Chapter 12
Outlier Detection for Pandemic-Related Data Using Compositional Functional Data Analysis

Christopher Rieser and Peter Filzmoser

Abstract With accurate data, governments can make the most informed decisions to keep people safer through pandemics such as the COVID-19 coronavirus. In such events, data reliability is crucial and therefore outlier detection is an important and even unavoidable issue. Outliers are often considered as the most interesting observations, because the fact that they differ from the data majority may lead to relevant findings in the subject area. Outlier detection has also been addressed in the context of multivariate functional data, thus smooth functions of several characteristics, often derived from measurements at different time points (Hubert et al. in Stat Methods Appl 24(2):177–202, 2015b). Here the underlying data are regarded as compositions, with the compositional parts forming the multivariate information, and thus only relative information in terms of log-ratios between these parts is considered as relevant for the analysis. The multivariate functional data thus have to be derived as smooth functions by utilising this relative information. Subsequently, already established multivariate functional outlier detection procedures can be used, but for interpretation purposes, the functional data need to be presented in an appropriate space. The methodology is illustrated with publicly available data around the COVID-19 pandemic to find countries displaying outlying trends.

12.1 Introduction

The crisis caused by COVID-19 in almost all areas of life has also revealed that an accurate data collection is a challenge that cannot be easily resolved due to political or logistic problems. However, the availability of clean and reliable data is a key step in fighting a pandemic. On the one hand, knowing the real number of tested, newly infected and dead people allows to investigate the causes of the observed

C. Rieser · P. Filzmoser (✉)
Institute of Statistics and Mathematical Methods in Economics, TU Wien, Wiedner Hauptstr. 8-10, 1040 Vienna, Austria
e-mail: peter.filzmoser@tuwien.ac.at

C. Rieser
e-mail: christopher.rieser@tuwien.ac.at

M. C. Boado-Penas et al. (eds.), *Pandemics: Insurance and Social Protection*,
Springer Actuarial, https://doi.org/10.1007/978-3-030-78334-1_12

developments and to take appropriate measures to stop the spread of an infection. On the other hand, insurance companies offering a protection linked to some specific events during a pandemic would like to have reliable data to avoid the possibility of moral hazard.

Many countries report the number of cases, deaths, tests, and further parameters (variables) related to the COVID-19 pandemic regularly over time, and the data are accessible in public data repositories. Rather than treating the data with tools from time series analysis, it is common to consider them as functional data, so that the measurements are represented by smooth functions over time. One could then analyse the multivariate information contained in the functions for the different variables, and compare the countries with respect to this information. Thus, countries for which the multivariate information differs from the main trend given by the majority of the countries are possible outliers. Instead of directly considering the reported number (represented by the functions), one could also focus on analysing relative information. This can be done by taking (log-)ratios between the variables. Thus, the source of information for the analysis would not consist in the number of cases, death, tests, etc., for a particular day in a particular country, but in the (log-)ratios between these numbers. This is what is done in compositional data analysis, and outlier detection in this context will focus on atypical behaviour in the multivariate information of such (log-)ratios. For example, if the development of the number of cases over time is similar in some countries, but in one country the number of deaths develops more rapidly, this could be much better visible in a (log-ratio) than in the reported values. Thus, treating COVID-19 data as compositional data and analysing relative rather than absolute information can be very beneficial for outlier detection.

In this paper we consider a new method for the detection of outliers in the compositional functional data setting. The detection of outliers in the p-dimensional multivariate data case has been intensively investigated throughout the years and many methods have been developed. Denote by $\mathbf{x}_k \in \mathbb{R}^p$, for $k = 1, ..., K$, the observed samples. A popular approach considers an outlier of these samples as a point $\mathbf{x}_{k_0}$ for which the robustified version of the Mahalanobis distance, $\sqrt{(\mathbf{x}_{k_0}-\mathbf{m})'\mathbf{C}^{-1}(\mathbf{x}_{k_0}-\mathbf{m})}$, where $\mathbf{m}$ respectively $\mathbf{C}$ are robust estimators for the mean and the covariance matrix, is above a certain threshold and thus far away from the centre $\mathbf{m}$ with respect to the covariance structure $\mathbf{C}$; see Rousseeuw (1985), Rousseeuw and Driessen (1999) and Hubert and Debruyne (2010). The idea of defining an outlier as a point being far away from the centre has been extended to more general measures related to statistical depth, see Tukey (1975), Serfling (2006) and Mosler (2012).

