Signal Interpretation in Hotelling's T^2 Control Chart for Compositional Data

Marina Vives-Mestres,
Josep Daunis-i-Estadella
and
Josep-Antoni Martín-Fernández
Department of Computer Science, Applied Mathematics and Statistics,
University of Girona

October 28, 2015

Abstract

Nowadays, control of concentrations of elements is of crucial importance in industry. Concentrations are expressed in terms of proportions or percentages which means that they are compositional data (CoDa). CoDa are defined as vectors of positive elements that represent parts of a whole and usually add to a constant sum. Classical T^2 control chart is not appropriate for CoDa, for which is better to use a compositional T^2 control chart (T^2_C CC). This paper generalizes the interpretation of the out-of-control signals of the individual T^2_C CC for more than three components. We propose two methods for identifying the ratio of components that mainly contribute to the signal. The first one is suitable for low dimensional problems and consists on finding the log ratio of components that maximizes the univariate T^2 statistic. The second one is an optimized method for large dimensional problems that simplifies the calculus by transforming the coordinates into the sphere. We illustrate the T^2_C CC signal interpretation with a practical example from the chemical and pharmaceutical industry.

Keywords: Composition, Hotelling's Statistic, Log ratio, Mixture, Multivariate Process Control, Signal Interpretation.

1 INTRODUCTION

One of the most familiar tool for multivariate statistical process control is Hotelling's T^2 control chart (Hotelling 1947). In standard control procedure for individual observations the well known Mahalanobis distance is plotted against time

$$T^2 = (\mathbf{x} - \boldsymbol{\mu})\boldsymbol{\Sigma}^{\mathbf{-1}}(\mathbf{x} - \boldsymbol{\mu})', \tag{1}$$

where $\mathbf{x}$ is a row p-dimensional vector observed at time t and $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are the mean and the variance-covariance matrix respectively. Usually $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are unknown and have to be estimated from a historical data set. It is assumed that $\mathbf{x}$ are mutually independent and multivariate normally distributed.

The T^2 control chart has the main advantage that enables monitoring multiple variables taking into account both the univariate and the interrelationship effects between them (Tracy et al. 1992; Montgomery 2009; Kenett et al. 2014). But it has an important disadvantage: it masks the cause of the out-of-control signals due to the dimensionality reduction from a p-dimensional vector to a unidimensional statistic. Identifying the cause of the anomaly is of crucial importance in order to apply appropriate remedial measures. Many methods can be found in literature for interpreting out-of-control signals in multivariate control charts (Das and Prakash 2008; Tan and Shi 2012 and references therein).

This paper considers the case in which the quality characteristic being monitored is a compositional vector $\mathbf{x} = (x_1, \ldots, x_p)$. Compositions are vectors of positive elements describing quantitatively the parts of some whole and usually adding to a constant sum (for simplicity, often taken to be 1) (Pawlowsky-Glahn and Buccianti 2011). Classical data units are weight or volume percent, ppm, ppb, molarities, or any other concentration units. For this reason compositional data (CoDa) are widely found in industries such as the chemical, pharmaceutical or asphalt and also in gas or water analysis among others. Note that in industry, the term CoDa is commonly referred to as mixture data or mixture composition (e.g., food, pharmaceutical,...). However, the term CoDa has a more general sense and can also refer to non-mixture data, such as data from the use-of-time surveys, household budgets, votes from elections or geochemistry among others.

The sample space of CoDa is the Simplex $\mathcal{S}^p$, a restricted space, where p represents the number of parts in the composition. When $p = 3$ the composition lies in an equilateral triangle in $\mathbb{R}^3$ (on the plane $x_1 + x_2 + x_3 = 1$, perpendicular to the vector $(1, 1, 1)$). It is more common to represent $\mathcal{S}^3$ in the ternary diagram.

Vives-Mestres et al. (2014a) proposed a T^2 control chart suitable for CoDa, denoted T^2_C control chart (T^2_C CC). They demonstrated that applying the T^2 procedure to raw data after deleting one component (hereafter referred as the classical method), is inconsistent with the definition of CoDa. The T^2_C CC is based on a transformation of the data into log ratios of components (called coordinates) that moves the data from restricted ($\mathcal{S}^p$) to non-restricted real space. The authors also furnished a simulation study comparing the average run length (ARL) of the classical and the CoDa approach and show that T^2_C outperforms the T^2 control chart in terms of in-control ARL.

Signal interpretation for the T^2_C CC has been studied for the easiest case of $p = 3$ in Vives-Mestres et al. (2014b). Authors showed that the interpretation of the conditional terms of the MYT decomposition method (Mason et al. 1995, 1997; Mason and Young 2002) in terms of the original components is misleading, and they proposed a method based on selecting the appropriate ilr basis, for each signaling observation, so that the unconditional term is maximum. The maximization function involved in the algorithm depends on the angle from the ilr basis to the abscissa. This feature makes difficult to generalize the algorithm for $p > 3$ because, as the number of dimension increases, also does the number of angles, and the maximization function gets trickier.

The main contribution of this paper is to present two generalized methods for interpreting the T^2_C CC signals for $p \geq 3$. The first one is suitable for low dimensional problems and is based on computing the univariate T^2 statistic on all possible combinations of ratios of components and retaining the maximum one. The second one transforms the coordinates into the sphere, where the maximum logarithm of a product of components is easily identified, and it is approximated by the closest log ratio of components. In both cases the selected log ratio of components is the main contributor to the out-of-control signal of the T^2_C CC.

The rest of this paper is organized as follows. Section 2 reviews the basic concepts for analysing compositional data based on the log-ratio transformation of components and describes the principles of the T^2_C CC. Section 3 illustrates graphically the idea behind the proposed methods for $p = 3$, develops those methods ans provides a performance analysis. An example of application of the proposed methods is presented on Section 4, and the last section is devoted to final remarks.

2 CoDa TREATMENT AND T^2_C CONTROL CHART

Compositions provide information about relative values of components; its total sum is not informative. Therefore every statement about a composition can be stated in terms of ratios of components (Aitchison 1986; Pawlowsky-Glahn and Buccianti 2011).

Aitchison proposed a new methodology based on a log-ratio transformation of components.

Logratios enable representation of CoDa in real space where standard unconstrained multivariate statistics can be applied. Inference therein is translatable back into compositional statements.

We use two main transformations: the centred logratio (clr), and the isometric logratio (ilr), which will be denoted by $\mathbf{z}$ and $\mathbf{y}$ respectively. Hereafter we will refer to transformed data as coordinates. The clr transformation ($\mathbf{z} = \mathrm{clr}(\mathbf{x})$), first proposed by Aitchison (1986), is defined as the logarithm of the ratio of parts over the geometric mean of the composition. The inverse transformation is $\mathrm{clr}^{-1}(\mathbf{z}) = \mathbf{x}$.

The clr coordinates live in $\mathbb{R}^p$ but they lie on a hyperplane. The dimensionality of the clr coordinates can be reduced by representing the data in $\mathbb{R}^{p-1}$ by the use of one of the infinite possible bases lying on the hyperplane. The ilr transformation ($\mathbf{y} = \mathrm{ilr}(\mathbf{x})$), first proposed by Egozcue et al. (2003), allows this representation and provides an orthonormal basis that enhances the interpretability of the data: the expression of the ilr coordinates represents ratios of components.

The ilr coordinates $\mathbf{y}$ and its respective orthonormal basis ($\mathbf{e}_1, \ldots, \mathbf{e}_{p-1}$ in $\mathcal{S}^p$) can be defined through a sequential binary partition (SBP). An example on how to construct a SBP for the case of $p = 3$ is illustrated in Vives-Mestres et al. (2014b). The orthogonal basis of the ilr coordinates in $\mathbb{R}^{p-1}$ is $\boldsymbol{\Psi} = (\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_{p-1})' = (\mathrm{clr}(\mathbf{e}_1), \ldots, \mathrm{clr}(\mathbf{e}_{p-1}))'$. The coordinates of a composition in the basis $\boldsymbol{\Psi}$ are called balances and the compositions of the basis ($\mathbf{e}_i$) are called balancing elements. Each basis element $\boldsymbol{\psi}_i$ is a log-contrast, that is, a linear combination of logarithms of components such as $\log(x_1^{\alpha_1} \cdots \cdot x_p^{\alpha_p})$, where $\sum \alpha_i = 0$, and defines a direction in $\mathbb{R}^p$. Hereafter we will refer to $\boldsymbol{\psi}_i$ as ilr direction. Note that given the composition $\mathbf{x}$, it holds $\mathbf{y} = \mathbf{z} \cdot \boldsymbol{\Psi}'$.

Figure 1 shows all possible ilr directions for $p = 3$ in $\mathbb{R}^2$ ($\boldsymbol{\psi}_1, \boldsymbol{\psi}_2, \ldots, \boldsymbol{\psi}_6$), and their corresponding balancing elements in $\mathcal{S}^3$ ($\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_6$). In the simplex, some directions are represented visually by curves because of the special geometry of the simplex.

