

SVDD control charts based on MEWMA technique for monitoring Compositional Data

Thi Thuy Van Nguyen^{*1,2}, Cédric Heuchenne¹, Kim Duc Tran³, Guillaume Tartare² and Kim Phuc Tran²

¹HEC Liège - Management School of the University of Liège, Belgium

²University of Lille, ENSAIT, GEMTEX, Lille, France

³IAD, Dong A University, Da Nang, Vietnam

^{*} Corresponding author. Email: ttv.nguyen@uliege.be

March 17, 2024

Abstract

The monitoring of compositional data (CoDa) using control charts has become increasingly important in the field of Statistical Process Control (SPC). This study introduces two approaches for CoDa monitoring, utilizing support vector data description (SVDD) control charts in conjunction with the multivariate exponentially weighted moving average (MEWMA) technique. The proposed approaches use two transformation methods, the Dirichlet density transformation, and the isometric log-ratio transformation. Through experimentation on simulated data, the effectiveness of the SVDD control charts in detecting anomalies is demonstrated surpassing the performance of the MEWMA control chart across various scenarios in monitoring CoDa. These findings contribute to the advancement of SPC and offer valuable insights for practitioners involved in CoDa monitoring across diverse applications.

Keywords: Compositional data, Control chart, Dirichlet density transformation, isometric log-ratio, Support Vector Data Description, Statistical Process Control.

1 Introduction

In Statistical Process Control (SPC) literature, control charts are a well-established tool for monitoring the stability and variability of process data. However, the application of control charts to compositional data (CoDa) has been relatively limited in the literature. This may be due to the special structure of CoDa, which represents the relative proportions of components that make up a whole, such as the composition of minerals in a rock or the ingredients in a product. The components of CoDa vectors are strictly positive and often represent proportions, percentages, concentrations, or frequencies of some whole. Their applications can be found in many domains such as chemical research, econometric and survey data analyses, and the food industry, see Aitchison [2]. The monitoring of CoDa with control charts presents challenges due to its special structure, which the traditional control charts, designed for continuous data,

cannot adequately address. To overcome this, the literature has seen the development of specialized control charts for monitoring CoDa. An early contribution to this field was presented in [4]. In this study, the author proposed a chi-square control chart for monitoring CoDa based on Dirichlet distribution. Guevara-González et al. [9] conducted a study examining the relationship between response and predictor variables using a Dirichlet regression and evaluated the performance of the method by plotting five different T^2 control charts. Building on this, Vives-Mestres et al. [34] investigated a T^2 control chart for monitoring CoDa with $p = 3$ and then Vives-Mestres et al. [35] extended this work for individual observations case. Two methods for interpretations of an out-of-control signal of individual T^2 control chart in case $p > 3$ were proposed in [36]. In [32], the authors introduced the MEWMA-CoDa control chart as a method for monitoring CoDa with arbitrary components. This control chart has been recognized for its effectiveness in identifying small to moderate shifts in processes, surpassing the performance of its competitor (T^2 -CoDa chart). Later, in the studies of Nguyen et al. [23] and Imran et al. [13], the authors enhanced the MEWMA control chart by incorporating variable sampling interval (VSI) and variable sample size (VSS) adaptive strategies for CoDa monitoring, demonstrating their outperformance over the traditional fixed sampling interval/sample size approaches. Imran et al. [16] further explored the zero-state and steady-state performance of the MEWMA-CoDa chart with VSI, revealing the influence of the number of variables and the subgroup size on the chart's performance. The literature has also seen the introduction of CUSUM-CoDa charts, with the study proposed by Imran et al. [14]. The results showed that this chart performed better than the T^2 -CoDa control chart, and was comparable to the MEWMA-CoDa chart. Besides the design of these control charts, the influence of measurement errors on the performances of T^2 -CoDa, MEWMA-CoDa, MCUSIM-CoDa, and VSI T^2 -CoDa were also investigated in [37], [38], [12] and [17], respectively. Addressing the challenge of autocorrelation in industrial processes, Zaidi et al. [40] proposed an enhanced residuals T^2 control chart for monitoring the autocorrelated CoDa using a time series autoregressive moving average (ARMA) model. To manage large and high-dimensional CoDa datasets, Imran et al. [15] developed a T^2 -CoDa control chart based on the principal component analysis (PCA) method to monitor CoDa explicitly. Recent advances have incorporated machine learning tools, with Zaidi et al. [39] and Imran et al. [18] utilizing artificial neural networks to detect abnormal patterns and analyze out-of-control signals in monitoring CoDa. Most of the control charts in these studies utilized the isometric log-ratio (ilr) transformation method to handle the special structure of CoDa. This method converts CoDa into a vector in $\mathbb{R}^{p-1}$ space, thereby removing constraints before applying traditional control charts. It should be noted that the design of these control charts assumes a restrictive assumption of multivariate normal distribution on the Simplex space, which may not always be applicable in real-world scenarios.

In the field of SPC, there have been various attempts to overcome the limitations imposed by the assumptions on data distribution. One approach that has received attention is the use of distribution-free methods, specifically the use of control charts based on one-classification algorithms. In [29], the authors proposed kernel distance-based charts, namely K-chart, using the Support Vector Data Description (SVDD) algorithm, which was originally proposed by [31] based on the idea of Support Vector Machines (SVM). Kernel methods are powerful tools in computational intelligence that have many advantages since they can handle non-linear relationships by implicitly mapping data to a higher dimensional space, leading to more accurate models without explicit transformation. This “kernel trick” can sometimes reduce computational costs associated with high-dimensional data. Furthermore, kernel methods can implicitly learn data structure by defining a feature space that best separates data, particularly in clas-

sification tasks. Based on the kernel methods and another type of one-classification algorithm, Kumar et al. [19] proposed a robust K-chart that can handle the autocorrelation in the observed data. In [20], the authors proposed an SVDD control chart using Mahalanobis kernels and showed its outperformance compared with the SVDD control chart using the Gaussian kernel and T^2 control chart. Later, in the study conducted by Tang et al. [30], the authors examined the impact of the Phase I dataset size on the efficiency of Phase II monitoring and found that the effectiveness of the control chart based on SVDD is significantly influenced by the quantity of Phase I sample. Recently, Ahsan et al. [1] proposed an innovative multivariate control chart by integrating SVDD and kernel density estimation (KDE) to address the challenges of network intrusion detection. The SVDD-KDE chart demonstrated enhanced performance in identifying shifts and outliers within network traffic, showing superior accuracy and lower false positive rates in both simulations and real-world datasets.

Our paper introduces two novel approaches for monitoring CoDa using SVDD control charts combined with the MEWMA technique. We propose employing two transformation approaches, namely the Dirichlet density transformation and the isometric log-ratio transformation to adapt CoDa for real-world monitoring scenarios. The rest of the paper is structured as follows: Section 2 presents the modeling of CoDa and the isometric log-ratio transformation. Section 3 focuses on the Dirichlet density transformation to convert CoDa to real data. Section 4 presents our innovative SVDD-MEWMA control charts for monitoring CoDa using the aforementioned transformation methods. The performances of the SVDD-MEWMA charts and comparison with the existing MEWMA control charts for CoDa monitoring under various scenarios are presented in Section 5. An illustration example in Section 6 presents the application of these proposed control charts in a real European plant's industrial CoDa set. The conclusion in Section 7 reflects on the study's contributions and suggests some directions for future research.

2 Modeling of Compositional Data

In the literature, CoDa refers to data where components represent relative proportions of a whole, such that its components sum to a constant value. This describes distribution where the focus is on the relative magnitudes of the parts, rather than their absolute values. In mathematical terms, a (row) vector, denoted as $\mathbf{x} = (x_1, x_2, \dots, x_p)$, is considered a p-part composition when its components are strictly positive numbers and convey relative information, see Aitchison [2], Pawlowsky-Glahn et al. [27]. The relative information here refers only to the proportional relationship between its components, disregarding their specific numerical values. The sum of these components of $\mathbf{x}$, $\sum_{i=1}^p x_i$, is a constant κ . For example, when κ equals 1, the measurements are expressed as ratios while $\kappa = 100$ corresponds to measurements in percentages. If $\mathbf{x}, \mathbf{y}$ are compositions and $\mathbf{x} = \lambda \mathbf{y}$ with λ is a constant, we say that $\mathbf{x}, \mathbf{y}$ are compositionally equivalent. Equivalently, if the closure of a p-part composition $\mathbf{x} = (x_1, x_2, \dots, x_p)$ to $\kappa > 0$ is defined by

$$\mathcal{C}(\mathbf{x}) = \left(\frac{\kappa \cdot x_1}{\sum_{i=1}^p x_i}, \frac{\kappa \cdot x_2}{\sum_{i=1}^p x_i}, \dots, \frac{\kappa \cdot x_p}{\sum_{i=1}^p x_i} \right) \quad (1)$$

then the two p-part compositions $\mathbf{x}$ and $\mathbf{y}$ are compositionally equivalent if $\mathcal{C}(\mathbf{x}) = \mathcal{C}(\mathbf{y})$ for

every positive constant κ . The sample space of CoDa is the simplex $\mathcal{S}^p$,

$$\mathcal{S}^p = \left\{ \mathbf{x} = (x_1, x_2, \dots, x_p) \mid x_i > 0, i = 1, \dots, p; \sum_{i=1}^p x_i = \kappa \right\} \quad (2)$$

In real space, the addition and the scalar multiplication of vectors using Euclidean geometry can be utilized to determine their properties or compute the distance between them. However, this geometry does not apply to CoDa in the $\mathcal{S}^p$ due to its special structure. To address this issue, Aitchison [2] introduced Aitchison geometry, which includes two fundamental operations for establishing a vector space structure in the simplex $\mathcal{S}^p$: perturbation and powering operators. Let $\mathbf{x}$ and $\mathbf{y}$ be two compositions on $\mathcal{S}^p$, the perturbation of $\mathbf{x}$ by $\mathbf{y}$ (denoted by $\oplus$) is defined by

$$\mathbf{x} \oplus \mathbf{y} = \mathcal{C}(x_1 y_1, \dots, x_p y_p) \in \mathcal{S}^p$$

and the powering of $\mathbf{x}$ by a constant $\alpha \in \mathbb{R}$ (denoted by $\odot$) is defined by

$$\alpha \odot \mathbf{x} = \mathcal{C}(x_1^\alpha, \dots, x_p^\alpha) \in \mathcal{S}^p$$

The simplex $\mathcal{S}^p$ equipped with the perturbation and the power operations, denoted as $(\mathcal{S}^p, \oplus, \odot)$, constitutes a vector space.