In recent years, many methods of multivariate statistics have been generalised to Functional Data Analysis (FDA). In FDA one considers data points to be whole functions, i.e. in the notation above, data points $\mathbf{x}_k : I \to \mathbb{R}^p$ are multivariate functions; for an overview of FDA we refer to Ramsay (2004), Ferraty and Vieu (2006) or Kokoszka and Reimherr (2017). Accordingly, the concept of outliers has been extended from the multivariate to the FDA setting, see Fraiman and Muniz (2001), Febrero et al. (2008), Sun and Genton (2011) and Hubert et al. (2015b).

In this paper we consider extending the ideas of outlyingness to functional data with image in the compositional data space. Thus, Sects. 12.1.1 and 12.1.2 provide

a short introduction to the concepts of compositional data analysis and functional data, respectively. Further, in Sect. 12.2 we consider smoothing for functional data with image in the compositional space. In Sect. 12.3 we look at how one can detect outliers for the latter setting. That is, we extend the methods of detecting outliers from the non-compositional FDA case to the compositional one. Furthermore, Sect. 12.4 contains an application of the method presented. The data is comprised of COVID-19 data of different countries over time. Each country represents a functional data point. We finish in Sect. 12.5 with a summary and some conclusions.

12.1.1 Compositional Data Analysis Concepts

Assume we have given a D-dimensional random vector $\mathbf{x}$ for which each entry is strictly positive, i.e. $\mathbf{x} \in \mathbb{R}_+^D$, where $\mathbb{R}_+^D$ denotes the D-dimensional real number space with strictly positive entries. In the framework of compositional data analysis (CODA) it is assumed that the ratios $\frac{x_j}{x_k}$, for any $j, k \in \{1, ..., D\}$, $j \neq k$, carry the relevant information, and thus only relative information is essential. As ratios do not change when multiplying $\mathbf{x}$ with a strictly positive scalar $\lambda > 0$, it holds that $\lambda\mathbf{x} =: \mathbf{y}$ carries the same information as $\mathbf{x}$. This motivates defining the equivalence relation

$$\mathbf{x} \sim \mathbf{y} \quad \Longleftrightarrow \quad \exists \lambda > 0 \quad \lambda\mathbf{x} = \mathbf{y} \text{ for any } \mathbf{x}, \mathbf{y} \in \mathbb{R}_+^D$$

which partitions the space $\mathbb{R}_+^D$ into equivalence classes. Choosing for each equivalence class the representative $\mathbf{x} = (x_1, ..., x_D)'$ satisfying $\sum_{j=1}^D x_j = 1$, leads to the set of equivalence classes called the D-part simplex

$$\mathcal{S}^D := \left\{ \mathbf{x} = (x_1, ..., x_D)' \in \mathbb{R}_+^D, \ \sum_{j=1}^D x_j = 1 \right\}.$$

The space $\mathcal{S}^D$ is turned into a Hilbert space—called the Aitchison geometry on the simplex, see Aitchison (1982)—by defining addition (perturbation), multiplication with a scalar (powering), an inner product and a norm for $\mathbf{x} = (x_1, ..., x_D)', \mathbf{y} = (y_1, ..., y_D)' \in \mathcal{S}^D$ and $\alpha \in \mathbb{R}$:

- Perturbation: $\mathbf{x} \oplus \mathbf{y} := (x_1 y_1, ..., x_D y_D)'$
- Powering: $\alpha \odot \mathbf{x} := (x_1^\alpha, ..., x_D^\alpha)'$
- Inner product:

$$\langle \mathbf{x}, \mathbf{y} \rangle_A := \frac{1}{2D} \sum_{j=1}^D \sum_{k=1}^D \log\left(\frac{x_j}{x_k}\right) \log\left(\frac{y_j}{y_k}\right)$$

- Norm: $\|\mathbf{x}\|_A := \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle_A}$.