The log-ratio methodology does not apply when the composition has some zero value. In our context we consider that these possible zeros are values below a detection limit. According to Palarea-Albaladejo and Martín-Fernández (2015) those elements can be replaced with specific techniques. Palarea-Albaladejo and Martín-Fernández (2013) give a review of these techniques.

The general definition of the T_C^2 CC is stated as follows: given $\mathbf{x} = (x_1, \ldots, x_p)$ a p-part composition and $\mathbf{y} = (y_1, \ldots, y_{p-1})$ its ilr coordinates, the T_C^2 statistic is defined as:

$$T_C^2 = (\mathbf{y} - \boldsymbol{\mu}_y)\boldsymbol{\Sigma}_y^{-1}(\mathbf{y} - \boldsymbol{\mu}_y)', \tag{2}$$

where $\mathbf{y}$ is the ilr coordinate of the observed composition and $\boldsymbol{\mu}_y$ and $\boldsymbol{\Sigma}_y$ are the mean vector and the variance matrix of the ilr coordinates. In practice it is necessary to estimate both values in

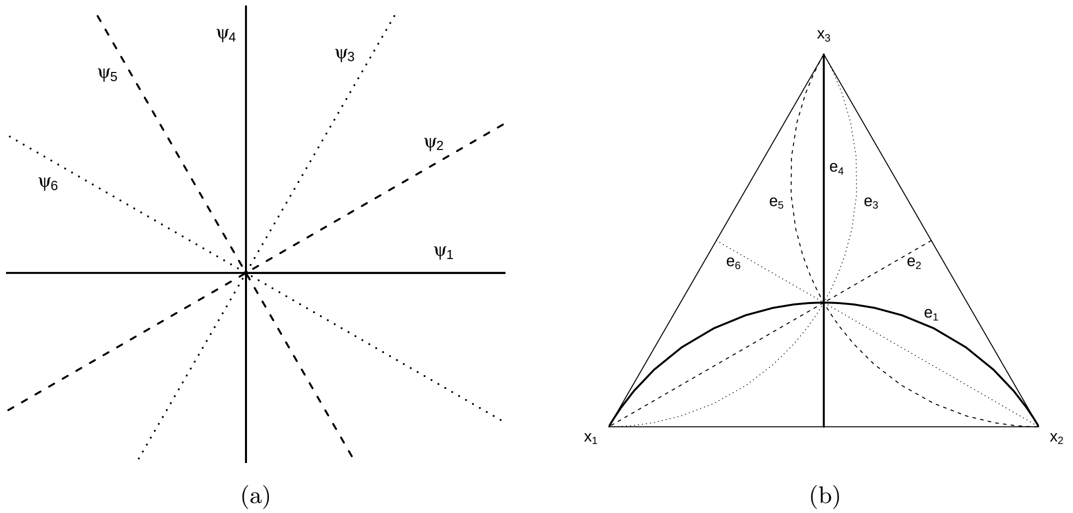


Figure 1: Representation in $\mathbb{R}^2$ (left) and in $\mathcal{S}^3$ (right) of all possible ilr directions defined through a SBP with $p = 3$, where $\boldsymbol{\psi}_i = \mathrm{clr}(\mathbf{e}_i)$.

Phase I as it is done in standard methods. It is assumed that the in-control observation vectors $\mathbf{y}$ are i.i.d. multivariate normal random vectors $\mathcal{N}(\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y)$ with common mean vector and covariance matrix, thus compositions $\mathbf{x}$ follow a normal distribution on the simplex (Mateu-Figueras et al. 2013): $\mathcal{N}_{\mathcal{S}}(\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y)$. In this article we assume that the covariance matrix remains in control.

Note that the T_C^2 statistic (Equation 2) is not affected by the basis used to construct the ilr coordinates. In practice the user would select a balance which is convenient for easy interpretation. The T_C^2 statistic can also be defined in terms of clr coordinates by replacing the $\mathbf{y}$'s from Equation 2 by $\mathbf{z}$'s after deleting one clr-component (it does not matter whichever it is), and replacing the mean and the covariance matrix of the ilr coordinates by the clr ones, after deleting the same previous component.

The T_C^2 is consistent with CoDa definition because: (i) fits better the distribution of the data set and encloses only values from the sample space, and (ii) fulfils the condition of subcompositional coherence: inference about some components must be the same whether the whole composition or a subcomposition is used (Vives-Mestres et al. 2014a, b).

3 INTERPRETATION OF T_C^2 SIGNALS

Our approach is inspired by the MYT decomposition method (Mason et al. 1995, 1997; Mason and Young 2002) that uses an orthogonal transformation to express the T^2 as the sum of two independent terms named unconditional and conditional terms. The unconditional term (T_i^2)

depends only on x_i and is the univariate T^2 statistic of x_i, and the conditional term ($T^2_{i.j}$) depends on the conditional density of x_i given x_j. The decomposition is performed on each signaling observation and conditional and unconditional terms are compared with its limiting values.

However, our method is different from the MYT because we are looking for the ilr coordinate that has the maximum unconditional term, i.e. the closest to the global T^2_C. By decomposing the T^2_C by the use of this coordinate, we obtain the highest weight on the unconditional term. Our interest is on interpreting only the unconditional term, which has a clear interpretation in terms of a log ratio of components responsible of the signal.

Another advantage of our method over the MYT is that no significance level of the decomposition terms is needed: we attribute the cause of the signal mainly to the ilr coordinate that has the maximum unconditional term.

Our method avoids the joint interpretation problem of the terms of the MYT decomposition because our focus is only on the unconditional term. Another problem with the MYT decomposition method is that, for p variables, there are $p!$ possible decompositions. Our method does not compute all possible decompositions, but instead we need to calculate the unconditional terms of all possible ilr coordinates and select the maximum one. To avoid the computational complexity for high dimensional problems we propose the method described in section 3.2 based on a spherizing transformation of the coordinates and a NN search.

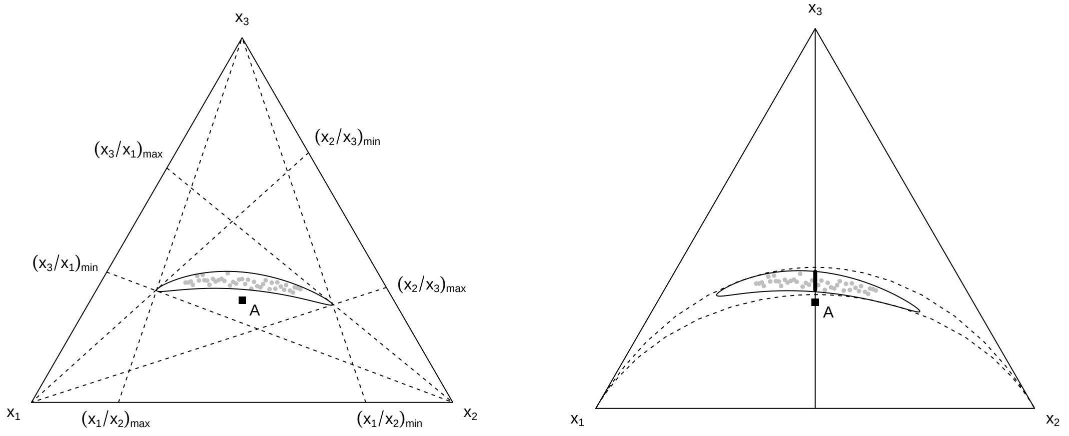
We graphically illustrate the purpose of our method by the use of a simulated data set of 41 observations in $\mathcal{S}^3$ following a normal distribution on the simplex $\mathcal{N}_{\mathcal{S}}(\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y)$ with parameters

$$\boldsymbol{\mu}_y = (0, 0) \quad \boldsymbol{\Sigma}_y = \begin{pmatrix} 0.129 & -0.011 \\ -0.011 & 0.002 \end{pmatrix}$$

Under the assumption of known parameters, the $UCL = \chi^2_{p=2} = 7.81$ with $\alpha = 0.05$. We add an extra observation $\mathbf{A} = (0.36, 0.36, 0.28)$ which is an outlier because $T^2_C = 39.6 > UCL$. The simulated data set and observation $\mathbf{A}$ (■) are represented in Figure 2 together with the control region of the T^2_C CC. It can be clearly seen that observation $\mathbf{A}$ is an outlier. Note that under classical approach, observation $\mathbf{A}$ will not signal because $T^2 = 3.93 < UCL$.

Univariate limits of a ratio of two components are defined by the projection of the control region from a vertex to the opposite edge (Figure 2a). For example, the projection of the control region from the vertex x_2 to the edge x_3x_1 represents the limits of the ratio x_3/x_1, where its maximum value $(x_3/x_1)_{\max}$ is on the side of vertex x_3 and the minimum value $(x_3/x_1)_{\min}$ is on the side of x_1. The same reasoning can be applied to the other limits of ratios of two components.