In practice, CoDa is often transformed into vector representations in Euclidean space, rather than being directly applied and deployed in Simplex space. This transformation process serves the purpose of eliminating the constraints of CoDa. One widely used transformation technique is the center log-ratio (clr) transformation, denoted as $\text{clr}(\mathbf{x})$. This method is an isometry that maps from $\mathcal{S}^p$ to a subspace $U \subset \mathbb{R}^p$, and is defined as

$$\begin{aligned} \text{clr}(\mathbf{x}) &= \left(\ln \frac{x_1}{g_m(\mathbf{x})}, \ln \frac{x_2}{g_m(\mathbf{x})}, \dots, \ln \frac{x_p}{g_m(\mathbf{x})} \right) \\ &= (\xi_1, \xi_2, \dots, \xi_p) = \boldsymbol{\xi} \end{aligned} \quad (3)$$

where $g_m(\mathbf{x}) = \left(\prod_{i=1}^p x_i \right)^{\frac{1}{p}} = \exp \left(\frac{1}{p} \sum_{i=1}^p \ln x_i \right)$ is the geometric mean of the composition and $\sum_{i=1}^p \xi_i = 0$. The inverse center log-ratio $\text{clr}^{-1}(\boldsymbol{\xi})$ recovering $\mathbf{x}$ from $\text{clr}(\mathbf{x})$ will be

$$\text{clr}^{-1}(\boldsymbol{\xi}) = \mathcal{C}(\exp(\boldsymbol{\xi})) = \mathcal{C}(\exp(\xi_1), \exp(\xi_2), \dots, \exp(\xi_p)).$$

Egozcue et al. [5] demonstrated that the constraints on the components of $\text{clr}(\mathbf{x})$ result in a singular variance-covariance matrix for random compositions. To tackle this problem, they introduced the isometric log-ratio (ilr) transformation, a new method that utilizes an orthogonal basis in the simplex $\mathcal{S}^p$. This ilr transformation is an isometry from $\mathcal{S}^p$ to $\mathbb{R}^{p-1}$, which helps to overcome the limitations of clr transformation by directly mapping to an orthogonal coordinate system in the Simplex space. Assuming $\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_{p-1}$ form an orthonormal basis of the simplex $\mathcal{S}^p$, any composition $\mathbf{x} \in \mathcal{S}^p$ can be represented as

$$\mathbf{x} = \bigoplus_{i=1}^{p-1} x_i^* \odot \mathbf{e}_i, \quad x_i^* = \langle \mathbf{x}, \mathbf{e}_i \rangle_a = \langle \text{clr}(\mathbf{x}), \text{clr}(\mathbf{e}_i) \rangle \quad (4)$$

where $\langle, \rangle_a$ denotes the Aitchison inner product. The ilr transformation of $\mathbf{x} \in \mathcal{S}^p$ is defined by

$$\text{ilr}(\mathbf{x}) = \mathbf{x}^* = (x_1^*, x_2^*, \dots, x_{p-1}^*). \quad (5)$$

Let $\mathbf{B}$ be a $(p-1, p)$ matrix with i^{th} row equal to $\text{clr}(\mathbf{e}_i)$, $i = 1, \dots, p-1$. This matrix is known as a contrast matrix associated with the orthonormal basis $\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_{p-1}$. The ilr transformation $\mathbf{x}^*$ of composition $\mathbf{x}$ can be computed in form of contrast matrix $\mathbf{B}$ and $\text{clr}(\mathbf{x})$ as follows

$$\mathbf{x}^* = \text{ilr}(\mathbf{x}) = (x_1^*, \dots, x_{p-1}^*) = \text{clr}(\mathbf{x}) \cdot \mathbf{B}^\top \quad (6)$$

In many potential candidates for an orthonormal basis in $\mathcal{S}^p$, Egozcue and Pawlowsky-Glahn [6] proposed a sequential binary partition for constructing an orthonormal basis. In this basis, $\mathbf{e}_i$ is defined to be $\mathcal{C}(e_{i,1}, \dots, e_{i,j}, \dots, e_{i,p})$ where

$$e_{i,j} = \begin{cases} \exp\left(\sqrt{\frac{1}{i(i+1)}}\right) & \text{if } j \leq i \\ \exp\left(-\sqrt{\frac{i}{i+1}}\right) & \text{if } j = i+1 \\ 1 & \text{otherwise} \end{cases} \quad (7)$$

Elements $e_{i,j}$ are called the *balancing elements* of this basis. Thus, if this type of orthonormal basis is chosen to transform $\mathbf{x}$, i.e., $\text{ilr}(\mathbf{x}) = \text{clr}(\mathbf{x}) \cdot \mathbf{B}^\top$, the coordinates x_i^* are called *balances* and can be obtained by

$$x_i^* = \sqrt{\frac{i}{i+1}} \ln \left(\frac{\left(\prod_{j=1}^i x_j\right)^{\frac{1}{i}}}{x_{i+1}} \right). \quad (8)$$

From its ilr coordinate $\mathbf{x}^*$, $\mathbf{x}$ can be recovered by using the inverse of ilr transformation:

$$\text{ilr}^{-1}(\mathbf{x}^*) = \text{clr}^{-1}(\mathbf{x}^* \mathbf{B}) = \mathcal{C}(\exp(\mathbf{x}^* \mathbf{B})).$$

For more detail, see Pawlowsky-Glahn et al. [27], chapter 4.

3 Dirichlet density transformation method to transform CoDa

According to Fan et al. [7], the marginal density ratios are highly effective in solving univariate binary classification problems. This concept can be extended to multivariate classification problems by applying the independent assumption between the feature variables given class labels. Specifically, each feature is independently transformed to its corresponding marginal density ratio and then concatenated together for training the classifier model. The decision boundary can be described as follows

$$\mathcal{D} = \left\{ \mathbf{x} : \beta_0 + \sum_{i=1}^p \beta_i \log \frac{f_i(x_i)}{g_i(x_i)} = 0, \quad \beta_0, \beta_1, \dots, \beta_p \in \mathbb{R} \right\}$$

where $\mathbf{x} = (x_1, \dots, x_p)^\top$ is p -dimensional feature vector, f_i, g_i are class conditional densities for class 1 and -1 of i^{th} feature, respectively, $\beta_i, i = 1, \dots, p$ are coefficients used to capture

the correlation between the features, and β_0 is the model's intercept. Building upon this, we extend this idea to multivariate one-classification problems. We employ the marginal density of each feature instead of using the original data and combine these marginal densities to train the anomaly detection problem. To handle the constraints and relative information of CoDa vectors, we implement the Dirichlet density estimation method based on the idea of the Dirichlet feature embedding method proposed by Gu et al. [8] on CoDa to remove the constant constraint before applying the SVDD algorithm for anomaly detection. The Dirichlet distribution is well-defined in Simplex space and has been extensively studied in many works such as [11], [10], [22] and [25]. Recall that, the Dirichlet distribution of order p with the parameter $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_p)$ has the probability density function f given by:

$$f(\mathbf{x}) = \frac{\Gamma(\alpha_+)}{\prod_{i=1}^p \Gamma(\alpha_i)} \prod_{i=1}^p x_i^{\alpha_i-1}, \quad (9)$$

where $\mathbf{x} = (x_1, x_2, \dots, x_p) \in S^p$, $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_p)^T \in \mathbb{R}_+^p$ is the parameter of Dirichlet distribution, $\alpha_+ = \sum_{i=1}^p \alpha_i$ and Γ denotes the gamma function. Instead of directly using the original CoDa, we will treat each CoDa feature as a whole and estimate its Dirichlet density within the Simplex space. These estimated densities will then be concatenated and utilized for training the model.

Let us consider a dataset $\mathbf{X}$ with n normal observations, denoted as $\mathbf{X}_i, i = 1, \dots, n$. Each observation is characterized by l CoDa vectors. Without loss of generality, we assume that the constant constraint on each CoDa vector $\mathbf{x} = (x_1, x_2, \dots, x_p) \in S^p$ is equal to 1, i.e., $\sum_{i=1}^p x_i = 1$. Additionally, we suppose that each of these l CoDa vectors may have a different number of parts, denoted as $p_1, p_2, \dots, p_l$. Consequently, each $\mathbf{X}_i$ can be represented as follows

$$\mathbf{X}_i = [x_{i1}^{(1)}, x_{i2}^{(1)}, \dots, x_{ip_1}^{(1)}, \dots, x_{i1}^{(l)}, x_{i2}^{(l)}, \dots, x_{ip_l}^{(l)}] \in S_l^{p_1, \dots, p_l}$$

Here $S_l^{p_1, \dots, p_l}$ denote the the multidimensional $(p_1, \dots, p_l)$ - Simplex space with dimension l , i.e., $S_l^{p_1, \dots, p_l} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_l : \mathbf{x}_i \in S^{p_i}\}$. The whole dataset can be presented as follows

$$\mathbf{X} = \begin{bmatrix} x_{11}^{(1)}, x_{12}^{(1)}, \dots, x_{1p_1}^{(1)} & \dots & x_{11}^{(l)}, x_{12}^{(l)}, \dots, x_{1p_l}^{(l)} \\ \vdots & \ddots & \vdots \\ x_{N1}^{(1)}, x_{N2}^{(1)}, \dots, x_{Np_1}^{(1)} & \dots & x_{N1}^{(l)}, x_{N2}^{(l)}, \dots, x_{Np_l}^{(l)} \end{bmatrix} \quad (10)$$

The summary procedure for transforming CoDa data using the Dirichlet density estimation method can be outlined as follows:

Algorithm 1: Dirichlet density transformation method

Input: An in-control multivariate CoDa set of size N with l compositional features:

$$\mathbf{X} = \{\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_N; \mathbf{X}_i \in S_l^{p_1, \dots, p_l}\}$$

Output: Dirichlet density transformed dataset $\hat{\mathbf{X}} = \{\hat{\mathbf{X}}_1, \hat{\mathbf{X}}_2, \dots, \hat{\mathbf{X}}_N; \hat{\mathbf{X}}_i \in \mathbb{R}^l\}$

1: **Estimating Dirichlet density:** Using Naïve Bayes Baye independence assumption and employ the MLE method to estimate Dirichlet density for l compositional features in $\mathbf{X}$. The estimated Dirichlet density $\hat{f}(\mathbf{x}_k^{(j)})$ of k^{th} individual and j^{th} compositional feature will be

$$\hat{f}(\mathbf{x}_k^{(j)}) = \frac{\Gamma(\hat{\alpha}_+^{(j)})}{\prod_{i=1}^{p_j} \Gamma(\hat{\alpha}_i^{(j)})} \prod_{i=1}^{p_j} (x_{ki}^{(j)})^{\hat{\alpha}_i^{(j)}-1}, \quad (11)$$

where $\hat{\alpha}_i^{(j)}$, $i = 1, \dots, p_j$, are the MLE of parameter $\alpha_i^{(j)}$ of the Dirichlet distribution for the j^{th} compositional feature, $\hat{\alpha}_+^{(j)} = \sum_{i=1}^{p_j} \hat{\alpha}_i^{(j)}$.

