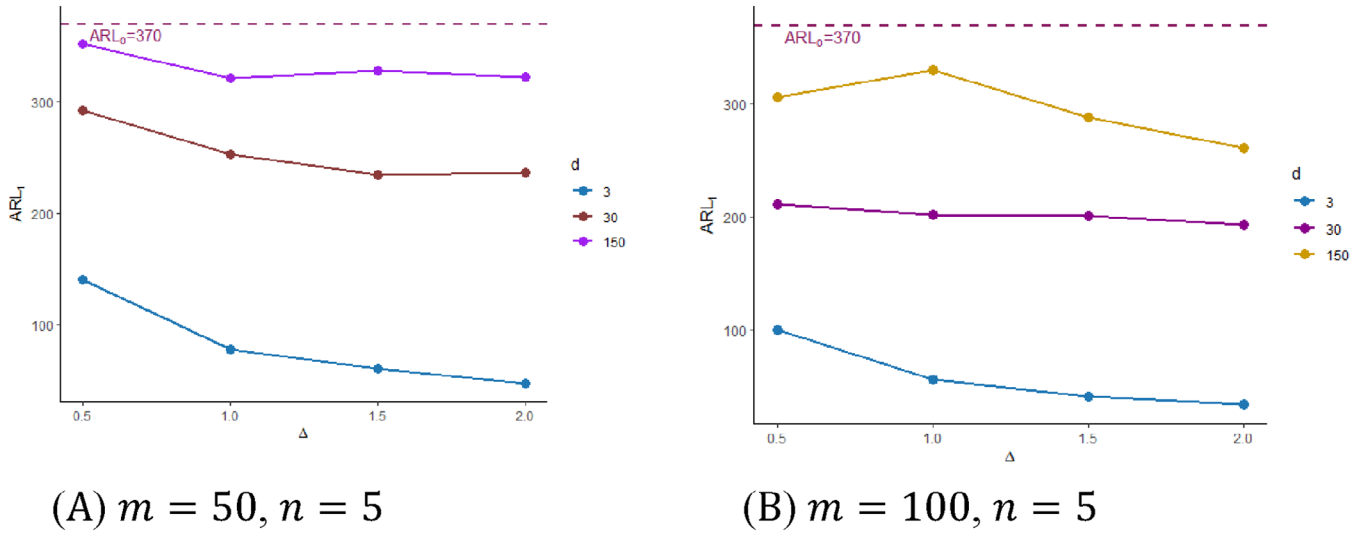

FIGURE 2 ARL_1 for different shift Δ in variance for multivariate normal distribution.

4.2.3 | Shift in the failure rate of multivariate exponential distribution

Failure rate modelling and monitoring is popular in reliability and process control literature.[59–61] In this section, we consider multivariate exponential distributions where the marginal failure rates have shifted from $\lambda_0 = 1$, without loss of generality, to $\lambda_1 = \lambda_0\,\delta$, for $\delta > 1$. The marginal exponential distributions are assumed to be connected by Clayton copula with parameter $\xi = 1$. In Figure 3, a decreasing trend is observed in the ARL_1 values plotted for $\delta = 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5$. However, ARL_1 values seldom encountered bias, especially for smaller values of δ. The empirical study in Figure 3 suggests that monitoring a highly skewed process with complex dependence structure would require a substantial number of reference data, preferably more than the dimension.

4.2.4 | Comparative remarks

In this section, we compare results of Mukherjee and Marozzi[29] on Phase II process monitoring based on the interpoint-distance methods with the proposed method (since they do not depend on the spatial sign or rank-based statistics). For readers' sake, we use similar notations as in Mukherjee and Marozzi[29] to briefly explain their method.

In essence, Mukherjee and Marozzi[29] have examined three distinct ways to calculate interpoint distances, each coupled with three different univariate distribution-free tests applied to these distance measures. One of the high-dimensional Shewhart-type (HDS) monitoring scheme in this paper called the HDSOR scheme, relies on the distance of each data point from the origin (OR). The other two HDS monitoring schemes depend on the interpoint distances from (i) a single reference data point (HDSIP), and (ii) the generalised median (HDSGM). The Wilcoxon statistic, Ansari–Bradley statistic and Lepage statistic were obtained on the interpoint distance values via the HDSOR, HDSIP and HDSGM methods. Whilst HDSGM scheme lacks robustness, it was noted that the HDSIP scheme has the uniformly better OC performance than the HDSOR scheme.[29] We consider the HDSOR and the HDSIP scheme for comparison.

Suppose that two random samples $\tilde{\boldsymbol{X}}$ and $\tilde{\boldsymbol{Y}}$ (Table 4) of size 5 are drawn from the bivariate normal distribution with mean $\boldsymbol{\mu}_X = (0,0)$ and $\boldsymbol{\mu}_Y = (1,1)$, respectively, and covariance matrix $\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$. We consider $\tilde{\boldsymbol{x}}_1$ and $\tilde{\boldsymbol{x}}_5$ as the *conditioning* observations as defined by Mukherjee and Marozzi.[29] Following the HDSIP scheme, interpoint Euclidean distances of all $\tilde{\boldsymbol{X}}$ and $\tilde{\boldsymbol{Y}}$ observations are obtained from these two points (Figure 4). 'Var1' and 'Var2' in Figure 4 refer to the first and second row elements of $\tilde{\boldsymbol{X}}$ and $\tilde{\boldsymbol{Y}}$ sample, respectively.

Based on these distance values from $\tilde{\boldsymbol{x}}_1$ and $\tilde{\boldsymbol{x}}_5$, we obtain the Wilcoxon rank-sum statistics HDSIP-$W_1 = 12$ and HDSIP-$W_2 = 21$, respectively. For two random samples of size 5, the critical value at 5% level of significance for the Wilcoxon rank-sum test is 20. Hence, for the same reference and test data, the process seems stable according to HDSIP-W_1 value, and not stable according to the HDSIP-W_2 value. Therefore, the choice of the *conditioning* observation influences a practitioner's