Observation $\mathbf{A}$ will not signal on any of the unconditional terms defined by the ratio of a pair of



(a) Univariate limits of the ratios of two components. Observation **A** is within all univariate limits.

(b) Univariate limits of $\log \frac{x_3}{\sqrt{x_1 x_2}}$ where observation **A** is the furthest from the projected data set.

Figure 2: Graphical interpretation of unconditional (left) and conditional (right) terms of the MYT decomposition in the ternary diagram ($p = 3$).

components. However, if we consider the univariate limits on the direction $\log \frac{x_3}{\sqrt{x_1 x_2}}$ (Figure 2b), it is easy to see that point **A** is out of these limits. In fact, this is the direction in which the outlier is the furthest from the data set according to the metric of the T^2_C control chart.

Using the ilr coordinates such that $y_1 = \sqrt{\frac{2}{3}} \log \frac{x_3}{\sqrt{x_1 x_2}}$ and $y_2 = \sqrt{\frac{1}{2}} \log \frac{x_1}{x_2}$ will produce a MYT decomposition that will give the higher weight on the unconditional term $T^2_{y_1}$, and the lower to $T^2_{y_2 \cdot y_1}$ that will be easily interpretable. In conclusion, the anomaly on **A** can be attributed to a bad relation between component x_3 with respect to the other two.

From now on, for simplicity, we will note a general ilr direction by $\boldsymbol{\psi}$, without sub-index. We propose two methods but, before applying them, it is necessary to have a list, denoted by $\mathcal{L}$, with all possible ilr directions for the given number of parts p. The size of $\mathcal{L}$ is equal to the number of possible combinations of ratios of components (NC) in any SBP, which is given by

$$NC = \sum_{i=2}^{p} C^i_p (C^{i-1}_i + C^{i-2}_i + \cdots + C^1_i) \qquad \text{for} \quad p \geq 3, \tag{3}$$

where C^n_m is the number of n combinations from a set of m elements.

This list has to be calculated only once for each problem (for each p) and it can be reused many times. The following scheme illustrates a simple procedure to generate this list. It can also be available upon request to the authors or in www.compositionaldata.com.

1. Generate a list with vectors of size p containing all possible combinations of 0, +1 and -1.

2. Delete the combinations that are not a partition: those whose elements are all non-strictly positive or all non-strictly negative. Let r and s be, respectively, the number of +1 and -1 in the combination.

3. For each list element (partition) calculate its balancing element $\boldsymbol{\psi}$ by

$$\boldsymbol{\psi} = (\psi_1, \ldots, \psi_p) \begin{cases} \psi_j = +\sqrt{\frac{s}{r(r+s)}} \\ \psi_k = -\sqrt{\frac{r}{s(r+s)}} \\ \psi_0 = 0 \end{cases}, \tag{4}$$

   where ψ_j is the coefficient for each part coded +1, ψ_k is the coefficient for each part coded -1 and ψ_0 is the coefficient for not involved parts. Note that $\|\boldsymbol{\psi}\| = 1$.

3.1 Computing all unconditional terms

Given an outlier $\mathbf{x} = (x_1, \ldots, x_p)$, and $\mathbf{z}$ its clr coordinates, the unconditional term $T^2_{\boldsymbol{\psi}}(\mathbf{z})$ on one of the ilr directions $\boldsymbol{\psi}$ from the list $\mathcal{L}$ is calculated as

$$T^2_{\boldsymbol{\psi}}(\mathbf{z}) = \frac{(z - \mu_z)^2}{\sigma^2_z}, \tag{5}$$

where z is the projection of the clr coordinates of the outlier $\mathbf{z}$ onto the ilr direction $\boldsymbol{\psi}$ and μ_z and σ^2_z are the mean and the variance, respectively, of the clr coordinates of the historical data set projected on the same direction. These elements can be calculated using the clr coordinates ($\mathbf{z}$) or via the ilr coordinates ($\mathbf{y}$).

$$z = \mathbf{z} \cdot \boldsymbol{\psi}' = \mathbf{y} \cdot \boldsymbol{\Psi} \cdot \boldsymbol{\psi}'$$
$$\mu_z = \boldsymbol{\mu}_z \cdot \boldsymbol{\psi}' = \boldsymbol{\mu}_y \cdot \boldsymbol{\Psi} \cdot \boldsymbol{\psi}'$$
$$\sigma^2_z = \boldsymbol{\psi} \cdot \boldsymbol{\Sigma}_z \cdot \boldsymbol{\psi}' = \boldsymbol{\psi} \cdot \boldsymbol{\Psi}' \cdot \boldsymbol{\Sigma}_y \cdot \boldsymbol{\Psi} \cdot \boldsymbol{\psi}'$$

where $\boldsymbol{\mu}_z$, $\boldsymbol{\mu}_y$ and $\boldsymbol{\Sigma}_z$, $\boldsymbol{\Sigma}_y$ are respectively the mean and the covariance matrix of the clr and ilr coordinates. The matrix $\boldsymbol{\Psi}$ is the orthogonal base of the ilr coordinates ($\mathbf{y}$).

Note that z from Equation 5 is the same whichever is the base used to construct the ilr coordinates, so it is the $T^2_{\boldsymbol{\psi}}(\mathbf{z})$ term, which means that our method is invariant to rotation of the ilr coordinates. The unconditional term in two opposite direction vectors is the same, so the list $\mathcal{L}$

can be reduced to the half by deleting those directions that are opposite. The resulting list $\mathcal{L}_2$ is of dimension $NC/2 \times p$.

After computing the $T^2_{\boldsymbol{\psi}}(\mathbf{z})$ for all elements of the list $\mathcal{L}_2$, the direction in which the $T^2_{\boldsymbol{\psi}}(\mathbf{z})$ is maximum indicates the ratio of components responsible of the signal. To obtain the ratio from a general direction $\boldsymbol{\psi}$, write the components of the positive coefficients in the numerator and the components of the negative coefficients in the denominator. Null coefficients indicate that the component is not involved in the ratio.

When $p = 4$ the number of $T^2_{\boldsymbol{\psi}}(\mathbf{z})$ terms to compute is $NC/2 = 25$, for $p = 7$ it is 966, and for $p = 10$ is 28501. Because $NC/2$ grows exponentially with p, this method is not very efficient for large p. For $p = 10$ the time for computing all unconditional terms in R-3.1.2 (R Core team 2013) is about 13 seconds, for $p = 11$ about 40 seconds and for p=12 about 131 seconds, using a personal computer with a 1. 1GHz Intel Core i7 processor. For $p \geq 11$ we suggest to use the second method described in the following section.

3.2 Spherizing the coordinates

Let a spherical distribution be the one with mean centred at the origin of real space and with covariance matrix equal to the identity matrix. Under this distribution, data are inscribed in a sphere. Spherical or spherized clr coordinates will be noted with the subscript s, and are calculated as follows

$$\mathbf{z}_s = (\mathbf{z} - \boldsymbol{\mu})\boldsymbol{\Sigma}^{-1/2}, \tag{6}$$

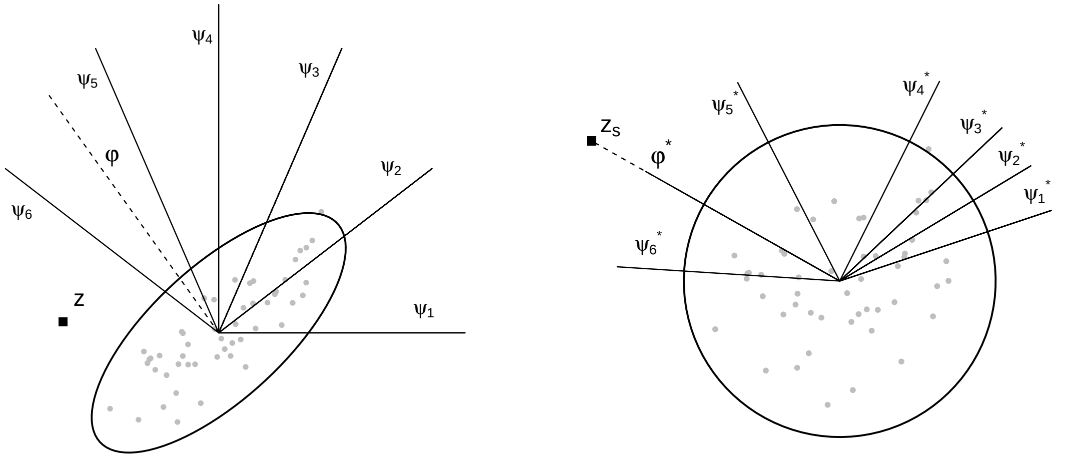
where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are the mean and the variance of the clr coordinates. Note that the global T^2_C for a given observation is the same whether we use original coordinates or spherized coordinates (clr or ilr): $T^2_C(\mathbf{z}) = T^2_C(\mathbf{z}_s) = T^2_C(\mathbf{y}) = T^2_C(\mathbf{y}_s)$. From now on, for simplicity, we will develop the method using clr coordinates ($\mathbf{z}$) although the covariance matrix of Equation 6 is singular. In practice we would use ilr coordinates.