2: **Transforming data:** Using parameters estimated in Step 1 to transform data $\mathbf{X}$ into its estimated probability density (or its logarithmic version). The resulting transformed data $\hat{\mathbf{X}}$ is no longer constrained by the unit sum and consists of N data points with a dimension of l

$$\hat{\mathbf{X}} = \begin{bmatrix} \hat{f}(\mathbf{x}_1^{(1)}) & \hat{f}(\mathbf{x}_1^{(2)}) & \dots & \hat{f}(\mathbf{x}_1^{(l)}) \\ \vdots & \ddots & \vdots & \\ \hat{f}(\mathbf{x}_N^{(1)}) & \hat{f}(\mathbf{x}_N^{(2)}) & \dots & \hat{f}(\mathbf{x}_N^{(l)}) \end{bmatrix} \quad (12)$$

4 SVDD control charts for monitoring CoDa

In ML, SVDD is a well-known algorithm for detecting anomalies (outliers). SVDD was first introduced by Tax and Duin [31] and widely used in anomaly detection due to its free assumption about the distribution of the observed data (see, for example, [3], [26], and [41], among many others). Although there have been many researches related to SVDD and its applications, most of them were developed for use in normal data. Due to the constraints of CoDa, these researches can not be applied directly without handling these constraints. In this section, we will briefly present the main idea of the SVDD algorithm and then propose SVDD control charts for monitoring CoDa using two types of transformation: the ilr transformation in section 2 and the Dirichlet density transformation in section 3. The control chart presented here utilizes the **weighted average** of all available observed-transformed data at the current time point, employing a memory-based approach, and will be used to detect distributional shifts.

4.1 SVDD using MEWMA-based vectors

Recall that, SVDD is a one-class classification algorithm that allows us to detect abnormal observations by modeling the normal ones. This algorithm constructs a spherical-shaped boundary around the dataset, specified by its center $\mathbf{a}$ and radius R . The primary objective is to minimize the radius R , thereby minimizing the volume of the sphere while ensuring that it contains the majority of training objects. SVDD is widely utilized in practical applications for identifying novel data points or outliers.

Consider an unlabeled training dataset collected over N sampling periods, where at each period $i = 1, \dots, N$, a sample of m observations of p features is gathered. The dataset is denoted as

$S = \{\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_N\}$, where each $\mathbf{X}_i$ is an $m \times p$ matrix. To construct our model, we first compute the sample mean $\bar{\mathbf{x}}_i$ of each $\mathbf{X}_i$. This process yields a series of mean feature vectors, represented by $\bar{S} = \{\bar{\mathbf{x}}_1, \bar{\mathbf{x}}_2, \dots, \bar{\mathbf{x}}_N\}$. We then generate the MEWMA vectors $\mathbf{w}_i$ using the formula:

$$\mathbf{w}_i = r(\bar{\mathbf{x}}_i - \boldsymbol{\mu}_0) + (1 - r)\mathbf{w}_{i-1}, \quad i = 1, 2, \dots, N \quad (13)$$

where $\boldsymbol{\mu}_0$ is the mean vector of the whole dataset, $\mathbf{w}_0 = \mathbf{0}$, $r \in (0, 1]$ is a fixed smoothing parameter. These MEWMA vectors will then be used to train the SVDD model.

The SVDD optimization problem to obtain a description of the dataset using MEWMA-based vectors can be described by

$$\min_{R, \mathbf{a}} \quad R^2 + C \sum_{i=1}^N \xi_i \quad (14)$$

$$\text{subject to} \quad \|\mathbf{w}_i - \mathbf{a}\|^2 \leq R^2 + \xi_i, \quad i = 1, \dots, N; \quad \xi_i \geq 0$$

where C is the trade-off parameter which is introduced to control the balance between the volume of the sphere and the errors. Introducing two Lagrange multipliers α_i, γ_i , the dual problem of problem (14) will be

$$\max_{\alpha} \quad \sum_{i=1}^N \alpha_i \langle \mathbf{w}_i, \mathbf{w}_i \rangle - \sum_{i,j=1}^N \alpha_i \alpha_j \langle \mathbf{w}_i, \mathbf{w}_j \rangle \quad (15)$$

$$\begin{aligned} \text{subject to} \quad & \sum_{i=1}^N \alpha_i = 1 \\ & 0 \leq \alpha_i \leq C \quad \text{for} \quad i = 1, \dots, N \end{aligned}$$

The dual problem, being convex quadratic, is significantly more tractable than the original problem (14). Solving this problem and noting that the parameters γ_i are omitted during the solving process, we obtain the center $\mathbf{a}$ and the radius R of the sphere as follows

$$\mathbf{a} = \sum_{SV} \alpha_s \mathbf{w}_s \quad (16)$$

$$R^2 = \langle \mathbf{w}_k, \mathbf{w}_k \rangle - 2 \sum_{i=1}^N \alpha_i \langle \mathbf{w}_k, \mathbf{w}_i \rangle + \sum_{i,j=1}^N \alpha_i \alpha_j \langle \mathbf{w}_i, \mathbf{w}_j \rangle \quad (17)$$

where SV denotes the index set of the support vectors and $\mathbf{w}_k$ in formula (17) is any object that satisfies $\mathbf{w}_k \in SV_{<C}$, the set of support vectors having $\alpha_k < C$, see [31] for more details.

In general, to monitor a new sample $\mathbf{Z}_1, \mathbf{Z}_2, \dots$, where each sample $\mathbf{Z}_t$ consists of m observations collected at time step t , we first compute their sample mean vectors $\bar{\mathbf{z}}_1, \bar{\mathbf{z}}_2, \dots$, and then transform them to the MEWMA vectors $\mathbf{w}'_1, \mathbf{w}'_2, \dots$, using formula (13). Following this transformation, we calculate the squared distance between these mean vectors and the center $\mathbf{a}$ of the sphere, which can be computed as follows:

$$\|\mathbf{w}'_t - \mathbf{a}\|^2 = \langle \mathbf{w}'_t, \mathbf{w}'_t \rangle - 2 \sum_{i=1}^N \alpha_i \langle \mathbf{w}'_t, \mathbf{w}_i \rangle + \sum_{i,j=1}^N \alpha_i \alpha_j \langle \mathbf{w}_i, \mathbf{w}_j \rangle \quad (18)$$

In practice, instead of using the inner product in the above formulas, we often use the kernel functions to obtain a more flexible data description. Vapnik [33], Smola et al. [28] proposed some kernel functions for the Support Vector Classifier. One of the most popular kernel functions can be mentioned is the Gaussian kernel function $K(\mathbf{w}_i, \mathbf{w}_j)$, is defined by

$$K_G(\mathbf{w}_i, \mathbf{w}_j) = \exp(-\|\mathbf{w}_i - \mathbf{w}_j\|^2 / \sigma^2) \quad (19)$$

where σ is used to control the complexity of the SVDD boundary.

If the kernel function $K(\mathbf{x}, \mathbf{y})$ is used, the squared distance in Equation 18 can be recomputed by replacing the inner product with the kernel function. Denote this kernel distance by $D_{\mathbf{a}}(\mathbf{w}')$, then

$$D_{\mathbf{a}}(\mathbf{w}'_t) = K(\mathbf{w}'_t, \mathbf{w}'_t) - 2 \sum_{i=1}^N \alpha_i K(\mathbf{w}'_t, \mathbf{w}_i) + \sum_{i,j=1}^N \alpha_i \alpha_j K(\mathbf{w}_i, \mathbf{w}_j) \quad (20)$$

4.2 Support Vector Data Description control chart

To construct the SVDD control chart, Sun and Tsung [29], Maboudou-Tchao et al. [20], Tang et al. [30], among other authors, proposed to use the kernel distance defined in Equation (20) as the chart statistic. The radius R from Equation (17) is a candidate for the upper control limit threshold, however, in the framework of SPC, the upper limit h based on pre-specified in-control ARL (denoted by ARL_0) is more appropriate. In this case, the control chart would give an out-of-control signal if the kernel distance $D_{\mathbf{a}}(\mathbf{w}_t)$ between the MEWMA vector $\mathbf{w}_t$ of the current sample mean $\bar{\mathbf{z}}_t$ and the center $\mathbf{a}$ of SVDD sphere is greater than h .

Since the chart statistic $D_{\mathbf{a}}$ does not have a known standard distribution, obtaining an appropriate value of h to achieve the pre-specified ARL_0 is challenging. To address this problem, Tang et al. [30], among others, suggested determining the value of h by simulation, employing bootstrap aggregation (bagging) techniques. These bagging methods help us to find the adjusted control limit to assure that the probability of ARL_0 attaining or exceeding the desired value is close to a high specified value $(1 - \epsilon) \cdot 100\%$, where ϵ is a small specified constant, eg, $\epsilon = 0.1$. In this work, we adopt the **bootstrap percentile** approach proposed by Tang et al. [30] to establish the control limit for our SVDD control chart.

Consider a training data set of size N , $S = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\} \subset \mathbb{R}^n$. The procedure to obtain the upper limit threshold h using the bagging method is summarized as follows:

Algorithm 2: Determining the upper limit threshold h for the SVDD control chart

Input: Training data set of size N , denoted as $S = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\}$

Output: Upper limit threshold h of SVDD control chart

- 1: Generate M independent bootstrap samples of size N from the training set.
- 2: **For** each bootstrap sample j^{th} where $j = 1, \dots, M$ **do**
- 3: Execute the SVDD algorithm to determine the SVDD sphere for the bootstrap sample.