Furthermore, the Aitchison geometry is (bijectively) isometric to $\mathbb{R}^{D-1}$. To show this, firstly define the centred log-ratio (clr)

$$\operatorname{clr} : \mathcal{S}^D \to \mathbb{R}^D, \quad \operatorname{clr}(\mathbf{x}) := \left(\log\left(\frac{x_1}{\sqrt[D]{\prod_{j=1}^D x_j}} \right), ..., \log\left(\frac{x_D}{\sqrt[D]{\prod_{j=1}^D x_j}} \right) \right)' \tag{12.1}$$

which satisfies the properties of being invariant under the above operations and the norm, i.e.

$$\operatorname{clr}(\mathbf{x} \oplus \mathbf{y}) = \operatorname{clr}(\mathbf{x}) + \operatorname{clr}(\mathbf{y}) \tag{12.2}$$

$$\operatorname{clr}(\alpha \odot \mathbf{x}) = \alpha \operatorname{clr}(\mathbf{x}) \tag{12.3}$$

$$\langle \mathbf{x}, \mathbf{y} \rangle_A = \langle \operatorname{clr}(\mathbf{x}), \operatorname{clr}(\mathbf{y}) \rangle_E, \tag{12.4}$$

see Filzmoser et al. (2018). However, as for any $\mathbf{x} \in \mathcal{S}^D$, the entries of $\operatorname{clr}(\mathbf{x})$ sum up to zero, $\sum_{i=1}^D \operatorname{clr}(\mathbf{x})_i = 0$, it follows that the clr mapping does not satisfy the property of being one-to-one onto $\mathbb{R}^D$. To obtain a bijective mapping, choose a $D-1$ dimensional basis $\mathbf{V} = (\mathbf{v}_1, ..., \mathbf{v}_{D-1})$, where $\mathbf{v}_j \in \mathbb{R}^D$, for $j = 1, \ldots, D-1$, are clr coefficients, and define the isometric log-ratio (ilr) mapping as

$$\operatorname{ilr}_{\mathbf{V}} : \mathcal{S}^D \to \mathbb{R}^{D-1}, \quad \operatorname{ilr}_{\mathbf{V}}(\mathbf{x}) := \mathbf{V}' \operatorname{clr}(\mathbf{x}). \tag{12.5}$$

The latter is a one-to-one mapping fulfilling (12.2), (12.3) and (12.4), see Filzmoser et al. (2018). As there are infinitely many possibilities to choose a basis $\mathbf{V}$, ilr coefficients are frequently considered to express all relative information of a composition appropriately in the usual Euclidean geometry, for which the common statistical tools have been designed. If an interpretation is desirable, the relative information is often re-expressed in terms of clr coefficients by $\operatorname{clr}(\mathbf{x}) = \mathbf{V} \operatorname{ilr}_{\mathbf{V}}(\mathbf{x})$, because they relate to the original compositional parts in terms of relative information of the part to an "average" (geometric mean), see (12.1).

12.1.2 Functional Data

In FDA we consider observations to be multivariate smooth functions $\mathbf{f} : [t_1, t_N] \to \mathbb{R}^D$. In practice, such observations often originate as time series, measured at certain time points t_i, with $i = 1, \ldots, N$, and thus they are not necessarily forming smooth functions. In this case, a preprocessing step is needed to find an estimate $\hat{\mathbf{f}}$ for $\mathbf{f}$ given $(t_i, \mathbf{y}_i)$, with $\mathbf{y}_i \in \mathbb{R}^D$, $i = 1, \ldots, N$, being noisy samples of $\mathbf{f}(t_i)$. We assume in the following Gaussian centred uncorrelated noise with equal variance. Although many methods exist to recover smooth functions, it is common that $\hat{\mathbf{f}}$ is estimated by smoothing spline methods. The literature on spline methods is vast and we refer to

Reinsch (1967), Wood (2017) and Yee (2015) for a good overview. The main idea is that given multivariate data $(t_i, \mathbf{y}_i)$ we find an estimate $\hat{\mathbf{f}}$ which is, on the one hand, sufficiently smooth but, on the other, also a good approximation to the data. It is common to look at the following vector valued smoothing problem

$$\hat{\mathbf{f}} := \arg\min_{\mathbf{f}} \sum_{i=1}^{N} \|\mathbf{y}_i - \mathbf{f}(t_i)\|_E^2 + \lambda \int_{t_1}^{t_N} \left\|\mathbf{f}''(t)\right\|_E^2 dt, \tag{12.6}$$