The advantage of working with spherized coordinates is that, following Equation 5, the unconditional term on any direction of the spherized space $\boldsymbol{\psi}_s = \boldsymbol{\psi}\Sigma^{-1/2}$ reduces to

$$T^2_{\boldsymbol{\psi}_s}(\mathbf{z}_s) = (z_s)^2 \tag{7}$$

where z_s is the projection of the clr coordinates of the spherized outlier onto $\boldsymbol{\psi}_s$, so that: $z_s = \mathbf{z}_s \cdot \boldsymbol{\psi}'_s$.

The global T^2_C of the spherized coordinate $\mathbf{z}_s$ is $T^2_C(\mathbf{z}_s) = \|\mathbf{z}_s\|^2$ (Equation 2), and it is equal to



(a) Coordinate space with control region and outlier $\mathbf{z}$ as well as all possible ilr directions $\boldsymbol{\psi}_1, ..., \boldsymbol{\psi}_6$ for $p = 3$.

(b) Spherized coordinate space with circular control region, outlier $\mathbf{z}_s$ and ilr directions transformed such that $\boldsymbol{\psi}_i^* = \boldsymbol{\psi}_i \Sigma^{1/2}$.

Figure 3: Graphical interpretation of the proposed procedure for interpreting out of control signals of the T_C^2 CC when $p \geq 11$.

the unconditional term (Equation 7) when z_s is maximum. The maximum value of z_s is achieved when the clr coordinates of the spherized outlier is projected onto the direction $\boldsymbol{\varphi}^*$ ($\|\boldsymbol{\varphi}^*\| = 1$) pointing from the origin to $\mathbf{z}_s$, that is, when $\boldsymbol{\varphi}^* = \mathbf{z}_s/\|\mathbf{z}_s\|$ it holds $T_C^2(\mathbf{z}_s) = (\frac{\mathbf{z}_s \cdot \mathbf{z}_s'}{\|\mathbf{z}_s\|})^2 = T^2_{\boldsymbol{\varphi}^*}(\mathbf{z}_s)$ (see demonstration in Appendix A).

Once $\boldsymbol{\varphi}^*$ is found , it is transformed back to the non spherized space ($\boldsymbol{\varphi}$) to interpret it. The transformation is done by $\boldsymbol{\varphi} = \boldsymbol{\varphi}^* \boldsymbol{\Sigma}^{-1/2}$ because it assures that both the unconditional term on the spherized and non spherized space are the same. It can be proved by developing Equation 5.

$$T^2_{\boldsymbol{\varphi}}(\mathbf{z}) = \frac{((\mathbf{z} - \boldsymbol{\mu})\boldsymbol{\varphi}')^2}{\boldsymbol{\varphi}\boldsymbol{\Sigma}\boldsymbol{\varphi}'} = \frac{(\mathbf{z}_s \boldsymbol{\Sigma}^{1/2}\boldsymbol{\varphi}')^2}{\boldsymbol{\varphi}\boldsymbol{\Sigma}\boldsymbol{\varphi}'} = \frac{(\mathbf{z}_s {\boldsymbol{\varphi}^*}')^2}{\boldsymbol{\varphi}^* {\boldsymbol{\varphi}^*}'} = T^2_{\boldsymbol{\varphi}^*}(\mathbf{z}_s) \tag{8}$$

Figure 3 shows the procedure described up to here. Figure 3a shows the original coordinate space together with the control region and the outlier $\mathbf{z}$ and Figure 3b shows the spherized coordinate space with the circular control region and the spherized outlier $\mathbf{z}_s$. The direction $\boldsymbol{\varphi}^*$ pointed by $\mathbf{z}_s$ (dashed line on Figure 3b) is transformed back to the original coordinate space $\boldsymbol{\varphi} = \boldsymbol{\varphi}^* \boldsymbol{\Sigma}^{-1/2}$ and results on the dashed line of Figure 3a.

The direction $\boldsymbol{\varphi}$ is a log-contrast and indicates the cause of the out of control signal. A general log-contrast is not easily interpretable in terms of original components, thus we propose to approximate $\boldsymbol{\varphi}$ by an ilr direction (representing a balance). The selected ilr direction has to have the property that the unconditional term on this ilr direction is maximum. To fulfill this requirement, the approximation has to be done in the spherized coordinate space, because there the unconditional term is simply the projection of $\mathbf{z}_s$ on a given direction (Equation 7).

To perform the approximation, the ilr directions $\boldsymbol{\psi}$ from the list $\mathcal{L}$ have to be transformed into the spherized space such that $\boldsymbol{\psi}^* = \boldsymbol{\psi}\Sigma^{1/2}$. Then we propose to use a nearest neighbour (NN) search algorithm to find the closest (in terms of angle) direction $\boldsymbol{\psi}^*$ to $\boldsymbol{\varphi}^*$, which is equivalent to find the closest direction (in terms of euclidean distance) of the normalized set $\boldsymbol{\psi}^*$ to $\boldsymbol{\varphi}^*$

Our aim here is not to provide a deep discussion on the NN search. We have used a kd-tree from the library "Approximate Nearest Neighbor Searching" in `www.cs.umd.edu/~mount/ANN/` with the option of exact nearest neighbour search implemented in C++ and also available under R (R Core team 2013). The computing time of the NN search for $p = 12$ is 0.55 seconds, using a personal computer with a 1. 1GHz Intel Core i7 processor.

In the example of Figure 3, $\boldsymbol{\varphi}^*$ would be approximated by $\boldsymbol{\psi}_6^*$. Note that, on the coordinate space, $\boldsymbol{\psi}_6$ does not corresponds to the closest direction to $\boldsymbol{\varphi}$ because following Equation 5, the unconditional term also implies the variance.

To summarize, the procedure to interpret the cause of the out-of-control signal in a T_C^2 CC when $p \geq 11$ stands as follows:

1. Compute the mean and the variance of the clr coordinates ($\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$).

2. Compute the clr coordinates of the atypical observation: $\mathbf{z}$.

3. Calculate the spherized clr coordinates of the outlier: $\mathbf{z}_s = (\mathbf{z} - \boldsymbol{\mu}) \cdot \boldsymbol{\Sigma}^{-1/2}$

4. Apply a NN search algorithm to find the closest ilr direction $\boldsymbol{\psi}^*$ to $\mathbf{z}_s$ in the normalized set $\mathcal{L}^* = \frac{\mathcal{L}\boldsymbol{\Sigma}^{1/2}}{\mathcal{L}\boldsymbol{\Sigma}\mathcal{L}'}$. The log ratio represented by this ilr direction is the responsible of the out-of-control signal.

Repeat steps 2 to 4 for each signaling observation. Note that given a dimension p, the number of points to query again $\mathbf{z}_s$ is NC, defined in Equation 3. The set $\mathcal{L}^*$ has to be calculated for each particular process and updated at each time the covariance matrix of the process changes.

3.3 Procedure performance

The two previously presented procedures always find the sample ilr coordinate that shifted most, that is, the ilr coordinate (ψ) that has the larger $T_{\boldsymbol{\psi}}^2(\mathbf{z})$. It can easily be seen in the method of Section 3.1 because all $T_{\boldsymbol{\psi}}^2(\mathbf{z})$ values are checked and the maximum one is retained for each outlier.

The second method (Section 3.2) also finds the ilr coordinate that shifted most, as it has been demonstrated mathematically. We have indeed checked that both methods give the same results when applied to a given problem.

However, there is a need to show how the methods can help on identifying the population raw components that shifted, and its performance. We suggest to check not only the ilr (ψ) that shifted most but also the following k ilr directions (i.e. $k = 5, 10$): in the first method it can be easily done by ordering the ilr directions by their $T^2_{\boldsymbol{\psi}}(\mathbf{z})$ and in the second one by looking for a subset of k nearest neighbours. The components shared by the first k ilr directions are indicators of the main responsible of the signal. To check the direction of the shift we suggest to compare the outlier with the geometric mean of the raw data.

To show the performance of the methods, we use a simulation example, inspired in Section 6.1 of Tan and Shi (2012), with $p = 12$. The known in control mean of the ilr coordinates taking the default ilr base of the R package 'compositions' (van der Boogart et al 2014) is $\boldsymbol{\mu}_0 = \mathbf{0}$ and the known covariance matrix $\boldsymbol{\Sigma}$ is taken from Appendix B of Tan and Shi (2012).