- 4: Use formula (18) to compute N statistics $D_{\mathbf{a}^{(j)}}(\mathbf{x}_1^{(j)}), \dots, D_{\mathbf{a}^{(j)}}(\mathbf{x}_N^{(j)})$ from j^{th} bootstrap sample, where $\mathbf{a}^{(j)}$ is the center of sphere obtained from j^{th} bootstrap sample.
- 5: Let ARL_0 be the pre-specified in-control ARL and $\beta = \frac{1}{ARL_0}$ be the desired false alarm rate. Calculate $100 \times (1 - \beta)th$ percentile of the set of N statistics. This value presents the estimated limit threshold obtained from the current bootstrap sample, denoted by $h^{(j)}$.
- 6: **End for**
- 7: Determine the final control limit h by identifying the $100 \times (1 - \epsilon)$ percentile of estimated limit thresholds $h^{(j)}, j = 1, \dots, M$, where ϵ is a small predetermined constant.

This established limit threshold h will be used to monitor new objects. An out-of-control signal is issued if the monitoring statistic for the new object exceeds h . This method ensures a consistent approach for finding the threshold that balances sensitivity and specificity in the SVDD model. The summary of building an SVDD control chart according to Maboudou-Tchao et al. [20] is as follows:

Algorithm 3: SVDD-based control chart

Input: An in-control dataset (training dataset) of size N , $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N$.

Output: SVDD-based control chart.

- 1: Apply the SVDD algorithm on the in-control dataset to obtain the parameters $\alpha_i, i = 1, \dots, N$ of the SVDD sphere.
- 2: Apply the bootstrap aggregating method as in **Algorithm 2** on the training dataset to find the upper limit threshold h of the control chart.
- 3: Monitor each new observation $\mathbf{z}$ by computing its distance statistic $D_{\mathbf{a}}(\mathbf{z})$

$$D_{\mathbf{a}}(\mathbf{z}) = K(\mathbf{z}, \mathbf{z}) - 2 \sum_{i=1}^N \alpha_i K(\mathbf{z}, \mathbf{x}_i) + \sum_{i,j=1}^N \alpha_i \alpha_j K(\mathbf{x}_i, \mathbf{x}_j) \quad (21)$$

An out-of-control signal is issued when $D_{\mathbf{a}}(\mathbf{z}) > h$

4.3 SVDD-EWMA control chart for monitoring multivariate CoDa

As mentioned before, due to the special structure of the CoDa vector as its components are non-negative and constrained by $\sum_{i=1}^p x_i = 1$ constant, we can not apply the SVDD algorithm directly to these special data. To address this, we propose utilizing the ilr and Dirichlet density transformation methods discussed in Section 2 and 3 to transform the special data into normal data. By applying these transformations, the multivariate CoDa training set will be converted into

normal data in $\mathbb{R}$. Once the data is transformed, we further employ the MEWMA in equation (13) to convert the data into MEWMA vectors. This transformation allows us to leverage all available observed data, both current and historical, to assess the performance of the process. By incorporating all the data up to the current time point, we can gain a more comprehensive understanding of the process and capture changes in the data more effectively. Next, we construct SVDD control charts utilizing MEWMA-transformed vectors. The main difference between the construction of the two proposed control charts lies in the data transformation method. The ilr-transformation, which is straightforward and well-documented, can be found in various works, such as Tran et al. [32], Nguyen et al. [23], [14] and Imran et al. [13]. In this section, we delve into a detailed construction of SVDD control chart for monitoring multivariate CoDa data via Dirichlet density transformation method.

Let's consider an in-control multivariate CoDa set of size N , represented by $\mathbf{X}_1, \dots, \mathbf{X}_N$, each $\mathbf{X}_i$ encompasses l CoDa vectors $\mathbf{x}_i^{(1)}, \dots, \mathbf{x}_i^{(l)}$ with the number of parts $p_1, p_2, \dots, p_l$. As detailed in Section 3, each observation $\mathbf{X}_i$ can be expressed as

$$\mathbf{X}_i = [\mathbf{x}_i^{(1)}, \dots, \mathbf{x}_i^{(l)}] = [x_{i1}^{(1)}, x_{i2}^{(1)}, \dots, x_{ip_1}^{(1)}, \dots, x_{i1}^{(l)}, x_{i2}^{(l)}, \dots, x_{ip_l}^{(l)}]$$

We also assume that the constraint on each CoDa vector is equal to 1, i.e., $\sum_{k=1}^{p_q} x_{ik}^{(q)} = 1$ for $i = 1, \dots, N$ and $q = 1, \dots, l$. The procedure for designing the SVDD control chart for monitoring the multivariate CoDa using the Dirichlet density transformation method (denoted by S_{Dir}) is given in Algorithm 4. For simplicity, we assume the sample size is equal to 1.

Algorithm 4: SVDD control chart for monitoring multivariate CoDa.

Input: An in-control multivariate CoDa dataset S (training set) of size N with l compositional features,

$$S = \{\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_N\}$$

Output: S_{Dir} control chart for monitoring multivariate CoDa data.

1: Use Algorithm 1 to estimate the Dirichlet propability densities of l CoDa features in S .

2: Transform each individual CoDa in S into the estimated probability densities, denote this new transformed data $\hat{S}$. The data $\hat{\mathbf{x}}_k \in \hat{S}, k = 1, \dots, N$ will have form

$$\hat{\mathbf{x}}_k = (\hat{f}(\mathbf{x}_k^{(1)}), \hat{f}(\mathbf{x}_k^{(2)}), \dots, \hat{f}(\mathbf{x}_k^{(l)})) \in \mathbb{R}^l, \quad k = 1, \dots, N$$

where $\hat{f}(\mathbf{x}_k^{(j)})$ is the estimated Dirichlet density for the j^{th} compositional feature of the k^{th} individual.

3: Transform data in $\hat{S}$ into MEWMA vectors $\hat{\mathbf{w}}_i$ using formula

$$\hat{\mathbf{w}}_i = r(\hat{\mathbf{x}}_i - \boldsymbol{\mu}) + (1 - r)\hat{\mathbf{w}}_{i-1}, \quad i = 1, 2, \dots, N$$

where $\hat{\mathbf{w}}_0 = \mathbf{0}$, r is the smoothing parameter, and $\boldsymbol{\mu}$ is the sample mean of $\hat{S}$. Denote this transformed data $\hat{S}_E$.

4: Apply Algorithm 3 on $\hat{S}_E$ to train the SVDD model, obtain parameters α and the upper limit threshold h of the SVDD chart. The SVDD optimization problem using this transformed data will be

$$\max_{\alpha} \sum_{i=1}^N \alpha_i K(\hat{\mathbf{w}}_i, \hat{\mathbf{w}}_i) - \sum_{i,j=1}^N \alpha_i \alpha_j K(\hat{\mathbf{w}}_i, \hat{\mathbf{w}}_j) \quad (22)$$

$$\begin{aligned} \text{subject to} \quad & \sum_{i=1}^N \alpha_i = 1 \\ & 0 \leq \alpha_i \leq C \quad \text{for } i = 1, \dots, N \end{aligned}$$

5: **For** each new multivariate CoDa observation $\mathbf{z}_t$ obtained at time point t **do**:

- transform $\mathbf{z}_t$ into its estimated Dirichlet probability density $\hat{\mathbf{z}}_t$, and then into its MEWMA vector $\hat{\mathbf{z}}_t^*$
- Compute the chart statistic $D(\hat{\mathbf{z}}_t^*)$

$$D(\hat{\mathbf{z}}_t^*) = K(\hat{\mathbf{z}}_t^*, \hat{\mathbf{z}}_t^*) - 2 \sum_{i=1}^N \alpha_i K(\hat{\mathbf{z}}_t^*, \hat{\mathbf{w}}_i) + \sum_{i,j=1}^N \alpha_i \alpha_j K(\hat{\mathbf{w}}_i, \hat{\mathbf{w}}_j) \quad (23)$$

- An out-of-control signal is issued when $D(\hat{\mathbf{z}}_t^*) > h$

5 Performance of the SVDD-EWMA control charts for monitoring CoDa

In this section, we will compare the performance of the two proposed control charts for monitoring multivariate CoDa: the SVDD-MEWMA control chart using Dirichlet density transformation (denoted by S_{Dir}), the SVDD-MEWMA control chart using ilr transformation (denoted by S_{ilr}) with the traditional MEWMA control chart for monitoring CoDa proposed by Tran et al. [32]. Our comparison is conducted through a simulation involving CoDa datasets generated from Dirichlet distributions, assessing the charts based on the out-of-control average run length (ARL_1) while constraining on the same pre-specified in-control value (ARL_0).

Previous research by Ning and Tsung [24] has indicated that the size of the Phase I data set has a significant impact on the performance of SVDD-based control charts and it is not feasible to set an excessively high value for in-control average run length ARL_0 unless the Phase I dataset size is sufficient. Otherwise, the substantial computation time required for the training stage of the SVDD model is also a consideration. Thus, our study selects a Phase I dataset size of 100 and investigates the performance of the considered control charts in case $ARL_0 = 100$ and $ARL_0 = 200$, as it is often used in the context of SVDD-based control charts, see, Ning and Tsung [24] and Tang et al. [30], among many others.

For designing the SVDD-based chart, we will employ a Grid Search to evaluate three kernel functions: Gaussian, linear, and polynomial. Specifically, for the Gaussian kernel, we'll determine the most effective kernel width σ and trade-off C . In the case of the linear kernel, our search will focus solely on the optimal trade-off C , while for the polynomial kernel, we will identify the most suitable degree d of the polynomial and the trade-off C . Each kernel's parameter will be evaluated over a pre-determined range. Therefore, to construct an SVDD-based chart, two types of parameters must be established: the control limit h and the kernel (and its parameters). The control limit h can be determined via the bootstrap method described in Algorithm 1, leaving the kernel (and its parameters) to be identified. However, determining the optimal kernel and its parameters presents a complex challenge. To date, there is no established method for efficiently selecting these parameters due to lack of information regarding the classification error. The determination of the most effective method for identifying the kernel and its parameters falls outside the scope of this work. To find a suitable kernel, we will execute the Grid Search method over three kernel functions to identify the one that most accurately represents the data, ensuring that the maximum amount of in-control data is enclosed within the hypersphere.