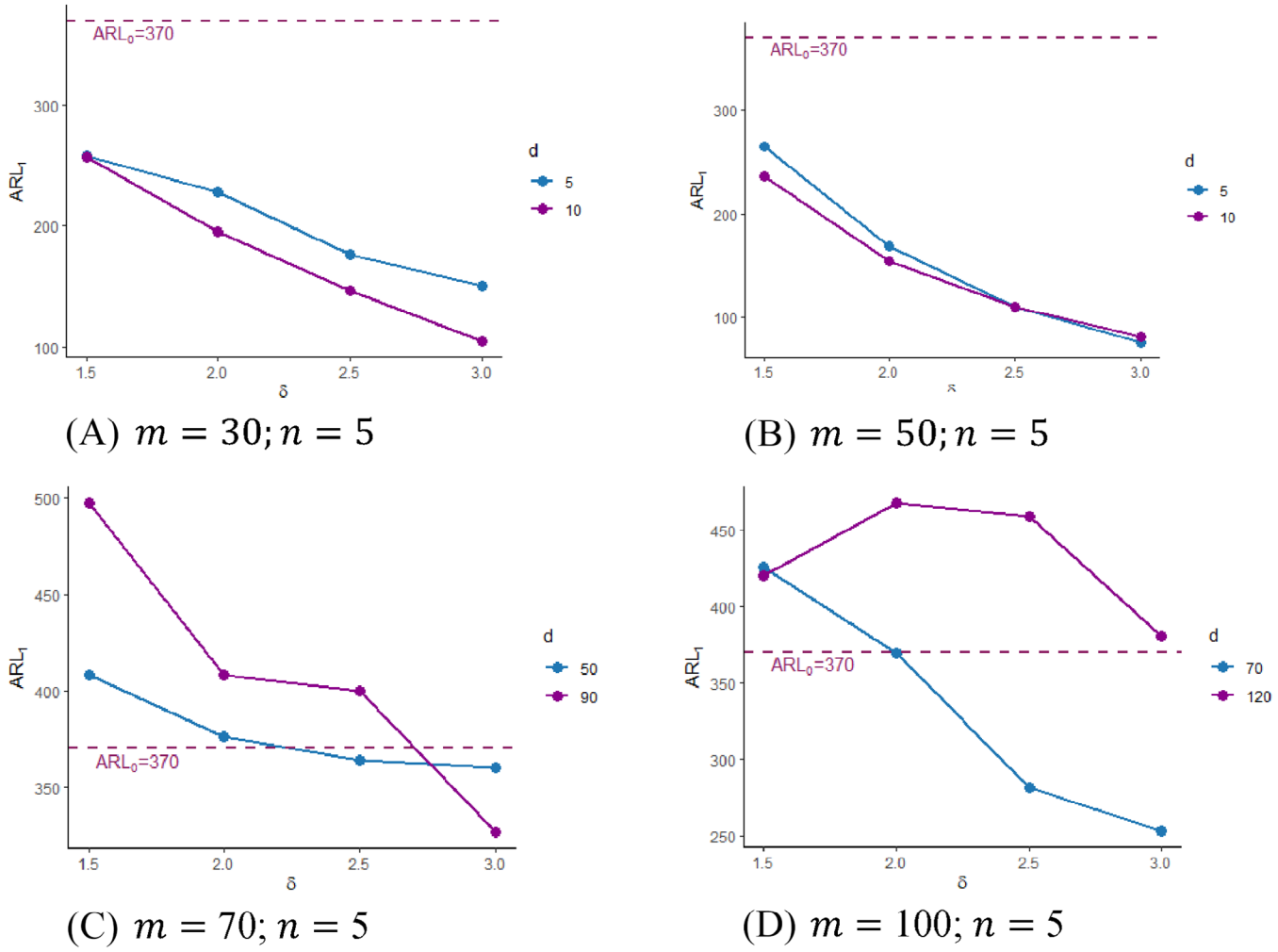

FIGURE 3 ARL_1 for different shift in failure rate λ_0 for multivariate exponential distribution connected by Clayton copula.

TABLE 4 Random samples from bivariate normal distribution with mean $\boldsymbol{\mu}_X = (0, 0)$ and $\boldsymbol{\mu}_Y = (1,1)$, respectively, and covariance matrix $\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$.

$\tilde{\boldsymbol{x}}_1$	$\tilde{\boldsymbol{x}}_2$	$\tilde{\boldsymbol{x}}_3$	$\tilde{\boldsymbol{x}}_4$	$\tilde{\boldsymbol{x}}_5$
0.688	0.872	0.102	0.254	2.185
−0.675	−2.119	−1.265	−0.374	−0.596
$\tilde{\boldsymbol{y}}_1$	$\tilde{\boldsymbol{y}}_2$	$\tilde{\boldsymbol{y}}_3$	$\tilde{\boldsymbol{y}}_4$	$\tilde{\boldsymbol{y}}_5$
2.436	0.638	2.759	1.325	1.652
−0.854	0.922	1.969	1.185	−0.380

decision. This is not desirable as the test statistic for the same reference and test data should be invariant over the choice of the *conditioning* observations. On the other hand, the rank-energy statistic $RE^2_{5,5}$ is 0.849 for the $\tilde{\boldsymbol{X}}$ and $\tilde{\boldsymbol{Y}}$ sample in Table 4 that is invariant of the practitioners' bias.

Mukherjee and Marozzi[29] demonstrated that HDSOR and HDSIP schemes, when using the Lepage statistic (called HDSOR-L and HDSIP-L, respectively), exhibit superior OC performance compared to their competitors. We next compare the OC performance of the proposed method with the HDSOR-L and HDSIP-L schemes for trivariate normal distribution with mean $\boldsymbol{\mu}_1$ (Equation 10) and covariance matrix $\boldsymbol{\Sigma}_1$ (Equation 12). The OC location and scale shift are considered as in Equation (10) and Equation (12).

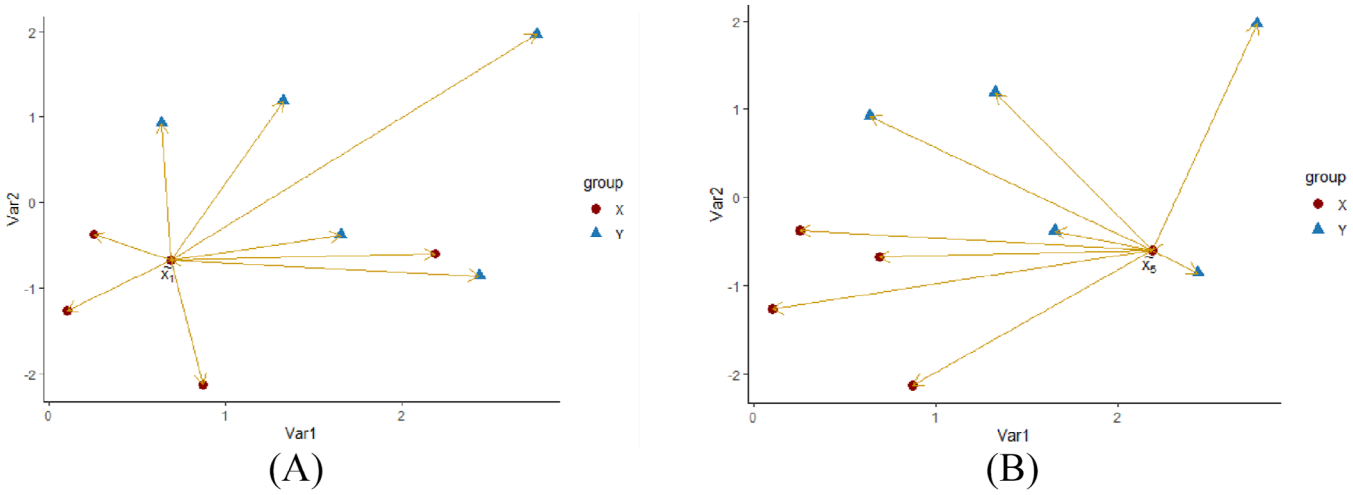

FIGURE 4 (A) Euclidean distances from $\tilde{x}_1$ with HDSIP-W$_1$ = 12; (B) Euclidean distances from $\tilde{x}_5$ with HDSIP-W$_2$ = 21.

TABLE 5 MRL_1 values for different location and scale shift for trivariate normal distribution.