Different out-of-control means $\boldsymbol{\mu}_i$ are defined on the original composition so that

$$\mathrm{ilr}^{-1}(\boldsymbol{\mu}_i) = (\underbrace{\delta_i\frac{1}{p}, \ldots, \delta_i\frac{1}{p}}_{i \text{ elements}}, \frac{1}{p}, \ldots, \frac{1}{p})$$

where δ_i corresponds to an increase of $i\Delta$ times the variance of the log ratio of the ith components against the other ones. For example, when $i = 2$ we shift to $\boldsymbol{\mu}_2$ which in terms of the original compositions it is equivalent to perturb the two first raw parts, for $i = 3$ it is a perturbation of the first three parts, etc. Following Tan and Shi (2012) we change $\Delta = 0.6, 1, 1.4$ for each $\boldsymbol{\mu}_i$, $i = 1, \ldots, 4$ according to the experimental design of Table 1 and we replicate each run in the design 100 times. For each replicate, we simulate an out of control signal from a $N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$ and apply the spherizing method.

Note that when there is a shift in one component, due to the constant sum restriction of CoDa, there is also a necessary change in some of the other components, which means that the responsible log ratio of the signal will include the shifted component together with other ones. Similarly, when we perturb $i > 1$ components with the same amount (δ_i), it does not necessary mean that the i components are responsible of the shift because it depends on the correlation structure between the ratios of the shifted components against the others as well as the correlation structure within the ratios of the shifted components. Based on these notes, we define three performance indicators as follows:

- **P1:** Cases in which the i shifted components appears in the maximum $T^2_{\boldsymbol{\psi}}(\mathbf{z})$.
- **P2:** Cases in which at least one of the shifted components appears in all the $k = 5$ highest $T^2_{\boldsymbol{\psi}}(\mathbf{z})$.
- **P3:** Minimum of the positions in which only non shifted components appear in the ordered list of $T^2_{\boldsymbol{\psi}}(\mathbf{z})$.

The performance indicator P3 is a kind of Type I error because it looks for responsible log ratios including only in control means (non shifted components) which means that they are incorrectly identified as out of control. The results of the simulation study are presented on Table 1.

Shifted mean	Shift Δ	P1 (%)	P2 (%)	P3
$\boldsymbol{\mu}_1$	0.6	100	98	2
	1.0	100	100	21
	1.4	100	100	644
$\boldsymbol{\mu}_2$	0.6	92	97	90
	1.0	100	100	2434
	1.4	100	100	> 4000
$\boldsymbol{\mu}_3$	0.6	67	92	1957
	1.0	74	100	> 4000
	1.4	77	100	> 4000
$\boldsymbol{\mu}_4$	0.6	62	79	161
	1.0	65	96	2198
	1.4	66	97	> 4000

Table 1: Performance indicators of the proposed procedure.

From Table 1 it can be seen that when there is a shift in only one component the proposed procedure clearly identifies that the problem is with that component, even checking only for the maximum $T^2_{\boldsymbol{\psi}}(\mathbf{z})$. When the size of the shift is small ($\Delta = 0.6$) in $\boldsymbol{\mu}_1$, the minimum position where the signal is attributed to only non shifted components is 2. However, this is the worst case, and along the 100 replications of this experimental design we obtain a median of the position of 404.5, that is, only in 2% of cases a log ratio including only non shifted components appears in one of the five highest $T^2_{\boldsymbol{\psi}}(\mathbf{z})$.

For $i > 1$ the percentage of cases in which all shifted components appear in the first responsible log ratio decreases, but by checking the $k = 5$ largest $T^2_{\psi}(\mathbf{z})$ we can get to the cause of the signal in a high percentage of cases. For all shifted means, as it would be expected, the performances increases as it does the shift size Δ. The minimum position where the signal is attributed to non shifted components is higher as the shift size increases and for $i > 1$ is located far enough so that the user will not check them to attribute the cause of the signal.

4 EXAMPLE

We analyse in this section an example of industrial application using the data from Gonzalez-de la Parra and Rodriguez-Loaiza (2003). The data describe the impurity profile of seven major organic impurities (denoted A to G) of a crystalline drug substance. The impurity profile, or impurities, are directly related to the chemical and physical method of manufacture thus is an important quality characteristic of the product.

The level of each impurity is reported in ppm, so it is a compositional data set not adding to a constant sum because not all components are measured. The authors provide a historical data set (HDS) of 30 observations and an evaluation data set (EDS) of 167 observations. Both data sets are reproduced in Table B1 and Table B2 respectively in Appendix B.

A clear advantage of the CoDa method over the classical one is that the former enables monitoring the proportion between components and check they keep within the expected range without taking into account if the total amount of components is high or low. In this example we assume that the total amount of impurities is not an important issue because is neither analysed not mentioned in the original article (Gonzalez-de la Parra and Rodriguez-Loaiza 2003).

Firstly, we applied the compositional T^2_C control chart to the HDS to determine the in-control state of the process and to identify a reference sample, which is known as Phase I. Secondly, we used the estimates computed from the reference sample to define the control limits to which the EDS was compared during Phase II.

Phase I The 30 observations from preliminary data set were considered to represent the impurity profile under the best monitored manufacturing operating conditions. Prior to consider them as the HDS, the authors Gonzalez-de la Parra and Rodriguez-Loaiza (2003) evaluated the raw data set for the presence of outliers, autocorrelation and multicollinearity. There were no univariate outliers but a significant first-order autoregressive model for impurities C and F was found, so the lag-one variables of this impurities, denoted by C_{t-1} and F_{t-1}, were added to the HDS. No multicollinearity problem was found, so the HDS composed of the raw data together with C_{t-1} and F_{t-1}, were used as a baseline to evaluate new observations using the classical method.

We use a compositional T_C^2 control chart by means of the clr transformation of the data, the impurity levels of each lot $\mathbf{x} = (\mathrm{A}, \mathrm{B}, \mathrm{C}, \mathrm{D}, \mathrm{E}, \mathrm{F}, \mathrm{G})$ were transformed into $\mathbf{z} = \mathrm{clr}(\mathbf{x}) = (z_{\mathrm{A}}, z_{\mathrm{B}}, z_{\mathrm{C}}, z_{\mathrm{D}}, z_{\mathrm{E}}, z_{\mathrm{F}}, z_{\mathrm{G}})$.

There was one out of control observation in the preliminary data set. The Hotelling's statistic of the coordinates of lot 20 is $T_C^2 = 17.58$ which is higher than the control limit UCL=16.70 obtained from the beta distribution with $p = 6$, $n = 30$ and $\alpha = 0.001$. Using univariate techniques we can verify that the value of z_{F} of observation 20 (-2,52) is at more than 3 standard deviations from the mean. Lot 20 was removed from the preliminary data set and not included in the HDS.

We used the method described in Section 3.1 to identify the cause of the anomaly in lot 20 because $p < 11$. The decomposition of the T_C^2 was computed on all ilr directions ($NC/2 = 966$) using Equation 5. The direction in which the T_{ψ}^2 was maximum is the one defined by the ratio $\propto \log \frac{\mathrm{F}}{(\mathrm{ABCE})^{1/4}}$. The decomposition on this direction equals to 16, which is a similar value to the global T_C^2.

Raw data on Table B1 shows that, in lot 20, the quantity of impurity F is very low while there are a lot of impurities A and B. If we look at the univariate values of the log ratio $\sqrt{\frac{4}{5}} \log \frac{\mathrm{F}}{(\mathrm{ABCE})^{1/4}}$, we see that the mean and the standard deviation calculated on the HDS are -0.14 and 0.56, respectively. On lot 20, the value of this log ratio is 2.11, which is at more than 3σ from the mean.

No significant autocorrelation was found on the clr coordinates and neither a strong relationship among the predictor coordinates. Under this conditions we set the Phase I T_C^2 control chart with critical value UCL=16.52 ($p = 6$, $n = 29$ and $\alpha = 0.001$). The T_C^2 control chart of the HDS is drawn in Figure 4 and do not present any out of control signal (once removed lot 20).

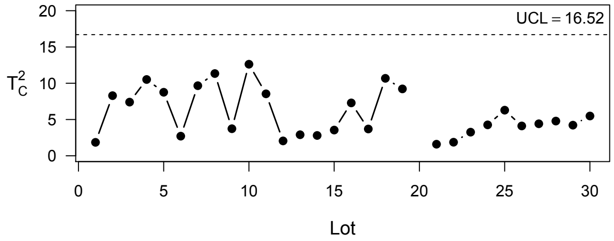


Figure 4: Compositional T_C^2 control chart for the historical data set from Gonzalez-de la Parra and Rodriguez-Loaiza (2003). Observation 20 was removed because it is an outlier.

Phase II The estimates of the mean vector and the covariance matrix of the clr-transformed HDS, were used to control the process in Phase II. The limit for testing new observations was calculated using the Fisher distribution (Tracy et al. 1992): UCL=42.68 for $\alpha = 0.001$. We used the clr-transformed data of the EDS of the impurity levels from Table B2 to plot the T_C^2 control chart.