5.1 Prepare the phase I data set

To utilize the concept of the SVDD algorithm and the Dirichlet transformation method in Section 3 in constructing the S_{Dir} control chart for monitoring CoDa, and to ensure its suitability in the perspective of SPC, we adopt the following data generation approach. We begin with an in-control dataset presumed to follow a Dirichlet distribution $\mathcal{D}(\alpha_0)$ with $\alpha_0 = (1, 2, 3)$. Each sample, composed of three in-control observations, is combined to form a "new observation". These samples constitute "new observations" of dimension 9 (or dimension 3 in the context of multivariate CoDa) in our training dataset. Through this process, we generate 4500 CoDa vectors following the $\mathcal{D}(\alpha_0)$ distribution and create new observations using this approach. The resulting dataset consists of 1500 data points of dimension 9 (or dimension 3 in multivariate CoDa context), which we refer to as $(1500, 1)$ or $(1500, 3)_{CoDa}$ dimension. We then apply the Dirichlet transformation method from Section 3 to transform the data into their Dirichlet densities. The final dataset used for training the SVDD algorithm has dimension $(1500, 3)$.

The training sets for the S_{ilr} and MEWMA control charts are generated as follows: Each sample of size $m = 3$ in the in-control dataset is grouped, and the mean of each group is considered as a "new observation" in the training set. The ilr transformation is then applied to this dataset to address the constraint issue. Consequently, the final dataset used for training the S_{ilr} and MEWMA control charts will have a dimension of $(1500, 2)$.

Table 1 illustrates the 20 samples of original data, their corresponding Dirichlet densities transformed data, and the sample means of their ilr transformed data. The first 9 columns present 3 CoDa features $\mathbf{x}_i^{(1)}$, $\mathbf{x}_i^{(2)}$, and $\mathbf{x}_i^{(3)}$, the next three columns z_{i1} , z_{i2} , z_{i3} present their corresponding Dirichlet densities while the last two columns x present the mean of their ilr transform.

5.2 Comparison between control charts

When the process is out of control, we are interested in detecting the change of in-control distribution $\mathcal{D}(\alpha_0)$ to $\mathcal{D}(\alpha_1)$. Some changes in the components of α_0 will be considered. Note

Table 1: Original data generated from $\mathcal{D}(\alpha_0)$ (the first 9 columns), their Dirichlet transforms (the next 3 columns) and their sample mean ilr transforms (the last 2 columns)

i	$\mathbf{x}_i^{(1)}$			$\mathbf{x}_i^{(2)}$			$\mathbf{x}_i^{(3)}$			Dirichlet transformation			ilr transformation	
	$(x_1$	x_2	$x_3)$	$(x_1$	x_2	$x_3)$	$(x_1$	x_2	$x_3)$	$z_{i,1}$	$z_{i,2}$	$z_{i,3}$	$\bar{x}_{i,1}^*$	$\bar{x}_{i,2}^*$
1	0.133	0.4	0.467	0.103	0.615	0.282	0.221	0.138	0.641	5.243	2.95	3.397	-0.571	-0.579
2	0.186	0.199	0.615	0.18	0.39	0.43	0.203	0.283	0.514	4.511	4.321	4.486	-0.275	-0.655
3	0.147	0.578	0.275	0.059	0.266	0.675	0.358	0.356	0.286	2.616	7.264	1.746	-0.675	-0.381
4	0.079	0.617	0.304	0.203	0.361	0.436	0.004	0.299	0.697	3.427	4.126	8.725	-1.663	-1.05
5	0.091	0.036	0.873	0.166	0.05	0.784	0.213	0.107	0.68	1.64	1.856	2.97	0.663	-1.736
6	0.073	0.283	0.644	0.304	0.114	0.582	0.029	0.35	0.621	7.047	2.314	8.093	-0.674	-1.214
7	0.035	0.775	0.19	0.326	0.174	0.5	0.317	0.077	0.606	1.67	2.613	1.688	-0.244	-0.609
8	0.086	0.445	0.469	0.115	0.279	0.606	0.071	0.564	0.365	5.88	6.151	4.508	-1.086	-0.734
9	0.024	0.083	0.893	0.125	0.344	0.531	0.33	0.474	0.196	3.98	5.822	1.09	-0.621	-0.883
10	0.029	0.345	0.626	0.059	0.244	0.697	0.205	0.35	0.445	8.114	7.121	4.157	-1.047	-1.117
11	0.231	0.634	0.135	0.532	0.073	0.395	0.159	0.228	0.613	0.696	0.686	5.137	0.143	-0.222
12	0.155	0.345	0.5	0.052	0.629	0.319	0.211	0.566	0.223	5.178	3.836	1.691	-1.009	-0.245
13	0.431	0.317	0.252	0.076	0.285	0.639	0.183	0.099	0.718	1.211	6.972	3.058	-0.092	-0.75
14	0.116	0.305	0.579	0.392	0.187	0.421	0.251	0.126	0.623	6.138	1.993	2.93	0.109	-0.768
15	0.299	0.448	0.253	0.042	0.152	0.806	0.175	0.127	0.698	1.709	5.921	3.709	-0.32	-0.946
16	0.087	0.312	0.601	0.102	0.516	0.382	0.149	0.339	0.512	6.752	4.536	5.335	-0.876	-0.716
17	0.167	0.487	0.346	0.035	0.541	0.424	0.176	0.365	0.459	3.493	5.85	4.606	-1.072	-0.521
18	0.163	0.459	0.378	0.143	0.168	0.689	0.181	0.362	0.457	3.942	4.788	4.543	-0.445	-0.652
19	0.323	0.167	0.51	0.025	0.01	0.965	0.116	0.296	0.588	2.601	0.573	6.136	0.147	-1.642
20	0.083	0.262	0.655	0.03	0.176	0.794	0.247	0.174	0.579	6.739	6.665	3.503	-0.607	-1.338

that the mean μ_0 and the covariance matrix Σ_0 of $\mathcal{D}(\alpha_0)$ distribution are

$$\mu_0 = \left(\frac{1}{6}, \frac{1}{3}, \frac{1}{2}\right) \quad \Sigma_0 = \begin{pmatrix} 0.0198 & -0.0079 & -0.0119 \\ -0.0079 & 0.0317 & -0.0238 \\ -0.0119 & -0.0238 & 0.0357 \end{pmatrix} \quad (24)$$

Figure 1 illustrates the ternary diagram of four samples from different Dirichlet distributions $\mathcal{D}(1, 2, 3)$, $\mathcal{D}(2, 2, 3)$, $\mathcal{D}(3, 2, 3)$, and $\mathcal{D}(4, 2, 3)$. As can be seen, the data distribution is impacted by the change in the component of the parameter. We will test the performance of the above SVDD-based control charts when the first component in parameter α_0 is changed. We considered 12 cases for α_1

$$\alpha_1 \in \{(0.3, 2, 3), (0.5, 2, 3), (0.7, 2, 3), (1.5, 2, 3), (2, 2, 3), (2.5, 2, 3), (3, 2, 3), (3.5, 2, 3), (4, 2, 3), (4.5, 2, 3), (5, 2, 3)\} \quad (25)$$

The choice of the 12 α_1 cases are done in a deliberate approach rather than making arbitrary choices, to test the performance of control charts in detecting shifts within a CoDa set generated from a Dirichlet distribution with the initial parameters of $\alpha_0 = (1, 2, 3)$. The objective is to simulate a comprehensive range of scenarios, reflecting both minor and major shifts in the relative proportions of components within the dataset. By adjusting the first parameter of α_0 in both directions—decreasing it to values of 0.3, 0.5, 0.7, and increasing it to 1.5, 2, 2.5, 3, 3.5, 4, 4.5, and 5, we aim to explore the effect of reducing/increasing the first component on the performance of the control charts. It's important to note that changes to this first component inherently affect the other components as well, leading to a wide spectrum of compositional changes. These choices of α_1 allow us to assess the control charts' performances across different levels of compositional changes, ensuring a thorough evaluation of their performance.

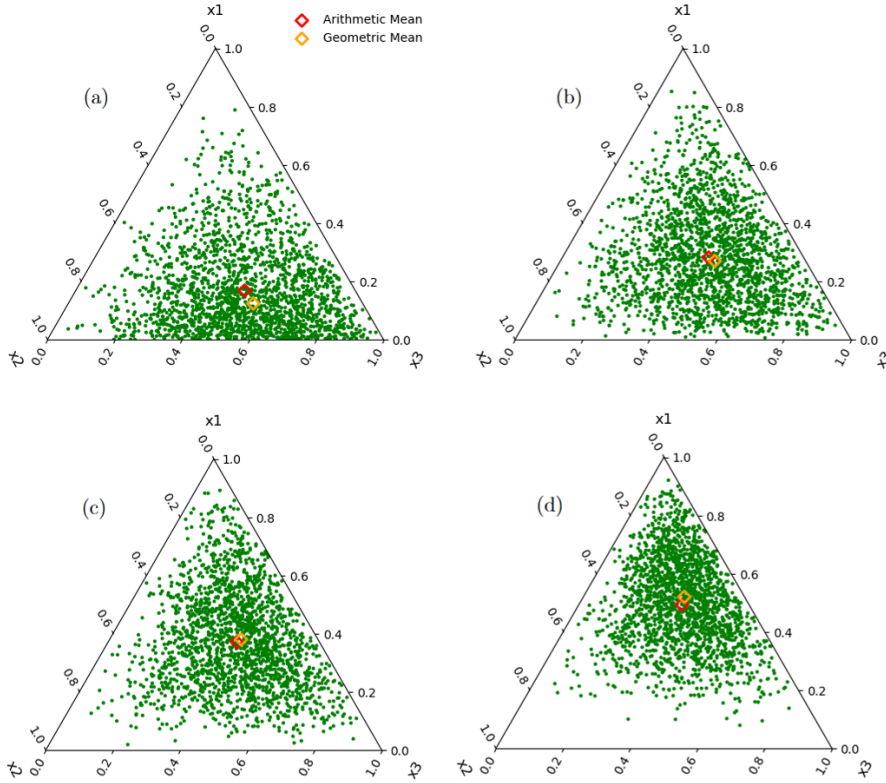


Figure 1: Ternary diagram of samples from Dirichlet distributions $\mathcal{D}(\alpha)$ with different parameters α : (a) $\alpha = (1, 2, 3)$; (b) $\alpha = (2, 2, 3)$; (c) $\alpha = (3, 2, 3)$; (d) $\alpha = (4, 2, 3)$

As stated before, the comparison between the control chart will be based on the values of out-of-control ARL_1 while constraining on the same in-control values of $ARL_0 = 100$ and $ARL_0 = 200$. The smoothing parameter r in the MEWMA vector determines the weight given to each observation, and a higher value places more emphasis on recent data points. The choice of the smoothing parameter depends on specific objectives, the characteristics of the process, and the level of responsiveness desired in the control chart. We will consider 6 different values of r for each value of α_1 , i.e.,

$$r \in \{0.2, 0.3, 0.4, 0.5, 0.6, 0.7\}.$$

Monte Carlo simulation with 10^4 iterations will be used to estimate the ARL_1 values in each scenario. Table 2 shows the values ARL_1 of S_{Dir} , S_{ilr} and MEWMA control charts for monitor-

ing CoDa in different scenarios.