	HDSOR-L	HDSIP-L	$RE^2_{m,n}$	HDSOR-L	HDSIP-L	$RE^2_{m,n}$
	$m = 100, n = 5$			$m = 300, n = 5$		
Decision limit	**11.07**	**10.99**	**1.744**	**11.39**	**11.37**	**1.778**
(δ, Δ)						
(0.5,1)	95	58	30	88	53	29
(1,1)	9	9	3	8	8	3
(0,1.5)	6	14	20	6	12	15
(0,2)	2	3	16	2	3	12
(0.5,1.5)	5	9	13	4	9	11
(0.5,2)	2	3	12	2	3	10
(1,1.5)	3	4	6	2	4	5
(1,2)	1	2	6	1	2	6

In our comparison, we consider $\rho = 0.5$, $\sigma = 1$, as in the OC setting of Mukherjee and Marozzi.[29] For location and scale shift, we consider $(\delta, \Delta) = (0.5,1), (1,1), (0,1.5), (0,2), (0.5,1.5), (1,1.5), (0.5,2), (1,2)$. It is to note that, $\delta = 0$ indicates a pure location shift and $\Delta = 1$ indicates a pure scale shift. For comparison purpose, we calculate the decision limit C for $RE^2_{m,n}$ statistic so that the IC median run length $MRL_0 = 250$. In Table 5, OC median run length (MRL_1) values for different (δ, Δ) are provided for the HDSOR-L, HDSIP-L and $RE^2_{m,n}$ schemes. Table 5 shows that the $RE^2_{m,n}$-based scheme outperforms both the HDSOR-L and the HDSIP-L schemes in detecting pure location shifts in trivariate normal distribution. However, when it comes to the scale shift, our method falls behind the other two monitoring schemes. This is because the Euclidean distance-based methods tend to be sensitive to the scale transformations and outliers. If the variables of a random vector are not on the same scale, one variable with larger value (or outliers) could influence the distance calculations.

4.3 | Sensitivity analysis

The OC performance of the proposed method depends on the dimension d, and sample sizes m and n. For Phase II monitoring, $n = 3,\ 5$, are standard choices for the test sample size. Therefore, it is of interest to see the performance of the proposed method for varying m and d for different location and scale shift in a multivariate process. We consider multivariate normal distribution to carry out a sensitivity analysis to understand the impact of the reference sample size and dimension on OC performance of the proposed method.

In Figure 5, we provide ARL_1 values for location shift (a) $\delta = 0.5$; (b) $\delta = 0.75$; (c) $\delta = 1.0$; (d) $\delta = 1.5$ in multivariate normal distribution. In Figure 6, ARL_1 values are displayed for variance shift (a) $\Delta = 0.5$; (b) $\Delta = 1.0$; (c) $\Delta = 1.5$;

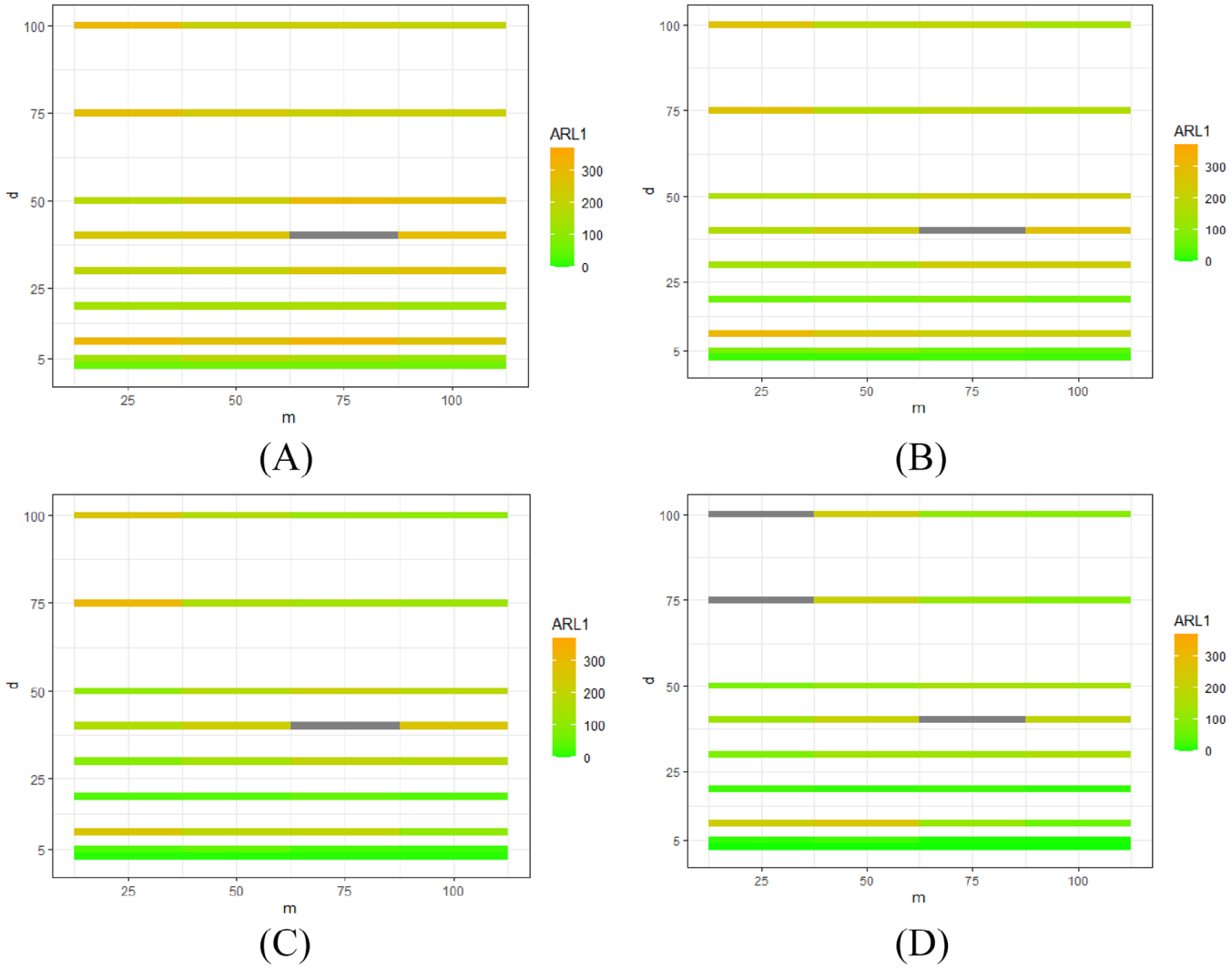

FIGURE 5 ARL_1 values for location shift (A) $\delta = 0.5$; (B)$\delta = 0.75$; (C) $\delta = 1.0$; (D)$\delta = 1.5$ in multivariate normal distribution for different sample size m and dimension d.

(d) $\Delta = 2.0$ in multivariate normal distribution. The location and the scale shift are considered as in Equations (10) and (12) with $\rho = 0.5$, $\sigma = 1$ for an IC Gaussian process.

Upon studying Figures 5 and 6, it is empirically evident that the proposed method exhibits bias, visible as the dark patches when $ARL_1 > ARL_0$, for some (m, d) combinations, particularly for smaller m. For small location shift ($\delta \leq 1$), the proposed method is able to detect the shift for high dimensional data with small reference sample size. If larger shift detection is intended, a larger reference sample size would be recommended. For a smaller scale shift in a high-dimensional data, the reference sample size of $m \geq 100$ would be recommended. For larger scale shift, the proposed method performs well for 'small reference sample, high dimension' settings. The empirical analysis suggests that, in general, for location and scale shift, a larger reference sample (such as $m \geq 100$) would be beneficial to avoid bias. It is also noteworthy that ARL_1 tends to increase when the dimension d is increasing. Consequently, we may suggest an empirical guideline for choosing (m, d) such that $\frac{d}{m} \leq 0.5$.