The EDS contains a zero in lot 116: impurity B is not present. We believed that this zero does not correspond to an absolute absence of this impurity, but corresponds to a low value below the detection limit, which means that this is a rounded zero. Palarea-Albaladejo et al. (2014) proposed to use a multiplicative replacement when the number of these null values is less than 10% and replace them by 2/3 of the threshold value. From the raw data we see that the minimum value in all impurities is 10 ppm, which we considered to be the detection limit. The missing value of impurity B in lot 116 was replaced by 20/3.

Figure 5 shows the compositional T_C^2 control chart for the EDS together with the classical T^2 as described in Gonzalez-de la Parra and Rodriguez-Loaiza (2003). There are 22 lots above the control limit in the T_C^2 control chart whereas there are 50 in the classical T^2: 20 lots are found as out-of-control signals under both approaches, 2 only appear in the T_C^2 and 30 only in T^2.

Many out of control signals appear at the beginning of the evaluation period, while at the end, the process returns to the baseline conditions specified by the HDS. This is because the quality assurance personnel looked for the reasons of the bad performance and found two factors contributing to the variability, which were fixed during the period of manufacture corresponding to lots 135 through 167. The vertical dashed line from Figure 5 indicates lot 135, from which the manufacturing conditions where improved.

The out-of-control signals found using the classical method are assigned to three instability periods (Gonzalez-de la Parra and Rodriguez-Loaiza 2003): period I is comprised of signals located in lots 20 through 47, period II from 63 to 80 and period III from 93 to 120. The three periods are shaded in Figure 5. Three isolated signals are not grouped within those periods.

Figure 5 shows that the out-of-control signals of the compositional T_C^2 control chart are present only in periods I and III. There is only one lot in period II and two lots that cannot be grouped. Table 2 compile the information of the 22 signaling lots of the T_C^2 control chart. The third and four columns are the T^2 statistic under the classical approach ($\text{UCL}_{T^2} = 68.29$) and the cause of the anomaly using the MYT decomposition method. Columns five to seven give the T_C^2 statistic, the maximum value of the statistic projected into an ilr direction (T_{ψ}^2) and the ratio represented by this direction. We have used the method described in Section 3.1 to identify the causes of the

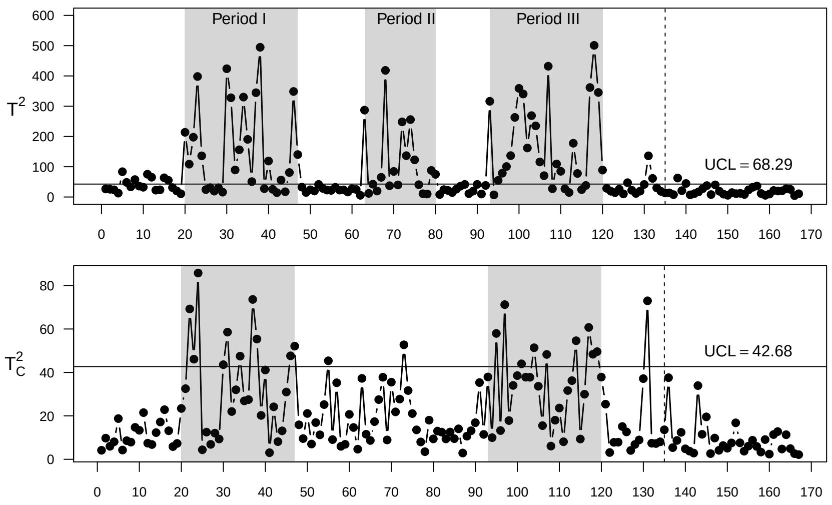


Figure 5: Hotelling's T^2 control chart (up) and compositional T_C^2 control chart (down) for the evaluation data set from Gonzalez-de la Parra and Rodriguez-Loaiza (2003). Three instability periods are found under the classical approach (Periods I, II and III) while there are only two under the CoDa approach.

signaling lots because $p < 11$.

From Table 2 we conclude that most signaling lots have in common a high level of impurities A and C, and a low level of impurities D and F. When impurity B appears in the responsible ratio, this is because it has a high level, and when impurity G appears in the ratio, it is because of a low level. Lot 24, for example, has a slightly different signaling ratios: it has the lowest value of impurity G while impurities A, B, C and E are high: all lie in the higher quartile (Q4). Similarly, lot 38 has values of impurities A, B, C, and E on the Q4 while the quantity of impurities D and G are low.

As stated before, the CoDa method enables monitoring that the ratio between components remains in control while the classical method focuses on the absolute quantities of each component. Lot 68 is a good example to show the difference between both approaches. The Hotelling's statistic under classical approach of this lot is $T^2 = 418.33$. By the use of the MYT decomposition method, the signal is attributed to a high level of impurities A and C. Indeed, the level of impurity A is 170, which is the third maximum value. This observation does not appear as an out of control signal in the T_C^2 control chart because it also have high levels of all other impurities: all appear to be in Q4 except from impurity D which is located in Q2. That means that, in this lot, the total amount of impurities is high, but the relationship between them is within the expected range.

Finally we discuss the case of lot 107 because under classical approach the cause of the anomaly

Period	Lot	T^2	Cause of T^2	T_C^2	T_ψ^2	Responsible ratio
I	22	197.60	A	69.23	63.35	A/F
I	23	398.12	A	46.11	40.03	AC/DF
I	24	135.93	A(BG)(CF$_{t-1}$)(GB)	85.78	69.23	ABCE/G
I	30	423.85	A	43.57	35.17	AC/DF
I	31	327.85	A	58.56	43.36	AC/DF
I	34	330.06	A	47.50	43.06	AC/DF
I	37	344.58	A	73.62	53.96	AC/DF
I	38	494.68	A	55.37	43.10	ABCE/DFG
I	46	348.66	A	47.65	32.69	AC/DF
I	47	140.19	A	52.12	38.56	AC/DF
-	55	21.86	-	45.37	32.63	AC/F
II	73	136.56	A	52.73	35.57	ABC/FG
III	95	54.94	-	57.97	39.59	ACG/DF
III	97	100.71	F(CF$_{t-1}$)(C$_{t-1}$F)(C$_{t-1}$F$_{t-1}$)	71.23	52.79	AC/DF
III	101	340.34	A	43.97	31.80	ABC/FG
III	104	235.01	AB	51.38	35.98	ABC/FG
III	107	432.04	G	48.29	31.00	AC/DF
III	114	77.74	A	54.60	39.88	ACG/DF
III	117	361.77	A	60.73	44.31	AC/DF
III	118	501.05	A	48.39	38.86	AC/DF
III	119	345.39	A	49.58	36.17	AC/DF
-	131	136.11	C	72.98	69.29	ACG/DF

Table 2: Signal interpretation of the 22 signaling observations of the T_C^2 control chart. The left hand side shows the information of the analysis using the classical method, and the right hand side the analysis using the CoDa method.

is attributed mainly to impurity G (although it also appears as an out of control observation in the unconditional terms of impurities A, B and C), while under CoDa approach impurity G does not appear in the ratio. Indeed, lot 107 has the highest value of impurity G, but also has high levels of A, B, C and F. In fact, this is the lot with more impurities: the sum of the impurities in lot 107 is 3440 ppm while on average there are 1908 ppm (or better we would say that the geometric mean of the total sum of impurities is 1853 ppm). The problem is not on the level of impurity G but on the log ratio between AC and DF.

To sum up, main attention has to be paid on the process to achieve a good relationship between the components A and C over D and F. We suggest to analyse the process in order to identify which modifications to the method of manufacture imply a bad relation between the impurities A and C over D and F on the final product.

5 FINAL REMARKS

The T_C^2 CC is suitable for monitoring a process in which the monitored quality characteristic is a composition, that is, a vector of components representing parts of a whole. In that case, it is necessary to transform data from the restricted sample space to the real space by the use of log ratios (coordinates): both the ilr and the clr coordinates can be used. The T_C^2 statistic is defined by the distance from each coordinate to the centre of the coordinates (geometric mean of the composition) by taking into account the correlation among them.

Two methods for identification of the main cause of the individual T_C^2 out-of-control signals have been proposed. The first method is based on computing the univariate T^2 statistic (unconditional term) of all ilr directions, that is, of all combinations of ratios of components: the largest unconditional term indicates the ratio responsible of the anomaly.

The second method transforms the coordinates onto the sphere such that the mean is centred at the origin and the covariance matrix is equal to the identity matrix. In that case, the maximum unconditional term equals the global T_C^2 when it is computed on the direction going from the origin to the spherized outlier. Interpreting this direction in terms of ratio of components may not be easy, thus it is approximated by the nearest log ratio using a nearest-neigbour (NN) search algorithm.