From this table, we can see

- When the smoothing parameter (r) increases, the performance of all three considered control charts tends to decrease.
- In general, the S_{ilr} control chart demonstrates superior performance compared to the S_{Dir} and MEWMA charts. However, in cases where the smallest scaled component in the CoDa vector is decreased (and the other two components are increased), the S_{ilr} chart's performance is slightly inferior to the MEWMA control chart.
- As the smoothing parameter (r) increases, the performance of the S_{Dir} control chart becomes increasingly similar to that of the MEWMA control chart. Specifically, when r reaches values of 0.6 and 0.7 (in case $ARL_0 = 100$) and when r reaches values of 0.5, 0.6, and 0.7 (in case $ARL_0 = 200$), the efficiency of the S_{Dir} chart surpasses the performance of MEWMA control chart, particularly when the smallest component in CoDa starts to increase.

6 Illustrative example

In this section, we demonstrate the application of the proposed S_{Dir} and S_{ilr} charts for detecting changes in the distribution of CoDa. To illustrate this, we will utilize a dataset that describes the particle-size distribution (expressed as a percentage by weight) for a plant located in Europe. This dataset has been analyzed in various studies related to CoDa, including works by Montgomery [21], Vives-Mestres et al. [35], Tran et al. [32], Imran et al. [14], among many others. By applying the S_{Dir} and S_{ilr} control charts, we aim to detect any changes in the particle-size distribution in that plant. According to the study conducted by Vives-Mestres et al. [35], the dataset contained four identified out-of-control points (#1, #26, #45, and #52). After identifying these out-of-control points, they were removed from the dataset to create the Phase I dataset. The resulting Phase I dataset, after removing the out-of-control points, is normalized and presented in Table 3. This table also includes their ilr transformations, their Dirichlet densities, and the monitoring statistics D_{ilr} and D_{Dir} which are derived using ilr and Dirichlet density transformations of the control charts, respectively. In this example, to estimate the Dirichlet distribution's parameter α of our dataset, we utilized the Python package provided by Eric J. Suh, available at Dirichlet MLE python library. Specifically, we employed the maximum likelihood estimation (MLE) method implemented within this package to estimate the Dirichlet distribution's parameter. The ilr -transformation $\mathbf{x}_i^* = (x_{i1}^*, x_{i2}^*)$ of the dataset is calculated using Equation (8), i.e.,

$$\begin{aligned} x_{i,1}^* &= \sqrt{\frac{1}{2}} \ln \left(\frac{M}{S} \right), \\ x_{i,2}^* &= \sqrt{\frac{2}{3}} \ln \left(\sqrt{\frac{MS}{L}} \right). \end{aligned} \quad (26)$$

~~The parameter α of the Dirichlet distribution for our dataset, is estimated using the MLE method.~~ To train the model, a sample size $m = \hat{n}$ as chosen and the control limits of the two

Table 2: ARL₁ values of different control charts in different cases of α_1 and ARL₀

α_1	r = 0.2						r = 0.3					
	ARL ₀ = 100			ARL ₀ = 200			ARL ₀ = 100			ARL ₀ = 200		
	S _{Dir}	S _{ilr}	EWMA	S _{Dir}	S _{ilr}	EWMA	S _{Dir}	S _{ilr}	EWMA	S _{Dir}	S _{ilr}	EWMA
(0.3, 2, 3)	2.58	1.30	1.30	2.53	1.33	1.26	3.11	1.38	1.30	3.68	1.44	1.34
(0.5, 2, 3)	5.73	2.63	2.58	5.79	2.83	2.34	7.49	3.01	2.48	10.05	3.33	2.81
(1.5, 2, 3)	11.21	3.82	17.10	11.68	4.79	11.22	18.9	5.36	14.1	30.48	7.38	25.83
(2, 2, 3)	2.97	1.54	2.96	2.97	1.71	2.50	4.44	1.90	3.20	5.66	2.23	4.47
(2.5, 2, 3)	1.65	1.15	1.56	1.67	1.21	1.42	2.15	1.28	1.69	2.41	1.37	1.97
(3, 2, 3)	1.26	1.04	1.20	1.27	1.06	1.14	1.47	1.09	1.25	1.57	1.13	1.37
(3.5, 2, 3)	1.10	1.01	1.06	1.11	1.01	1.05	1.22	1.03	1.10	1.25	1.04	1.14
(4, 2, 3)	1.04	1.003	1.02	1.04	1.005	1.01	1.09	1.009	1.03	1.13	1.01	1.06
(4.5, 2, 3)	1.017	1.001	1.009	1.015	1.001	1.006	1.04	1.002	1.01	1.06	1.004	1.02
(5, 2, 3)	1.01	1.0002	1.003	1.01	1.0006	1.002	1.02	1.001	1.004	1.03	1.001	1.01
α_1	r = 0.4						r = 0.5					
	ARL ₀ = 100			ARL ₀ = 200			ARL ₀ = 100			ARL ₀ = 200		
	S _{Dir}	S _{ilr}	EWMA	S _{Dir}	S _{ilr}	EWMA	S _{Dir}	S _{ilr}	EWMA	S _{Dir}	S _{ilr}	EWMA
(0.3, 2, 3)	3.64	1.48	1.37	4.53	1.57	1.43	3.99	1.85	1.44	5.05	1.96	1.54
(0.5, 2, 3)	9.07	3.67	2.84	12.5	4.14	3.30	9.97	6.44	3.18	14	7.34	3.67
(1.5, 2, 3)	28.2	7.43	24.68	47.09	11.17	51.89	41.48	8.47	41.42	75.19	13.02	92.81
(2, 2, 3)	6.26	2.39	5.26	8.89	2.94	8.68	9.53	2.74	9.05	14.25	3.54	17.97
(2.5, 2, 3)	2.75	1.45	2.34	3.4	1.63	3.24	3.82	1.6	3.59	4.88	1.84	5.71
(3, 2, 3)	1.74	1.16	1.52	1.98	1.25	1.86	2.44	1.23	2.09	2.57	1.35	2.86
(3.5, 2, 3)	1.36	1.06	1.24	1.46	1.09	1.37	1.57	1.08	1.52	1.75	1.14	1.89
(4, 2, 3)	1.17	1.01	1.09	1.23	1.03	1.17	1.36	1.03	1.25	1.39	1.06	1.45
(4.5, 2, 3)	1.08	1.008	1.04	1.11	1.01	1.07	1.16	1.01	1.12	1.21	1.02	1.23
(5, 2, 3)	1.04	1.002	1.02	1.05	1.004	1.03	1.08	1.005	1.06	1.11	1.01	1.12
α_1	r = 0.6						r = 0.7					
	ARL ₀ = 100			ARL ₀ = 200			ARL ₀ = 100			ARL ₀ = 200		
	S _{Dir}	S _{ilr}	EWMA	S _{Dir}	S _{ilr}	EWMA	S _{Dir}	S _{ilr}	EWMA	S _{Dir}	S _{ilr}	EWMA
(0.3, 2, 3)	4.35	1.78	1.52	5.35	1.85	1.61	5.58	1.87	1.58	7.19	1.98	1.71
(0.5, 2, 3)	10.12	5.35	3.5	14.49	6.15	4.26	13.66	5.76	3.89	20.31	6.69	4.73
(1.5, 2, 3)	57.77	11.26	59.54	113.26	18.48	142.34	61.98	14.10	80.32	121.92	24.85	197.75
(2, 2, 3)	14.15	3.45	15.37	23.29	4.83	34.03	17.94	4.44	24.86	29.9	6.68	59.96
(2.5, 2, 3)	5.47	1.88	5.89	7.53	2.35	11.24	6.83	2.32	9.78	10.05	3.01	21.90
(3, 2, 3)	2.85	1.36	3.08	3.61	1.56	5.01	3.6	1.54	4.97	4.77	1.9	9.81
(3.5, 2, 3)	1.93	1.15	2.02	2.3	1.25	2.97	2.31	1.26	3.02	2.86	1.43	5.25
(4, 2, 3)	1.49	1.07	1.55	1.68	1.11	2.06	1.71	1.12	2.09	1.98	1.20	3.32
(4.5, 2, 3)	1.28	1.02	1.29	1.37	1.05	1.58	1.41	1.05	1.64	1.57	1.10	2.36
(5, 2, 3)	1.15	1.01	1.16	1.21	1.02	1.34	1.24	1.02	1.37	1.34	1.04	1.85