5 | APPLICATION IN SENSORLESS DRIVE DIAGNOSIS

Sensorless drive is a type of a control system for electric motors, commonly used in variable speed drive (VSD) applications such as heating, ventilation, air conditioning system, also known as HVAC systems, and electric vehicles. It operates

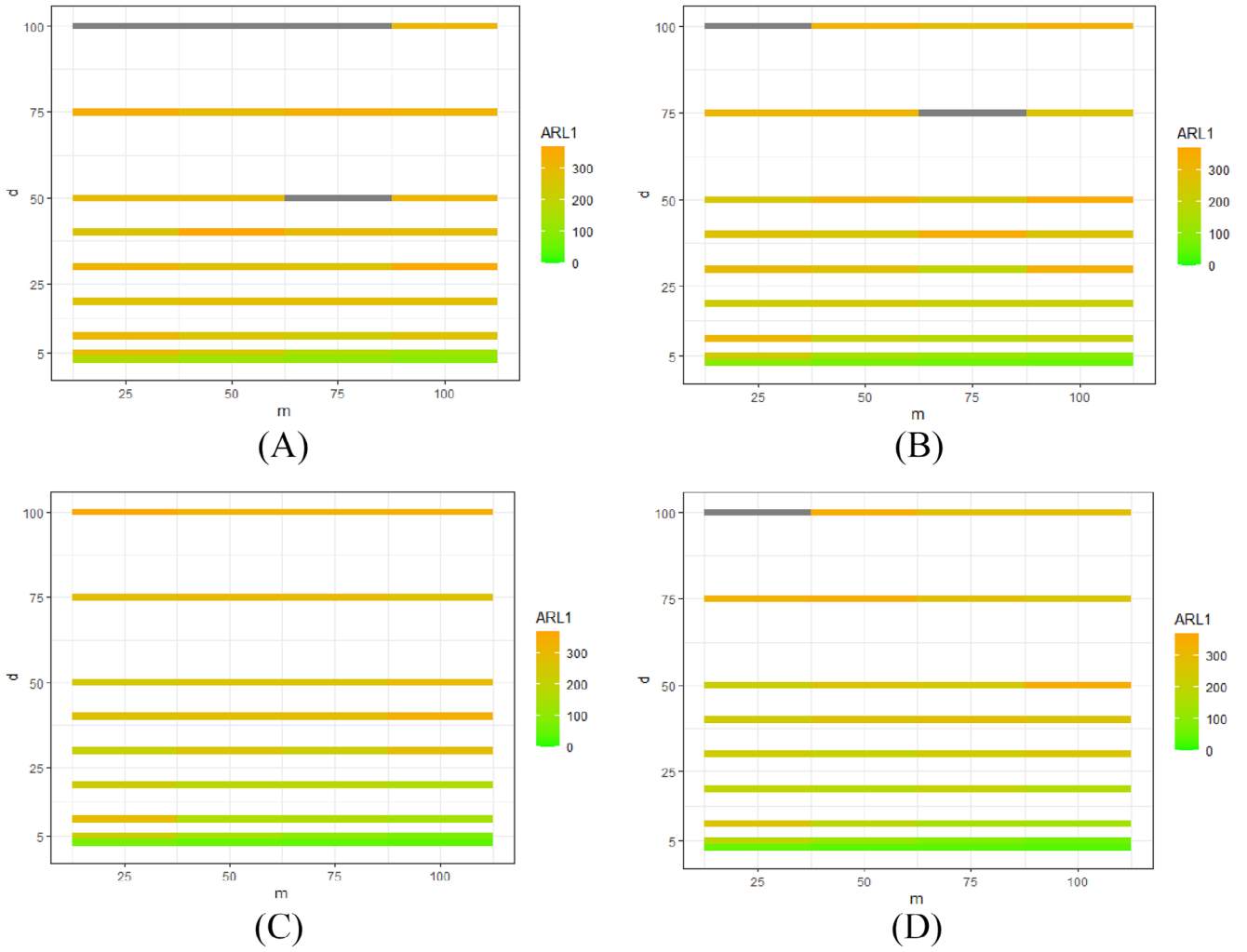

FIGURE 6 ARL_1 values for variance shift (A) $\Delta = 0.5$; (B)$\Delta = 1.0$; (C) $\Delta = 1.5$; (D)$\Delta = 2.0$ in multivariate normal distribution for different sample size m and dimension d.

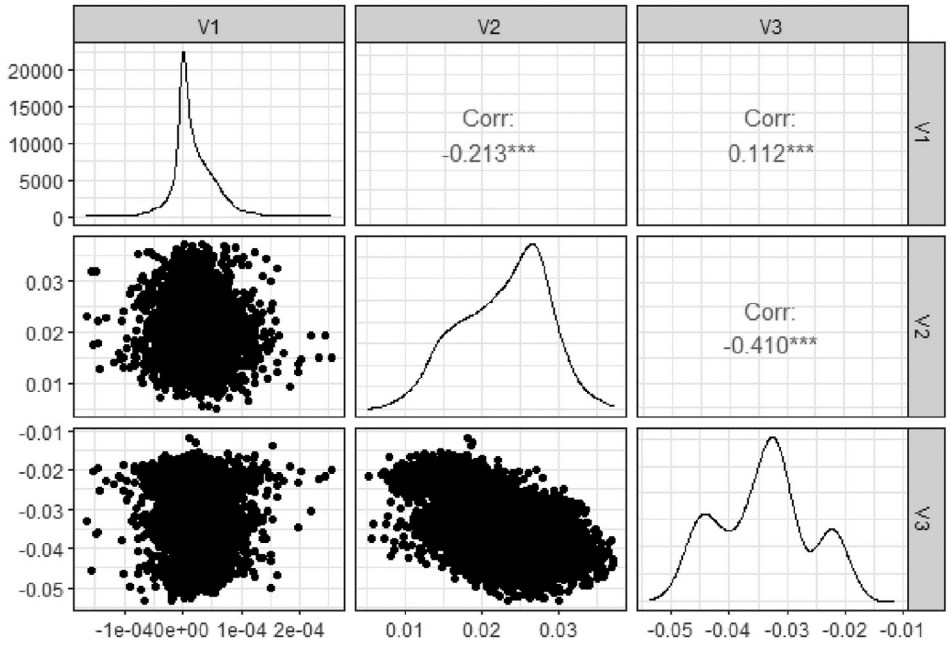

FIGURE 7 Correlation plot for the sensorless drive data.