Both approaches provide the closest univariate T^2 statistic of a log ratio to the global T_C^2 such that the decomposition using this ratio will give the highest weight to the unconditional term. The first method is more intuitive and requires a single generation of the list of all possible log ratios that can be reused for other problems of the same dimension (p). As p increases, this method performs slower because the number of combinations of log ratios is hight. This is why we suggest to use the second method for $p \geq 11$. However, the second method requires to update the list of all possible log ratios for each problem, that is, each time that the covariance matrix of the process changes. So we suggest to use the first method for low dimensional problems ($p < 11$).

More current signal interpretation methods, for example, based on Bayesian approaches like the one presented in Tan and Shi (2012), may be of interest for the case of the T^2_C control chart. We leave this improvement for a further development.

The T^2_C control chart is useful to detect out-of-control ratios of components. In some applications, the quality characteristic of the process is a composition, but not all elements of the sample are measured, that is, the vector of components do not add to a constant. In those cases, the vector of components is a subcomposition (also known as a not closed composition). If there is an interest not only on the ratio of components but also on the total amount, then it may be necessary to include in the composition the remainder or include a new variable with the total.

Acknowledgments

This work has been partially financed by the Ministerio de Ciencia e Innovación (Ref: MTM2012-33236) and the Agència de Gestió d'Ajuts Universitaris i de Recerca (AGAUR), Generalitat de Catalunya (Ref: 2014SGR551). The authors appreciate the comments from the referees that helped improve the article.

Appendix A

Let $\mathbf{z}$ be a point in $\mathbf{R}^p$. Consider a normal distribution with mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\Sigma}$. The direction $\boldsymbol{\varphi}$ in $\mathbf{R}^p$ is the one in which the decomposition of the T^2 of point $\mathbf{z}$ is maximum, that is $T^2(\mathbf{z}) = T^2_{\boldsymbol{\varphi}}(\mathbf{z})$.

The T^2 of the projection of point $\mathbf{z}$ on the direction $\boldsymbol{\varphi}$ where $\|\boldsymbol{\varphi}\| = 1$ is (from Equation 5)

$$T^2_{\boldsymbol{\varphi}}(\mathbf{z}) = \frac{(z_i - \mu_i)^2}{\sigma_i^2} = \frac{((\mathbf{z} - \boldsymbol{\mu})\boldsymbol{\varphi}')^2}{\boldsymbol{\varphi}\boldsymbol{\Sigma}\boldsymbol{\varphi}'}$$

The point $\mathbf{z}$ can be spherized by computing $\mathbf{z}_s = (\mathbf{z} - \boldsymbol{\mu})\boldsymbol{\Sigma}^{-1/2}$. The same projection of the spherized point $\mathbf{z}_s$ is calculated taking into account that the mean of the spherized distribution is a vector of zeros and the variance is the identity matrix

$$T^2_{\boldsymbol{\varphi}}(\mathbf{z}_s) = \frac{(\mathbf{z}_s\boldsymbol{\varphi}')^2}{\boldsymbol{\varphi}\boldsymbol{\varphi}'}$$

Both expressions are equivalent if we consider the linear transformation $h(\boldsymbol{\varphi}) = \boldsymbol{\varphi}^* = \boldsymbol{\Sigma}^{1/2}\boldsymbol{\varphi}$. We are looking for the maximum of

$$T^2_{\boldsymbol{\varphi}}(\mathbf{z}) = T^2_{\boldsymbol{\varphi}^*}(\mathbf{z}_s) = \frac{\left(\mathbf{z}_s\boldsymbol{\varphi}^{*'}\right)^2}{\boldsymbol{\varphi}^*\boldsymbol{\varphi}^{*'}} = T^2_{h(\boldsymbol{\varphi})}(\mathbf{z}_s)$$

Applying the chain rule

$$\frac{\partial T^2_{\boldsymbol{\varphi}}(\mathbf{z})}{\partial \boldsymbol{\varphi}_k} = D_k(T^2_{\boldsymbol{\varphi}}(\mathbf{z})) = D_k(T^2_{h(\boldsymbol{\varphi})}(\mathbf{z}_s)) = DT^2_{h(\boldsymbol{\varphi})} \cdot D_k(h(\boldsymbol{\varphi}))$$

The term $D_k(h(\boldsymbol{\varphi}))$ is a constant. The first term has to be equal to zero.

$$DT^2_{\boldsymbol{\varphi}^*}(\mathbf{z}_s) = \frac{(\boldsymbol{\varphi}^*\boldsymbol{\varphi}^{*'})\mathbf{z}_s - (\mathbf{z}_s\boldsymbol{\varphi}^{*'})\boldsymbol{\varphi}^*}{(\boldsymbol{\varphi}^*\boldsymbol{\varphi}^{*'})^2}$$

The numerator is equal to zero only in the case in which $\mathbf{z}_s = \boldsymbol{\varphi}^*$ thus $T^2(\mathbf{z}) = T^2_{\boldsymbol{\varphi}}(\mathbf{z}) = T^2_{\boldsymbol{\varphi}^*}(\mathbf{z}_s)$ in the direction pointed by the spherized point $\mathbf{z}_s$.

The direction $\boldsymbol{\varphi}^*$ in $\mathbf{R}^p$ is the one in which the decomposition of the T^2 of point $\mathbf{z}_s$ is maximum, that is $T^2(z_s) = T^2_{\boldsymbol{\varphi}^*}(z_s)$.

Appendix B

Table B1: Historical data set of the impurity levels (ppm) from Gonzalez-de la Parra and Rodriguez-Loaiza (2003).

Lot	A	B	C	D	E	F	G
1	30	140	550	790	350	170	1110
2	30	10	420	500	220	200	730
3	20	210	540	380	210	180	900
4	10	170	550	390	260	190	580
5	30	10	540	660	190	60	610
6	20	10	230	670	160	70	380
7	30	20	200	840	210	140	280
8	20	40	370	500	20	100	1210
9	40	40	230	800	160	100	480
10	10	30	390	690	10	90	340
11	10	10	440	660	240	60	510
12	20	110	410	960	90	110	780
13	20	210	520	810	280	120	1200
14	20	20	450	520	70	120	640
15	10	20	310	500	110	100	520
16	10	180	520	510	20	110	620
17	20	110	360	560	50	60	730
18	10	50	190	570	60	20	710
19	10	160	200	770	320	70	1190
20	40	140	160	610	140	10	620
21	20	40	320	730	70	70	480
22	40	20	360	700	130	120	640
23	30	10	280	810	120	70	410
24	40	30	310	610	90	70	480
25	20	60	150	590	320	110	900
26	20	140	230	460	220	130	1000
27	30	60	260	460	40	100	740
28	30	100	220	520	50	120	510
29	40	50	360	590	80	110	440
30	20	20	180	610	240	110	780

Table B2: Evaluation data set of the impurity levels (ppm) from Gonzalez-de la Parra and Rodriguez-Loaiza (2003).

Lot	A	B	C	D	E	F	G
1	30	240	580	580	100	190	1310
2	10	130	710	730	260	260	1120
3	40	110	750	610	130	160	800
4	20	20	430	550	210	120	270
5	20	260	1090	360	130	190	620
6	10	50	500	520	80	110	830
7	20	70	760	490	100	80	920
8	20	100	570	400	60	130	1300
9	30	310	720	480	130	110	690
10	10	240	670	530	110	50	1080
11	30	230	680	410	130	60	1190
12	40	50	570	540	200	120	1230
13	30	180	710	710	80	90	980
14	10	10	280	470	20	20	530
15	40	90	480	310	150	90	620
16	30	90	620	250	240	100	660

17	30	30	470	370	230	80	370
18	40	10	440	520	150	130	490
19	40	50	470	450	70	110	470
20	140	100	400	540	250	140	830
21	100	110	450	310	160	110	720
22	140	10	130	400	90	10	480
23	160	60	490	380	170	70	940
24	70	210	730	360	270	120	100
25	30	40	250	440	120	140	360
26	50	290	370	400	170	210	900
27	10	10	300	350	110	100	550
28	40	40	400	310	210	180	830
29	10	30	430	370	130	40	670
30	160	170	670	440	180	100	1130
31	140	70	620	250	180	80	800
32	70	90	470	320	60	110	770
33	110	40	610	430	100	80	490
34	140	10	710	380	90	60	740
35	110	130	430	400	60	140	490
36	70	180	370	290	70	100	500
37	150	260	720	320	80	50	700
38	170	270	690	410	230	140	460
39	20	200	620	470	240	50	740
40	90	10	190	370	80	20	580
41	40	20	310	550	140	130	600
42	10	10	250	540	50	10	630
43	30	30	520	370	60	90	500
44	10	150	350	380	90	30	550
45	40	170	620	270	100	70	530
46	140	210	780	340	50	120	640
47	90	200	430	270	140	60	420
48	20	180	350	250	50	80	290
49	30	110	310	340	90	90	400
50	20	200	270	360	120	50	250
51	10	30	350	390	100	140	290
52	40	350	320	480	80	60	730
53	10	10	430	290	120	130	550
54	30	30	380	380	230	30	360
55	40	30	240	540	110	10	330
56	40	50	260	410	90	60	450
57	30	180	290	350	120	20	740
58	10	190	220	410	70	110	430
59	10	230	190	440	100	70	1140
60	40	110	200	380	290	40	730
61	20	10	340	450	40	70	1120
62	30	20	180	530	40	80	520
63	160	10	400	390	10	140	490
64	20	30	440	380	30	50	920
65	40	60	580	510	20	140	670
66	40	80	580	500	10	60	550
67	60	190	80	540	30	90	710
68	160	160	890	420	120	180	1180
69	40	10	360	480	40	90	660
70	80	230	400	310	10	80	560
71	50	10	490	490	120	40	730
72	120	30	660	460	30	110	790
73	80	310	770	470	130	50	540
74	110	150	730	430	50	110	840
75	70	320	780	470	70	140	1160