Table 3: The normalized Phase I dataset after removing the out-of-control points from Tran et al. [32], their corresponding ilr and Dirichlet densities transformation, and their statistics D_{ilr} , D_{Dir}

i	M	S	L	x_{i1}^*	x_{i2}^*	Dirichlet density	D_{ilr}	D_{Dir}	i	M	S	L	x_{i1}^*	x_{i2}^*	Dirichlet density	D_{ilr}	D_{Dir}
1	0.926	0.042	0.032	2.19	1.48	212.292	<i>0.0124</i>	<i>3.102</i>	27	0.836	0.09	0.074	1.58	1.07	53.748	<i>0.0725</i>	<i>67.436</i>
2	0.917	0.031	0.052	2.4	0.96	188.172			28	0.848	0.084	0.068	1.63	1.12	99.442		
3	0.869	0.096	0.035	1.56	1.72	66.359	<i>0.0494</i>	<i>29.007</i>	29	0.871	0.066	0.063	1.82	1.09	253.732	<i>0.0676</i>	<i>42.629</i>
4	0.904	0.067	0.029	1.84	1.75	165.819			30	0.872	0.067	0.061	1.81	1.12	261.387		
5	0.921	0.033	0.046	2.35	1.09	217.724	<i>0.036</i>	<i>35.824</i>	31	0.873	0.061	0.066	1.88	1.02	260.438	<i>0.0725</i>	<i>26.634</i>
6	0.915	0.041	0.044	2.2	1.21	323.485			32	0.848	0.09	0.062	1.59	1.22	89.261		
7	0.903	0.047	0.05	2.09	1.16	386.404	<i>0.0144</i>	<i>21.82</i>	33	0.874	0.061	0.065	1.88	1.03	269.595	<i>0.0712</i>	<i>5.1</i>
8	0.851	0.065	0.084	1.82	0.84	100.415			34	0.868	0.072	0.06	1.76	1.17	226.607		
9	0.897	0.061	0.042	1.9	1.4	347.602	<i>0.0258</i>	<i>48.375</i>	35	0.888	0.064	0.048	1.86	1.31	344.851	<i>0.0636</i>	<i>50.154</i>
10	0.925	0.037	0.038	2.28	1.29	233.181			36	0.898	0.053	0.049	2	1.22	400.23		
11	0.918	0.039	0.043	2.23	1.21	294.301	<i>0.0359</i>	<i>67.436</i>	37	0.869	0.073	0.058	1.75	1.2	229.502	<i>0.0771</i>	<i>27.168</i>
12	0.917	0.046	0.037	2.12	1.4	305.84			38	0.838	0.09	0.072	1.58	1.09	59.302		
13	0.903	0.059	0.038	1.93	1.47	322.787	<i>0.0576</i>	<i>67.436</i>	39	0.892	0.052	0.056	2.01	1.1	375.7	<i>0.0765</i>	<i>54.636</i>
14	0.945	0.029	0.026	2.46	1.51	55.261			40	0.845	0.086	0.069	1.62	1.11	85.538		
15	0.945	0.028	0.027	2.49	1.47	55.3	<i>0.0642</i>	<i>24.953</i>	41	0.844	0.082	0.074	1.65	1.04	84.053	<i>0.0868</i>	<i>34.537</i>
16	0.887	0.034	0.079	2.31	0.64	106.218			42	0.843	0.068	0.089	1.78	0.81	67.236		
17	0.846	0.088	0.066	1.6	1.16	86.574	<i>0.0425</i>	<i>30.909</i>	43	0.898	0.02	0.082	2.69	0.4	17.861	<i>0.0708</i>	<i>41.207</i>
18	0.907	0.053	0.04	2.01	1.39	358.74			44	0.904	0.029	0.067	2.43	0.72	116.364		
19	0.902	0.073	0.025	1.78	1.9	93.631	<i>0.0621</i>	<i>40.048</i>	45	0.901	0.04	0.059	2.2	0.95	290.95	<i>0.0785</i>	<i>52.746</i>
20	0.927	0.035	0.038	2.32	1.27	207.773			46	0.836	0.077	0.087	1.69	0.87	52.07		
21	0.915	0.057	0.028	1.96	1.71	181.211	<i>0.0868</i>	<i>38.757</i>	47	0.88	0.056	0.064	1.95	1.02	301.818	<i>0.0849</i>	<i>62.972</i>
22	0.918	0.053	0.029	2.02	1.66	200.901			48	0.847	0.069	0.084	1.77	0.86	87.715		
23	0.906	0.061	0.033	1.91	1.6	249.271	<i>0.074</i>	<i>35.443</i>	49	0.93	0.019	0.051	2.75	0.78	35.526	<i>0.0868</i>	<i>34.541</i>
24	0.873	0.055	0.072	1.95	0.91	226.237			50	0.914	0.036	0.05	2.29	1.05	268.418		
25	0.826	0.104	0.07	1.47	1.17	26.124	<i>0.0711</i>	<i>44.321</i>	51	0.862	0.088	0.05	1.61	1.39	129.488	<i>0.0614</i>	<i>30.178</i>
26	0.835	0.105	0.06	1.47	1.3	36.366			52	0.872	0.069	0.059	1.79	1.16	258.637		
$h_{\text{Dir}} = 67.436; h_{\text{ilr}} = 0.08683$																	

charts S_{Dir} and S_{ilr} , denoted by h_{Dir} and h_{ilr} , respectively, have been obtained

$$\alpha = (109.44, 7.29, 6.87) \quad (27)$$

and

$$h_{\text{Dir}} = 67.436; h_{\text{ilr}} = 0.08683 \quad (28)$$

For the Phase II dataset, a dataset with a sample size $m = 2$ is also conducted. The simulation will reflect an in-control process for the first 10 samples and then be followed by a shift in distribution between the 10th and 11th samples due to an assignable cause. Specifically, the first 10 samples will be generated to follow a Dirichlet distribution characterized by $\alpha = (109.44, 7.29, 6.87)$. Subsequently, the next set of 10 samples of Dirichlet distribution with parameters

$$\alpha_1 = (100, 10, 8) \quad (29)$$

Table 4: The Phase II dataset for testing the proposed control charts

i	M	S	L	x_{i1}^*	x_{i2}^*	Dirichlet density	D_{ilr}	D_{Dir}	i	M	S	L	x_{i1}^*	x_{i2}^*	Dirichlet density	D_{ilr}	D_{Dir}
1	0.92	0.04	0.03	2.19	1.43	240.95	0.0204	6.43	21	0.89	0.06	0.05	1.91	1.28	372.7	0.0307	20.16
2	0.86	0.08	0.06	1.69	1.14	150.07			22	0.8	0.15	0.05	1.21	1.54	1.69		
3	0.89	0.05	0.06	2.05	0.98	320.73	0.0312	34.78	23	0.88	0.03	0.09	2.28	0.58	78.19	0.0373	30.77
4	0.85	0.09	0.05	1.55	1.36	83.73			24	0.86	0.09	0.05	1.61	1.35	124.2		
5	0.89	0.03	0.08	2.37	0.61	87.38	0.0384	30.08	25	0.79	0.12	0.09	1.32	0.99	1.77	0.0733	70.12
6	0.9	0.03	0.07	2.41	0.71	116.44			26	0.84	0.08	0.08	1.66	0.93	57.63		
7	0.83	0.1	0.08	1.5	1.08	28.25	0.0215	60.4	27	0.88	0.07	0.05	1.8	1.35	299	0.0714	36.88
8	0.94	0.03	0.03	2.45	1.53	56.83			28	0.88	0.08	0.05	1.71	1.4	214.52		
9	0.9	0.05	0.05	2.11	1.07	363.69	0.0436	43.29	29	0.86	0.07	0.07	1.8	1.02	195.73	0.0798	37.11
10	0.91	0.03	0.06	2.39	0.78	151.45			30	0.86	0.09	0.05	1.63	1.31	133.92		
11	0.87	0.09	0.05	1.64	1.46	151.01	0.0134	9.68	31	0.83	0.1	0.07	1.51	1.13	34.17	0.0991	62.98
12	0.9	0.06	0.04	1.92	1.34	371.12			32	0.86	0.09	0.05	1.6	1.33	113.53		
13	0.91	0.04	0.06	2.27	0.93	254.29	0.0271	34.79	33	0.78	0.11	0.11	1.41	0.8	1.49	0.1267	102.8
14	0.88	0.06	0.06	1.95	1.06	331.59			34	0.81	0.1	0.1	1.51	0.86	9.16		
15	0.91	0.03	0.06	2.47	0.73	106.18	0.0624	20.05	35	0.89	0.04	0.07	2.19	0.79	195.4	0.1086	77.25
16	0.86	0.05	0.09	1.96	0.71	90.65			36	0.87	0.07	0.06	1.75	1.2	233.57		
17	0.89	0.07	0.04	1.81	1.52	261.81	0.0322	51.75	37	0.85	0.08	0.07	1.64	1.13	104.65	0.1063	73.45
18	0.9	0.05	0.05	2.04	1.27	392.15			38	0.88	0.07	0.05	1.82	1.33	310.86		
19	0.89	0.03	0.09	2.4	0.53	56.11	0.0351	37.55	39	0.86	0.09	0.06	1.62	1.27	119.32	0.1012	79.14
20	0.86	0.1	0.04	1.54	1.57	73.91			40	0.88	0.05	0.06	2	0.99	309.09		
$h_{\text{Dir}} = 67.436; h_{\text{ilr}} = 0.08683$																	

will reflect a change in the process. The Phase II dataset with its statistics D based on ilr and Dirichlet transformations and Equation (20) are given in Table 4.

Figure 2 shows the S_{ilr} and S_{Dir} control charts. As can be seen from this figure, the S_{Dir} control chart detects the out-of-control situation at sample point #13 where this data point surpasses the control limit $h_{\text{Dir}} = 67.436$. the S_{ilr} chart signals an out-of-control condition later at sample point #16. In this example, the S_{Dir} chart demonstrates a more prompt response in detecting anomalies, signaling the out-of-control condition earlier than the S_{ilr} chart.

7 Conclusion

This study presents a new methodology for transforming multivariate CoDa into real vectors in $\mathbf{R}^n$, namely Dirichlet density transformation, thereby eliminating inherent constraints and reducing data dimensionality. Furthermore, we propose two distinct distribution-free approaches for monitoring CoDa, utilizing a combination of SVDD control charts and the multivariate exponentially weighted moving average technique and two transformation methods: the Dirichlet density transformation and the isometric log-ratio transformation. A notable advantage of

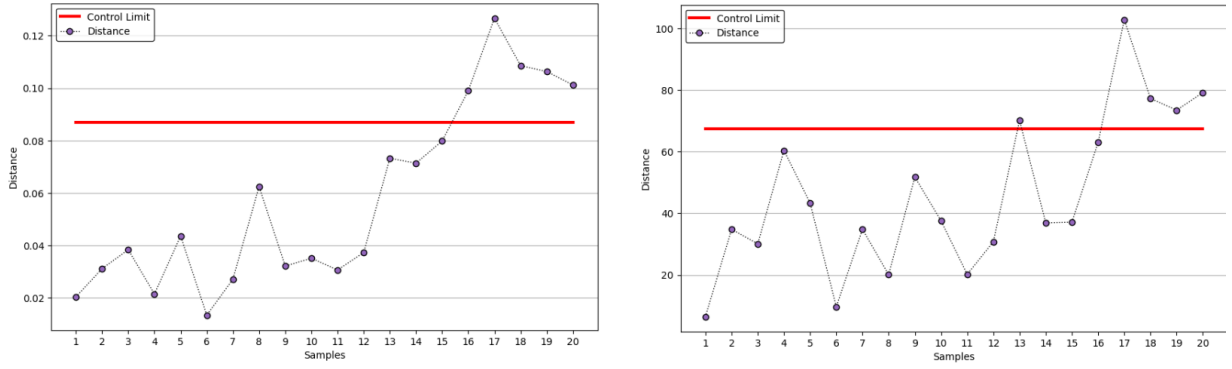


Figure 2: The S_{illr} (left side) and S_{Dir} (right side) control charts for the illustrative example of Phase II dataset

these control charts is their ability to operate without requiring any distribution assumptions. Through extensive simulation studies, the proposed S_{illr} control chart consistently demonstrates superior performance compared to the S_{Dir} control chart and the MEWMA control chart for monitoring CoDa in most of the scenarios.