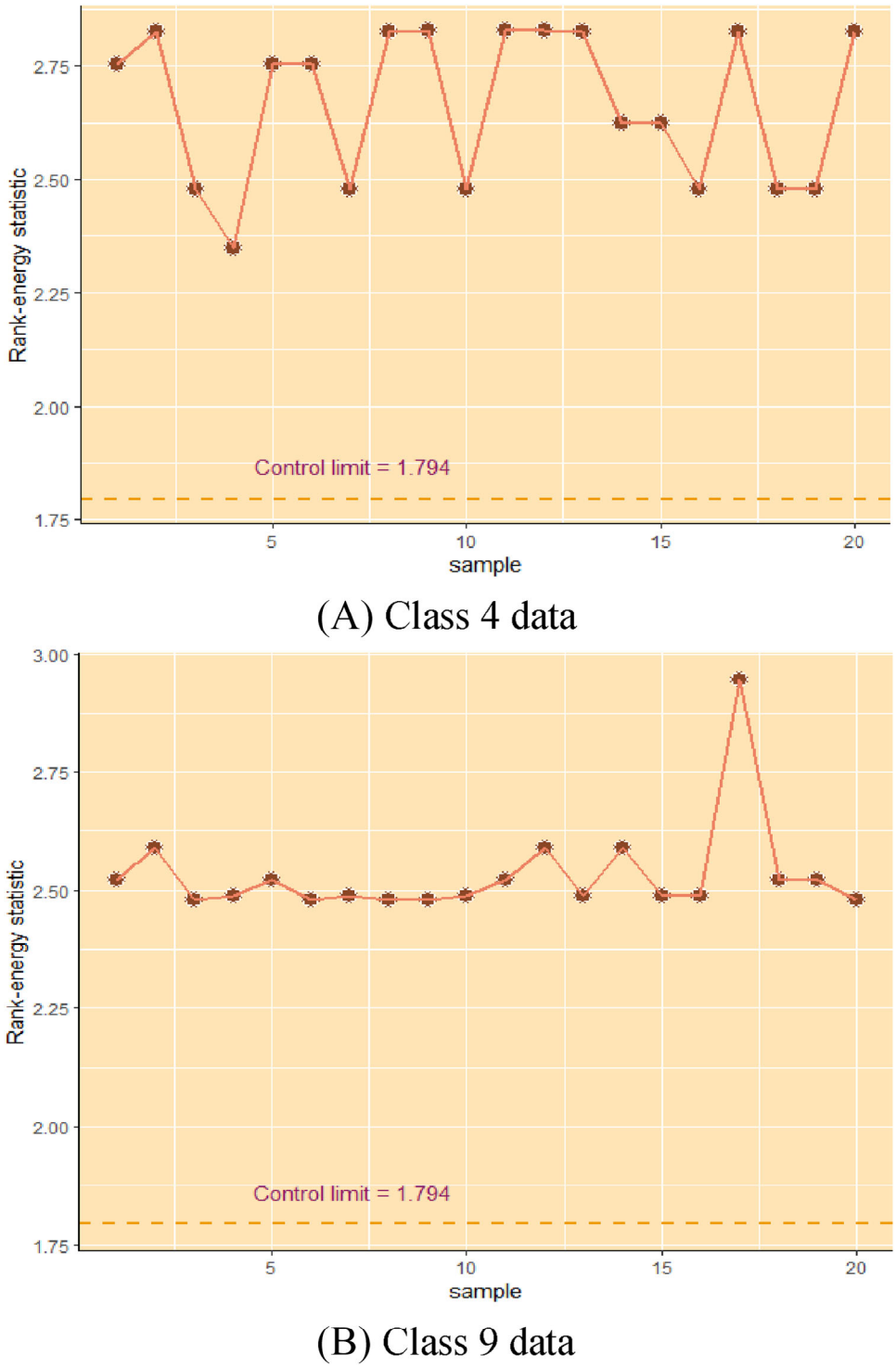

FIGURE 8 $RE^2_{100,5}$ statistics for the Class 4 and Class 9 sensorless drive data.

without using physical sensors to monitor motor parameters. In sensor-based VSD systems, physical sensors are used to monitor motor parameters such as speed, torque, or position. Whilst generally reliable, these sensor-based systems can be expensive, sensitive to environmental conditions, and require complex installation. On the other hand, sensorless drive technology uses motor's electrical parameters to estimate rotor position and speed and is widely employed in electric and hybrid vehicle applications. Recent research on sensorless drive systems includes,[62,63] among others. Sensorless drive fault detection is vital for increasing reliability and maintainability characteristics of electrical motors.

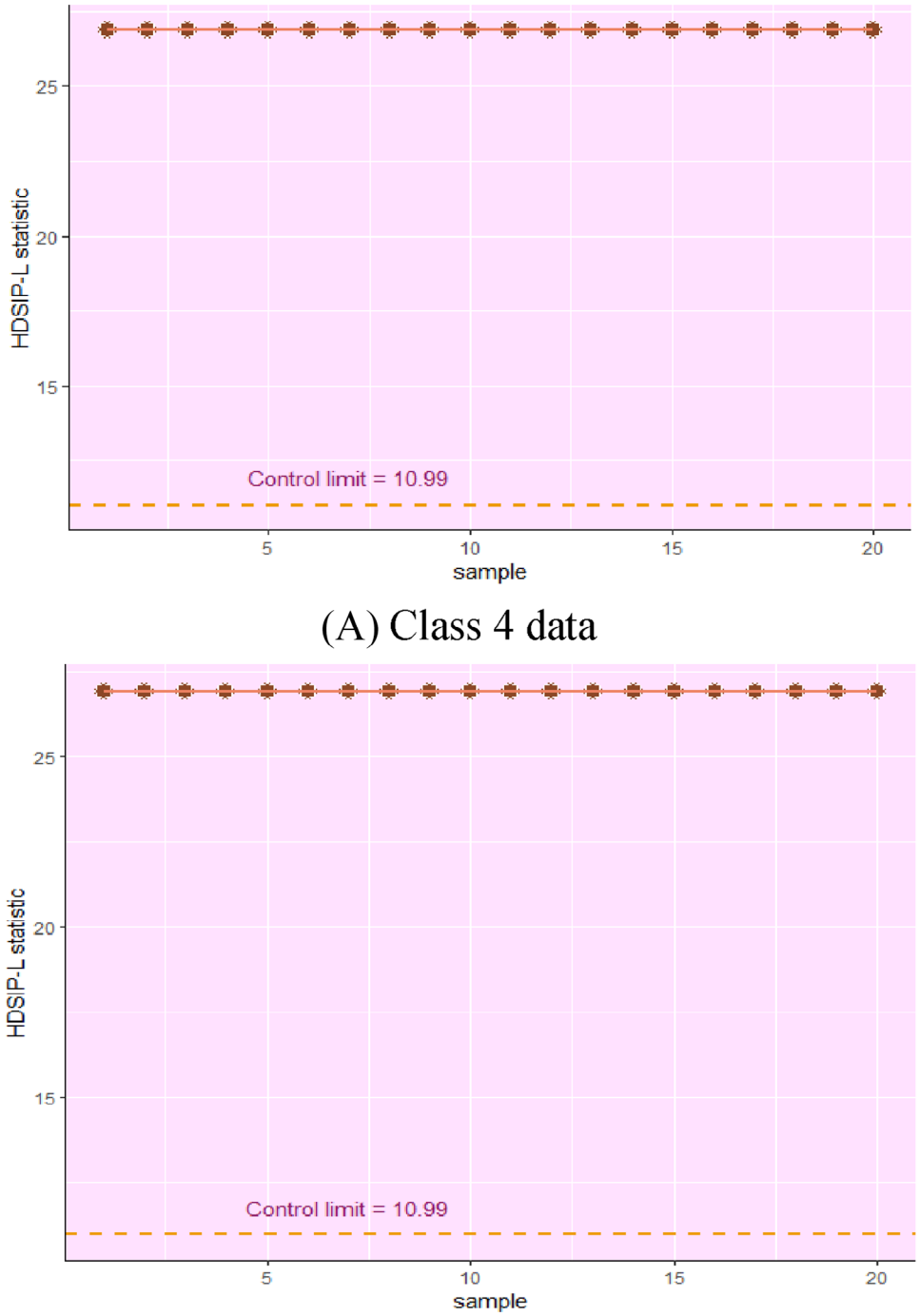

FIGURE 9 HDSIP-L values for the Class 4 and Class 9 sensorless drive data with the first reference data point taken as the *referencing* point.