76	40	60	630	360	30	110	560
77	20	20	440	610	20	40	450
78	10	40	380	600	40	80	540
79	20	80	540	370	40	330	1400
80	40	40	350	550	30	50	570
81	10	30	360	650	10	90	280
82	40	10	490	510	40	110	910
83	40	10	600	700	40	100	830
84	20	30	350	670	10	140	330
85	10	20	750	550	110	120	770
86	10	10	840	740	80	80	800
87	30	30	190	550	120	110	610
88	40	20	120	690	110	90	260
89	10	20	400	510	70	30	990
90	10	10	290	1320	100	60	770
91	20	40	240	520	130	10	900
92	20	30	340	310	70	40	730
93	130	10	320	490	740	60	350
94	10	40	240	530	60	20	570
95	30	30	400	320	160	10	600
96	10	30	790	510	110	50	630
97	40	20	710	440	140	10	610
98	30	60	830	330	80	90	780
99	130	70	490	360	120	110	580
100	170	130	600	500	140	110	660
101	160	280	730	620	60	120	440
102	90	330	810	470	50	90	610
103	120	340	820	580	90	110	640
104	110	380	790	540	120	100	370
105	60	330	840	500	40	120	350
106	50	80	700	450	140	100	860
107	150	330	820	360	30	140	1610
108	10	80	560	620	90	70	1130
109	70	150	710	470	160	140	980
110	80	70	610	530	140	70	790
111	10	140	190	450	80	30	470
112	30	10	170	510	140	10	500
113	120	80	360	560	10	50	840
114	70	100	280	260	10	20	670
115	10	100	190	320	30	50	1060
116	10	0	560	440	10	130	900
117	170	200	450	380	10	50	870
118	180	100	630	420	310	110	610
119	150	240	570	380	30	80	1390
120	80	160	430	450	60	40	570
121	40	10	150	420	10	30	740
122	10	20	290	380	40	80	500
123	30	40	460	470	50	70	370
124	20	20	350	250	30	80	260
125	10	10	310	390	10	60	670
126	20	10	660	370	20	70	430
127	10	30	460	440	50	90	400
128	30	10	280	410	90	50	490
129	10	10	390	390	80	80	720
130	40	20	530	370	20	20	630
131	40	10	830	410	10	10	960
132	10	30	330	410	30	30	620
133	10	160	630	520	70	80	630
134	20	30	180	700	60	30	320

135	10	20	360	640	70	20	460
136	40	40	190	970	40	10	530
137	10	20	230	740	80	40	480
138	30	10	690	520	110	180	610
139	10	40	760	480	10	70	580
140	10	160	190	470	50	80	720
141	20	140	230	450	40	100	570
142	10	50	180	540	50	50	430
143	20	10	300	430	40	10	700
144	20	40	270	360	30	30	720
145	40	20	340	380	20	40	800
146	10	40	360	590	140	80	640
147	40	40	440	360	60	230	580
148	30	20	390	500	50	70	530
149	10	30	430	600	100	150	590
150	40	20	430	660	60	100	360
151	40	20	210	370	70	70	560
152	10	10	160	620	90	130	540
153	40	60	450	480	50	110	420
154	20	10	320	370	110	80	430
155	10	40	580	410	90	80	700
156	30	20	200	480	160	220	550
157	10	10	370	350	80	50	460
158	40	20	240	580	70	100	480
159	30	10	190	620	90	130	690
160	10	40	260	470	70	90	620
161	40	230	330	440	150	160	500
162	10	70	700	470	90	140	320
163	10	180	440	400	70	100	590
164	20	170	770	550	90	100	510
165	10	180	490	530	190	140	970
166	30	20	310	590	170	80	740
167	10	30	280	330	70	70	580

References

Aitchison, J. (1986), *The Statistical Analysis of Compositional Data. Monographs on Statistics and Applied Probability* (re-edited in 2003 with additional material), London, UK: Chapman and Hall Ltd.

Buccianti, A. (2011) Natural Laws Governing the Distribution of the Elements in Geochemistry: the Role of the Log-ratio Approach in *Compositional Data Analysis: Theory and Applications*, eds. V. Pawlowsky-Glahn and A. Buccianti; Chichester, UK: John Wiley and Sons; pp. 255–266.

van den Boogaart, G. K., Tolosana, R. and Bren M. (2014) compositions: Compositional Data Analysis, *R package version 1.40-1.* http://CRAN.R-project.org/package=compositions

Das, N. and Prakash, V. (2008) Interpreting the Out-of-control Signal in Multivariate Control Chart

– a Comparative Study. *The International Journal of Advanced Manufacturing Technology*, 37, 966–979.

Egozcue, J. J., Pawlowsky-Glahn, V., Mateu-Figueras, G. and Barceló-Vidal, C. (2003) Isometric Logratio Transformations for Compositional Data Analysis. *Mathematical Geology*, 35, 279–300.

Gonzalez-de la Parra, M. and Rodriguez-Loaiza, P. (2003) Application of the Multivariate T^2 Control Chart and the Mason-Tracy-Young Decomposition Procedure to the Study of the Consistency of Impurity Profiles of Drug Substances. *Quality Engineering*, 16, 127–142.

Hotelling, H. (1947) Multivariate Quality Control , in *Techniques of Statistical Analysis*, eds. C. Eisenhart, H. Hastay, and W. A. Wallis, New York: McGraw-Hill, pp 111–184.

Kenett, R. S., Zacks, S. and Amberti, D. (2014), *Modern Industrial Statistics: with applications in R, MINITAB and JMP* (2nd. ed.), Chichester, West Sussex: John Wiley and Sons.

Mason, R. and Young, J. (2002), *Multivariate Statistical Process Control with Industrial Applications* (1st ed.), Alexandria, Virginia: American Statistical Association and Society for Industrial and Applied Mathematics.

Mason, R., Tracy, N. and Young, C. (1995) Decomposition of T^2 for Multivariate Control Chart Interpretatio., *Journal of Quality Technology*, 27, 99–108.

Mason, R., Tracy, N. and Young, C. (1997) A Practical Approach for Interpreting Multivariate T^2 Control Chart Signals. *Journal of Quality Technology*, 29, 396–406.

Mateu-Figueras, G., Egozcue, J. J. and Pawlowsky-Glahn, V. (2013) The Normal Distribution in some Constrained Sample Spaces. *Statistics and Operations Research Transactions*, 37, 29–56.

Montgomery, D. C. (2013), *Statistical Quality Control: a Modern Introduction* (7th. ed.), Asia: John Wiley and Sons.

Palarea-Albaladejo, J., Martín-Fernández, J. A.(2013) Values below detection limit in compositional chemical data. *Analytica Chimica Acta*, 764, 32-43.

Palarea-Albaladejo, J., Martín-Fernández, J. A.(2015) zCompositions - R package for multivariate imputation of nondetects and zeros in compositional data sets. *Chemometrics and Intelligent Laboratory Systems*, 143, 85-96.

Pawlowsky-Glahn, V. and Buccianti, A. (2011), *Compositional Data Analysis: Theory and Applications* (1st. ed.), Chichester, UK: John Wiley.

R Core Team, (2013), R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing*, Vienna, Austria. URL http://www.R-project.org/.

Tan, M. H. Y. and Shi J. (2012) A Bayesian Approach for Interpreting Mean Shifts in Multivariate Quality Control. *Technometrics*, 54(3), 294–307.

Tracy, N., Young, J. and Mason, R. (1992) Multivariate Control Charts for Individual Observations. *Journal of Quality Technology*, 24, 88–95.

Vives-Mestres, M., Daunis-i-Estadella, J. and Martín-Fernández, J. A. (2014a) Individual T^2 Control Chart for Compositional Data. *Journal of Quality Technology*, 46, 127–139.

Vives-Mestres, M., Daunis-i-Estadella, J. and Martín-Fernández, J. A. (2014b) Out-of-Control Signals in Three-Part Compositional T^2 Control Chart. *Quality and Reliability Engineering International*, 30, 337–346.