Future research in the field of CoDa monitoring could focus on optimizing the parameters of the SVDD sphere and the smoothing parameter in the MEWMA vector. Although the S_{Dir} control chart using the Dirichlet transformation method does not surpass the S_{illr} control chart, it remains a noteworthy approach, particularly in scenarios where a multivariate CoDa set with numerous components is involved. Due to the diverse applications of CoDa in real life, SPC researchers must prioritize the development of online monitoring techniques. These efforts can lead to improved process control, proactive interventions, and enhanced quality assurance across industries, ensuring efficient utilization of CoDa.

References

- [1] M. Ahsan, H. Khusna, Wibawati, and M.H. Lee. Support vector data description with kernel density estimation (svdd-kde) control chart for network intrusion monitoring. *Scientific Reports*, 13(1):19149, 2023.
- [2] J. Aitchison. *The Statistical Analysis of Compositional Data (Monographs on Statistics and Applied Probability)*. Chapman & Hall Ltd., London, (Reprinted in 2003 with additional material by The Blackburn Press), 1986.
- [3] A. Banerjee, P. Burlina, and R. Meth. Fast hyperspectral anomaly detection via svdd. In *2007 IEEE International Conference on Image Processing*, volume 4, pages IV–101 – IV–104, 2007.
- [4] R. A. Boyles. Using the chi-square statistic to monitor compositional process data. *Journal of Applied Statistics*, 24(5):589–602, 1997.
- [5] J. J. Egozcue, V. Pawlowsky-Glahn, G. Mateu-Figueras, and C. Barceló-Vidal. Isometric logratio transformations for compositional data analysis. *Mathematical Geology*, 35(3): 279–300, 2003.
- [6] J.J. Egozcue and V. Pawlowsky-Glahn. Groups of Parts and Their Balances in Compositional Data analysis. *Mathematical Geology*, 37(7):795–828, 2005.
- [7] J. Fan, Y. Feng, J. Jiang, and X. Tong. Feature augmentation via nonparametrics and selection (fans) in high dimensional classification, 2016.
- [8] Jie Gu, Bin Cui, and Shan Lu. A classification framework for multivariate compositional data with dirichlet feature embedding. *Knowledge-Based Systems*, 212:106614, 2021.
- [9] Rubén Darío Guevara-González, José Alberto Vargas-Navas, and Dorian Luis Linero-Segrera. Profile monitoring for compositional data. *Revista Colombiana de Estadística*, 37(1):159–181, 2014.
- [10] R. K. S Hankin. A generalization of the dirichlet distribution. *Journal of Statistical Software*, 33, 2010.
- [11] H. S. Hariharan and R. P. Velu. On estimating dirichlet parameters - a comparison of initial values. *Journal of Statistical Computation and Simulation*, 48:47–58, 1993.
- [12] M. Imran, J. Sun, F. S. Zaidi, Z. Abbas, and H. Z. Nazir. Effect of measurement errors on the multivariate cusum coda control chart for the manufacturing process, 2022.
- [13] M. Imran, J. Sun, F. S. Zaidi, Z. Abbas, and H. Z. Nazir. On designing efficient multivariate exponentially weighted moving average control chart for compositional data using variable sample size. *Journal of Statistical Computation and Simulation*, pages 1–22, 2022.
- [14] M. Imran, J. Sun, F. S. Zaidi, Z. Abbas, and H. Z. Nazir. Multivariate cumulative sum control chart for compositional data with known and estimated process parameters. *Quality and Reliability Engineering International*, 38(5):2691–2714, 2022.
- [15] M. Imran, H.L. Dai, F.S. Zaidi, K.P. Tran, Z. Abbas, and H.Z. Nazir. Incorporating principal component analysis into hotelling t2 control chart for compositional data monitoring. *Computers & Industrial Engineering*, 186:109755, 2023.

- [16] M. Imran, J. Sun, X. Hu, F.S. Zaidi, and A. Tang. Investigating zero-state and steady-state performance of mewma-coda control chart using variable sampling interval. *Journal of Applied Statistics*, pages 1–22, 2023.
- [17] M. Imran, J. Sun, F.S. Zaidi, Z. Abbas, and H.Z. Nazir. Evaluating the performance of variable sampling interval hotelling t2 charting scheme for compositional data in the presence of measurement error. *Quality and Reliability Engineering International*, 2023.
- [18] M. Imran, H.L. Dai, F.S. Zaidi, X. Hu, K.P. Tran, and J. Sun. Analyzing out-of-control signals of t2 control chart for compositional data using artificial neural networks. *Expert Systems with Applications*, 238:122165, 2024.
- [19] S. Kumar, A. K. Choudhary, M. Kumar, R. Shankar, and M. K. Tiwari. Kernel distance-based robust support vector methods and its application in developing a robust k-chart. *International Journal of Production Research*, 44(1):77–96, 2006.
- [20] E. M. Maboudou-Tchao, I. R. Silva, and N. Diawara. Monitoring the mean vector with mahalanobis kernels. *Quality Technology & Quantitative Management*, 15(4):459–474, 2018.
- [21] D.C. Montgomery. *Statistical Quality Control: a Modern Introduction, 7th Edn.* Wiley, New York, 2013.
- [22] K. W. Ng, G. L. Tian, and M. L. Tang. *Dirichlet and Related Distributions: Theory, Methods and Applications.* John Wiley & Sons, 2011.
- [23] T.T.V. Nguyen, C. Heuchenne, and K. P. Tran. Anomaly detection for compositional data using vsi mewma control chart. *IFAC-PapersOnLine*, 55(10):1533–1538, 2022. ISSN 2405-8963. 10th IFAC Conference on Manufacturing Modelling, Management and Control MIM 2022.
- [24] X. Ning and F. Tsung. Improved design of kernel distance-based charts using support vector methods. *IIE Transactions*, 45(4):464–476, 2013.
- [25] A. Ongaro and S. Migliorati. A generalization of the dirichlet distribution. *Journal of Multivariate Analysis*, 114:412–426, 2013.
- [26] E. J. Pauwels and O. Ambekar. One class classification for anomaly detection: Support vector data description revisited. In Petra Perner, editor, *Advances in Data Mining. Applications and Theoretical Aspects*, pages 25–39, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg.
- [27] V. Pawlowsky-Glahn, J. J. Egozcue, and R. Tolosana-Delgado. *Modeling and Analysis of Compositional Data.* John Wiley & Sons, 2015.
- [28] A. Smola, B. Scholkopf, and K. Muller. The connection between regularization operators and support vector kernels. *Neural Networks*, pages 637–649, 1998.
- [29] R. Sun and F. Tsung. A kernel-distance-based multivariate control chart using support vector methods. *International Journal of Production Research*, 41(13):2975–2989, 2003.
- [30] A. Tang, P. Castagliola, X. Hu, and F. Xie. An assessment for the conditional performance of an support vector data description (svdd)-based chart. *Quality and Reliability Engineering International*, pages 1–17, 2022.

- [31] D. M. J. Tax and R. P. W. Duin. Support Vector Data Description. *Machine Learning*, 54(1):45–66, 2004.
- [32] K. P. Tran, P. Castagliola, G. Celano, and Michael B.C. Khoo. Monitoring compositional data using multivariate exponentially weighted moving average scheme. *Quality and Reliability Engineering International*, 34(3):391–402, 2017.
- [33] V. N. Vapnik. *Statistical Learning Theory*. John Wiley & Sons, 1998.
- [34] M. Vives-Mestres, J. Daunis-I-Estadella, and J.A. Martin-Fernandez. Out-of-Control Signals in Three-Part Compositional T^2 Control Chart. *Quality and Reliability Engineering International*, 30(3):337–346, 2014.
- [35] M. Vives-Mestres, J. Daunis-I-Estadella, and J.A. Martin-Fernandez. Individual T^2 Control Chart for Compositional Data. *Journal of Quality Technology*, 46(2):127–139, 2014.
- [36] M. Vives-Mestres, J. Daunis-I-Estadella, and J.A. Martin-Fernandez. Signal Interpretation in Hotelling’s T^2 Control Chart for Compositional Data. *IIE Transactions*, 48(7):661–672, 2016.
- [37] F. S. Zaidi, P. Castagliola, K. P. Tran, and M. B. C. Khoo. Performance of the hotelling t_2 control chart for compositional data in the presence of measurement errors. *Journal of Applied Statistics*, 46(14):2583–2602, 2019.
- [38] F. S. Zaidi, P. Castagliola, K. P. Tran, and M. B. C. Khoo. Performance of the mewma-coda control chart in the presence of measurement errors. *Quality and Reliability Engineering International*, 36(7):2411–2440, 2020.
- [39] F.S. Zaidi, H.L. Dai, M. Imran, and K.P. Tran. Analyzing abnormal pattern of hotelling t_2 control chart for compositional data using artificial neural networks. *Computers & Industrial Engineering*, 180:109254, 2023.
- [40] F.S. Zaidi, H.L. Dai, M. Imran, and K.P. Tran. Monitoring autocorrelated compositional data vectors using an enhanced residuals hotelling t_2 control chart. *Computers & Industrial Engineering*, 181:109280, 2023.
- [41] Z. Zhang and X. Deng. Anomaly detection using improved deep svdd model with data structure preservation. *Pattern Recognition Letters*, 148:1–6, 2021.