To demonstrate the usage of the proposed multivariate monitoring scheme, we consider a sensorless drive diagnosis dataset from the 'rebmix' package available in the Comprehensive R Archive Network (CRAN). This dataset contains 58,509 data points (in rows) and 4 variables (in columns). Among these 4 variables, the first three variables are continuous, and the fourth one is a categorical variable denoting 11 different classes. The Class 1 corresponds to 'healthy' drives, whilst the remaining classes correspond to faulty drives. Any of these faulty drive data could be considered as a test data to illustrate the shift detection ability of the proposed method. We designate the Class 1 data as the reference data and the Class 4 and Class 9 data as the test data.

In Figure 7, we present a plot illustrating the pairwise correlation among the variables in the sensorless drive data along with their densities. The marginal densities of the continuous variables in Figure 7 show that the normality assumption for the sensorless drive data is inappropriate and therefore distribution-free SPC method should be considered. It is evident that there is significant correlation among the variables. In order to make the variables (or features) unit-free, we standardise each variable of the reference and test data. The first 100 data points from Class 1 are considered as the reference data and the data points from Class 4 and Class 9 data are considered as test data, respectively. The decision limit C for $RE^2_{100,5}$ statistic is 1.794 (see Table 3). The test data are divided into subgroups of size $n = 5$ and $RE^2_{100,5}$ statistic is computed. The resulting $RE^2_{100,5}$ statistic values are plotted in Figure 8.

We also calculate the HDSIP-L values[29] for the same reference and test data, using the first data point in the reference data as the *referencing* point. For a three-dimensional dataset, with reference and test sample size 100 and 5, respectively, the decision limit is 10.99 for $AR\,L_0 = 370.43$. Figure 9 illustrates the HDSIP-L values, indicating that HDSIP-L values are unable to capture the fluctuations in the data. The reason for this is that the Euclidean distances of the test samples are much larger than the reference samples, and therefore, the Lepage statistic takes an extreme value, and this effect is the same for all subgroups.

Using this case study, we have demonstrated our methodology and its usefulness in sensorless drive technology. The importance of capturing data fluctuations in VSD applications is important for various reasons such as optimisation, safety, energy management, etc. As observed in Figure 8, faulty drives from Class 9 demonstrate relative stability when compared to those from Class 4, despite both being defective. However, the HIDSIP-L statistics, presented in Figure 9, do not highlight these differences, instead they suggest a similar level of defectiveness across both classes of faulty drives.

6 | CONCLUSION

We propose a new rank-energy (RE^2) statistic-based[43] multivariate monitoring scheme. The existing nonparametric multivariate monitoring schemes often lack robustness, invariance properties and are dependent on the quality of the training data. The proposed method is distribution-free, invariant to affine transformations, whilst demonstrating proficient shift detection capabilities.

The proposed method also has certain limitations. We have observed that with the increase in dimension, there is a corresponding rise in the ARL_1, indicating a decline in the detection power. Despite the ability to detect shifts in 'high dimension, low sample size' settings, the proposed method sometimes exhibits bias, especially for skewed multivariate processes. Numerical experiments indicate that a larger reference sample may be beneficial in monitoring the high-dimensional data, reducing bias and increasing detection power. Therefore, reducing bias from a rank-energy statistic-based monitoring scheme without increasing the reference sample size could be considered as an important topic for further research.

ACKNOWLEDGEMENTS

The authors express their gratitude to the anonymous reviewers whose valuable comments and suggestions have enhanced the quality of this paper. Additionally, the authors would like to thank the high-performance computation (HPC) facility at the University of the Free State for their assistance with the computational work involved in this study. The authors did not receive support from any organization for the submitted work.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in UC Irvine Machine Learning Repository at https://archive.ics.uci.edu/dataset/325/dataset+for+sensorless+drive+diagnosis.

ORCID

Niladri Chakraborty https://orcid.org/0000-0001-9853-4067
Maxim Finkelstein https://orcid.org/0000-0002-3018-8353

REFERENCES

1. Hotelling H. Multivariate quality control illustrated by air testing of sample bombsights. In: Eisenhart C, Hastay MW, Wallis WA, eds. *Techniques of Statistical Analysis*. McGraw Hill; 1947:111-184.
2. Qiu P. *Introduction to Statistical Process Control*. Chapman and Hall/CRC; 2014. doi:10.1201/b15016

3. Zhang C, Chen N, Wu J. Spatial rank-based high-dimensional monitoring through random projection. *J Qual Technol*. 2020;52:111-127. doi:10.1080/00224065.2019.1571336
4. Ahsan M, Mashuri M, Lee MH, et al. Robust adaptive multivariate Hotelling's T2 control chart based on kernel density estimation for intrusion detection system. *Expert Syst Appl*. 2020;145:113105. doi:10.1016/j.eswa.2019.113105
5. Chakraborty N, Lui CF, Maged A. A distribution-free change-point monitoring scheme in high-dimensional settings with application to industrial image surveillance. *Commun Stat Simul Comput*. 2023:1-17. doi:10.1080/03610918.2023.2202371
6. Wu T, Wang R, Yan H, et al. Adaptive change point monitoring for high-dimensional data. *Stat Sin*. 2022;32:1583-1610. doi:10.5705/ss.202020.0438
7. Ge Z, Song Z. Multivariate statistical process control. *Advances in Industrial Control*. Springer London; 2013. doi:10.1007/978-1-4471-4513-4
8. Champ CW, Jones-Farmer LA, Rigdon SE. Properties of the T 2 control chart when parameters are estimated. *Technometrics*. 2005;47:437-445. doi:10.1198/004017005000000229
9. Jensen WA, Jones-Farmer LA, Champ CW, et al. Effects of parameter estimation on control chart properties: a literature review. *J Qual Technol*. 2006;38:349-364. doi:10.1080/00224065.2006.11918623
10. Psarakis S, Vyniou AK, Castagliola P. Some recent developments on the effects of parameter estimation on control charts. *Qual Reliab Eng Int*. 2014;30:1113-1129. doi:10.1002/qre.1556
11. Ebadi M, Chenouri S, Lin DKJ, et al. Statistical monitoring of the covariance matrix in multivariate processes: a literature review. *J Qual Technol*. 2022;54:269-289. doi:10.1080/00224065.2021.1889419
12. Qiu P. Some perspectives on nonparametric statistical process control. *J Qual Technol*. 2018;50:49-65. doi:10.1080/00224065.2018.1404315
13. Maboudou-Tchao EM. High-dimensional data monitoring using support machines. *Commun Stat Simul Comput*. 2021;50:1927-1942. doi:10.1080/03610918.2019.1588312
14. Tran KP, ed. *Control Charts and Machine Learning for Anomaly Detection in Manufacturing*. Springer International Publishing; 2022. doi:10.1007/978-3-030-83819-5
15. He S, Jiang W, Deng H. A distance-based control chart for monitoring multivariate processes using support vector machines. *Ann Oper Res*. 2018;263:191-207. doi:10.1007/s10479-016-2186-4
16. Li Z, Tian L, Yan X. An ensemble framework based on multivariate statistical analysis for process monitoring. *Expert Syst Appl*. 2022;205:117732. doi:10.1016/j.eswa.2022.117732
17. Qiu P, Xie X. Transparent sequential learning for statistical process control of serially correlated data. *Technometrics*. 2022;64:487-501. doi:10.1080/00401706.2021.1929493
18. Lee WJ, Mendis GP, Triebe MJ, et al. Monitoring of a machining process using kernel principal component analysis and kernel density estimation. *J Intell Manuf*. 2020;31:1175-1189. doi:10.1007/s10845-019-01504-w
19. Xiao B, Li Y, Sun B, et al. Decentralized PCA modeling based on relevance and redundancy variable selection and its application to large-scale dynamic process monitoring. *Process Saf Environ Prot*. 2021;151:85-100. doi:10.1016/j.psep.2021.04.043
20. Liu RY. *Control Charts for Multivariate Processes*. Source: Journal of the American Statistical Association; 1995.
21. Qiu P, Hawkins D. A rank-based multivariate CUSUM procedure. *Technometrics*. 2001;43:120-132. doi:10.1198/004017001750386242
22. Qiu P, Hawkins D. A nonparametric multivariate cumulative sum procedure for detecting shifts in all directions. *J R Stat Soc*. 2003;52:151-164. doi:10.1111/1467-9884.00348
23. Qiu P. Distribution-free multivariate process control based on log-linear modeling. *IIE Trans*. 2008;40:664-677. doi:10.1080/07408170701744843
24. Marden J. Multivariate rank tests. In: Ghosh S, ed. *Multivariate Analysis, Design of Experiments, and Survey Sampling*. CRC Press; 1999:401-432. doi:10.1201/9781482289824
25. Randles RH. A simpler, affine-invariant, multivariate, distribution-free sign test. *J Am Stat Assoc*. 2000;95:1263-1268. doi:10.1080/01621459.2000.10474326
26. Hettmansperger TP, McKean JW. *Robust Nonparametric Statistical Methods*. CRC Press; 2010. doi:10.1201/b10451
27. Holland MD, Hawkins DM. A control chart based on a nonparametric multivariate change-point model. *J Qual Technol*. 2014;46:63-77. doi:10.1080/00224065.2014.11917954
28. Bae SJ, Do G, Kvam P. On data depth and the application of nonparametric multivariate statistical process control charts. *Appl Stoch Models Bus Ind*. 2016;32:660-676. doi:10.1002/asmb.2186
29. Mukherjee A, Marozzi M. Nonparametric Phase-II control charts for monitoring high-dimensional processes with unknown parameters. *J Qual Technol*. 2022;54:44-64. doi:10.1080/00224065.2020.1805378
30. Bush HM, Chongfuangprinya P, Chen VCP, et al. Nonparametric multivariate control charts based on a linkage ranking algorithm. *Qual Reliab Eng Int*. 2010:663-675. doi:10.1002/qre.1129
31. Hettmansperger TP, Randles RH. A practical affine equivariant multivariate median. *Biometrika*. 2002;89:851-860. doi:10.1093/biomet/89.4.851
32. Huwang L, Lin L, Yu C. A spatial rank–based multivariate EWMA chart for monitoring process shape matrices. *Qual Reliab Eng Int*. 2019;35:1716-1734. doi:10.1002/qre.2471
33. Li J, Zhang X, Jeske DR. Nonparametric multivariate CUSUM control charts for location and scale changes. *J Nonparametr Stat*. 2013;25:1-20. doi:10.1080/10485252.2012.726992
34. Li Z, Zou C, Wang Z, et al. A multivariate sign chart for monitoring process shape parameters. *J Qual Technol*. 2013;45:149-165. doi:10.1080/00224065.2013.11917923

35. Liang W, Xiang D, Pu X. A robust multivariate EWMA control chart for detecting sparse mean shifts. *J Qual Technol*. 2016;48:265-283. doi:10.1080/00224065.2016.11918166
36. Wang J, Su Q, Fang Y, et al. A multivariate sign chart for monitoring dependence among mixed-type data. *Comput Ind Eng*. 2018;126:625-636. doi:10.1016/j.cie.2018.09.053
37. Zi X, Zou C, Zhou Q, et al. A directional multivariate sign EWMA control chart. *Qual Technol Quant Manag*. 2013;10:115-132. doi:10.1080/16843703.2013.11673311
38. Zou C, Wang Z, Tsung F. A spatial rank-based multivariate EWMA control chart. *Nav Res Logist (NRL)*. 2012;59:91-110. doi:10.1002/nav.21475
39. Zou C, Tsung F. A multivariate sign EWMA control chart. *Technometrics*. 2011;53:84-97. doi:10.1198/TECH.2010.09095
40. Chen N, Zi X, Zou C. A distribution-free multivariate control chart. *Technometrics*. 2016;58:448-459. doi:10.1080/00401706.2015.1049750
41. Zhang C, Chen N, Zou C. Robust multivariate control chart based on goodness-of-fit test. *J Qual Technol*. 2016;48:139-161. doi:10.1080/00224065.2016.11918156
42. Hallin M, del Barrio E, Cuesta-Albertos J, et al. Distribution and quantile functions, ranks and signs in dimension d: A measure transportation approach. *The Annals of Statistics*. 2021;49(2):1139-1165. doi:10.1214/20-AOS1996
43. Deb N, Sen B. Multivariate rank-based distribution-free nonparametric testing using measure transportation. *J Am Stat Assoc*. 2023;118:192-207. doi:10.1080/01621459.2021.1923508
44. Chernozhukov V, Galichon A, Hallin M, et al. Monge–Kantorovich depth, quantiles, ranks and signs. *Ann Stat*. 2017;45(1):223-256. doi:10.1214/16-AOS1450
45. Hallin M. Measure transportation and statistical decision theory. *Annu Rev Stat Appl*. 2022;9:401-424. doi:10.1146/annurev-statistics-040220-105948
46. Hallin M, Mordant G. On the finite-sample performance of measure-transportation-based multivariate rank tests. In: *Robust and Multivariate Statistical Methods: Festschrift in Honor of David E. Tyler*. Springer International Publishing; 2022:87-119. doi:10.1007/978-3-031-22687-8_5
47. Jones-Farmer LA, Woodall WH, Steiner SH, et al. An overview of phase I analysis for process improvement and monitoring. *J Qual Technol*. 2014;46:265-280. doi:10.1080/00224065.2014.11917969
48. Qiu P. Big data? Statistical process control can help!. *Am Stat*. 2020;74:329-344. doi:10.1080/00031305.2019.1700163
49. Li W, Pu X, Tsung F, et al. A robust self-starting spatial rank multivariate EWMA chart based on forward variable selection. *Comput Ind Eng*. 2017;103:116-130. doi:10.1016/j.cie.2016.11.024
50. Xue L, Qiu P. A nonparametric CUSUM chart for monitoring multivariate serially correlated processes. *J Qual Technol*. 2021;53:396-409. doi:10.1080/00224065.2020.1778430
51. Train KE. *Discrete Choice Methods with Simulation*. Cambridge University Press; 2009. doi:10.1017/CBO9780511805271
52. Christophe D, Petr S, randtoolbox: Generating and Testing Random Numbers. R package version 2.0.4. 2023. doi:10.32614/CRAN.package.randtoolbox
53. Sofikitou E, Koutras M. In: Koutras MV, Triantafyllou IS, eds. *Distribution-Free Methods for Statistical Process Monitoring and Control*. Springer International Publishing; 2020. doi:10.1007/978-3-030-25081-2
54. Joe H. *Dependence Modeling With Copulas*. Chapman and Hall/CRC; 2014. doi:10.1201/b17116
55. Nelsen RB. *An Introduction to Copulas*. Springer New York; 2006. doi:10.1007/0-387-28678-0
56. Finkelstein M. *Failure Rate Modelling for Reliability and Risk*. Springer London; 2008. doi:10.1007/978-1-84800-986-8
57. Chen Z, Fan J, Wang K. Multivariate Gaussian processes: definitions, examples and applications. *METRON*. 2023;81:181-191. doi:10.1007/s40300-023-00238-3
58. Maier M, Rupenyan A, Bobst C, et al. Self-optimizing grinding machines using Gaussian process models and constrained Bayesian optimization. *Int J Adv Manuf Technol*. 2020;108:539-552. doi:10.1007/s00170-020-05369-9
59. Chakraborty N, Mahmood T. Failure rate monitoring in generalized gamma-distributed process. *Qual Technol Quant Manag*. 2021;18:718-739. doi:10.1080/16843703.2021.1953241
60. Finkelstein M, Cha JH, Chackraborty N. Balancing load and performance for different failure models. *Appl Stoch Models Bus Ind*. 2022;38:323-333. doi:10.1002/asmb.2662
61. Finkelstein M, Cha JH. Stochastic modeling for reliability. *Springer Series in Reliability Engineering*. Springer London; 2013. doi:10.1007/978-1-4471-5028-2
62. Wang W, Niu S, Zhao X. A novel field and armature synchronous pulse injection method for sensorless drive control of 12/10 DC vernier reluctance machine. *IEEE Trans Energy Convers*. 2023;38(3):2126-2135. doi:10.1109/TEC.2023.3268328
63. Yuan L, Ding Y, Jiang Y, et al. Multi-dimensional vector control for six-phase interior permanent magnet synchronous motor sensorless drive system. *Energ Rep*. 2023;9:1126-1133. doi:10.1016/j.egyr.2023.05.082
64. Joe H. *Dependence Modeling With Copulas*. Chapman and Hall/CRC; 2014. doi:10.1201/b17116

How to cite this article: Chakraborty N, Finkelstein M. Distribution-free multivariate process monitoring: A rank-energy statistic-based approach. *Qual Reliab Engng Int*. 2024;40:4068–4087. https://doi.org/10.1002/qre.3619

APPENDIX

A1

R-Algorithm for Monte Carlo simulation to estimate the decision limit for a given ARL_0:

Input: Design parameters m, n, d; Number of iterations B_1 and B_2; A given value for ARL_0 so that $\alpha_0 \approx 1/ARL_0$.

Output: The decision limit.

1. Call the necessary packages.
2. Call the computestatistic() function.[43]
3. Define the mean vector $\boldsymbol{\mu}_{\boldsymbol{d}\times 1}$ and the covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{d}\times\boldsymbol{d}}$ as in Section 4.1.
4. # Define a function to calculate the $RE^2_{m,n}$ in an iterative way.
   Test_stat = function(count, m, n, d, reference sample){
   Draw test sample $\boldsymbol{Y}_{\boldsymbol{n}\times\boldsymbol{d}}$ of size n from a multivariate normal with mean $\boldsymbol{\mu}_{\boldsymbol{d}\times 1}$ and covariance $\boldsymbol{\Sigma}_{\boldsymbol{d}\times\boldsymbol{d}}$.
   Calculate the $RE^2_{m,n}$ statistic using the computestatistic() function using the reference and the test sample.
   Return the $RE^2_{m,n}$ statistic.
   }
5. **Fix** C
   for $\boldsymbol{i}$ **in 1**:B_1 **{**
6. Draw a reference sample $\boldsymbol{X}_{\boldsymbol{d}\times\boldsymbol{m}}$ from multivariate normal distribution with mean $\boldsymbol{\mu}_{\boldsymbol{d}\times 1}$ and covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{d}\times\boldsymbol{d}}$.
7. Calculate B_2 number of conditional $RE^2_{m,n}$ values using the Test_stat() function.
8. Calculate the 100 $(1-\alpha)$ percentile of the conditional $RE^2_{m,n}$ to estimate C.
9. Estimate the conditional $ARL_0 \approx \frac{1}{P_X[RE^2_{m,n}\geq C]}$
   }
10. Estimate the ARL_0 by taking average of B_1 number of conditional ARL_0 values. The decision limit is obtained iteratively so that $ARL_0 \approx 370$.

A2

To estimate the ARL_1, we change the mean vector and the covariance matrix or change the distribution. The rest of the steps in the algorithm are same as the Algorithm in A1 except that it is not necessary to estimate the decision limit.

A3

We provide a brief discussion on the Gaussian and Clayton copula. A d-dimensional copula C is a cumulative distribution function (c.d.f.) given by $C : [0,1]^d \to [0,1]$ with uniform marginals. According to the Sklar's Theorem,[64] for any d-dimensional c.d.f. with continuous marginals, there exists a copula that can uniquely describe the dependency structure. The Gaussian copula is defined as:

$$C^{Gauss}_{\boldsymbol{\Sigma}}(u_1,\dots,u_d) = \boldsymbol{\Phi}_{\boldsymbol{\Sigma}}\left(\Phi^{-1}(u_1),\dots,\Phi^{-1}(u_d)\right),(u_1,\dots,u_d)\in[0,1]^d, \tag{A.1}$$

where $\Phi_{\boldsymbol{\Sigma}}$ is the d-variate normal c.d.f. with correlation matrix $\boldsymbol{\Sigma}$, and $\Phi(.)$ is a univariate standard normal c.d.f.

The Archimedean copula is defined as:

$$C^{Arch}(u_1,\dots,u_d) = \psi\left(\psi^{-1}(u_1)+\psi^{-1}(u_2)+\cdots+\psi^{-1}(u_d)\right),(u_1,\dots,u_d)\in[0,1]^d, \tag{A.2}$$

where $\psi(.)$ is a generator function of C^{Arch}. The Clayton copula is a special case of the Archimedean copula with the generator

$$\psi(u_i) = \frac{u_i^{-\xi}-1}{\xi},\ \xi>0,\ i=1,2,3,\dots,d. \tag{A.3}$$

AUTHOR BIOGRAPHIES

Dr. Niladri Chakraborty is a lecturer in the Department of Mathematical Statistics and Actuarial Sciences, University of the Free State, South Africa. He earned his PhD in Mathematical Statistics from the University of Pretoria in 2017. Prior to joining the University of the Free State, he worked as a postdoctoral researcher at the City University of Hong Kong. His main research interests are in statistical process control and nonparametric inference.

Prof. Maxim Finkelstein holds the title of Distinguished Professor in Statistics/Mathematical Statistics at the Department of Mathematical Statistics and Actuarial Sciences, University of the Free State, South Africa. He also serves as a visiting researcher at Max Planck Institute for Demographic Research, Rostock, Germany, and has been a visiting research professor at the ITMO University, St Petersburg, Russia since 2014. Prof. Finkelstein earned his PhD in OR (mathematical theory of reliability) from Leningrad Elektropribor Institute in 1979, followed by a Doctor of Science (habilitation) from St Petersburg Elektropribor Institute in 1993. His research primarily explores stochastic shock processes, hazard rates in heterogeneous populations, models of imperfect repair, survival in dynamic environments, and wear and degradation processes. He has also served as a member of the editorial board for several national and international reputed journals.