where $\lambda > 0$ is a fixed smoothing parameter, and $\|\cdot\|_E$ denotes the Euclidean norm. The idea is that with increasing λ, the second derivative $\mathbf{f}''$ is forced to zero, i.e. towards a linear function. From Problem (12.6) it can be deduced that the solution is of the form $\mathbf{f}(t) := \sum_{i=1}^{N} \mathbf{a}_i b_i(t)$, see Yee (2015), with b_i being basis functions of the cubic spline space, and $\mathbf{a}_i$ being fixed vectors in $\mathbb{R}^D$. Plugging this basis expansion into (12.6) shows that the penalty function acts as regularisation penalty on $\mathbf{a}_i$ restraining the flexibility of the latter. In reality, one never uses the full basis expansion as given above, but rather a different and equally flexible expansion with less basis functions to save coefficients and avoid unnecessary computation in the case of a lot of data, for example a B-spline basis. Plugging in a specific basis expansion $\mathbf{f}(t) := \sum_{i=1}^{N} \mathbf{a}_i b_i(t)$ we can see that the problem is a convex problem, and solving this vector valued problem is discussed in Yee (2015).

12.2 Smoothing for CODA Time Series

In this section we consider functional observations with image in $\mathcal{S}^D$, i.e. functions $\mathbf{u} : [t_1, t_N] \to \mathcal{S}^D$. As before, we assume that only a set of discrete samples $(t_i, \mathbf{x}_i)$ is given, with $i = 1, \ldots, N$ and $\mathbf{x}_i \in \mathcal{S}^D$, where $\mathbf{x}_i$ is a sample of $\mathbf{u}(t_i)$. To construct a smooth estimate $\hat{\mathbf{u}}$ of $\mathbf{u}$, we firstly define derivatives and smoothing splines in a compositional context. For a function $\mathbf{u} : [t_1, t_N] \to \mathcal{S}^D$, its derivative at a time point t is defined as

$$\mathbf{u}'(t) := \lim_{h \to 0} \frac{1}{h} \odot \mathbf{u}(t + h) \ominus \mathbf{u}(t). \tag{12.7}$$

Accordingly, one can define higher order derivatives inductively, e.g. $\mathbf{u}''(t) := (\mathbf{u}')'(t)$. For a reference on compositional calculus we refer to Pawlowsky-Glahn and Buccianti (2011). In accordance with the previous section, define $\hat{\mathbf{u}}$ as

$$\hat{\mathbf{u}} := \arg\min_{\mathbf{u}} \sum_{i=1}^{N} \|\mathbf{x}_i \ominus \mathbf{u}(t_i)\|_A^2 + \lambda \int_{t_1}^{t_n} \left\|\mathbf{u}''(t)\right\|_A^2 dt, \tag{12.8}$$

where $\lambda > 0$ is again a fixed smoothing parameter controlling the smoothness.

Using the continuity of $\text{ilr}_{\mathbf{V}}$ and (12.2), it follows that

$$\begin{aligned}
\text{ilr}_{\mathbf{V}}(\mathbf{u}')(t) &= \text{ilr}_{\mathbf{V}}\left(\lim_{h\to 0}\left\{\frac{1}{h}\odot \mathbf{u}(t+h)\ominus \mathbf{u}(t)\right\}\right)\\
&= \lim_{h\to 0}\text{ilr}_{\mathbf{V}}\left(\frac{1}{h}\odot \mathbf{u}(t+h)\ominus \mathbf{u}(t)\right)\\
&= \lim_{h\to 0}\frac{\text{ilr}_{\mathbf{V}}(\mathbf{u}(t+h))-\text{ilr}_{\mathbf{V}}(\mathbf{u}(t))}{h}\\
&= \text{ilr}_{\mathbf{V}}(\mathbf{u})'(t)
\end{aligned}$$

holds. With the same arguments, the equation $\text{ilr}_{\mathbf{V}}(\mathbf{u}'')(t) = \text{ilr}_{\mathbf{V}}(\mathbf{u})''(t)$ follows.

Therefore, defining $\mathbf{f} := \text{ilr}_{\mathbf{V}}(\mathbf{u})$, Problem (12.8) can be reformulated using the latter, as well as the properties (12.2) and (12.4):

$$\arg\min_{\mathbf{u}} \sum_{i=1}^{N} \|\mathbf{x}_i \ominus \mathbf{u}(t_i)\|_A^2 + \lambda \int_{t_1}^{t_n} \left\|\mathbf{u}''(t)\right\|_A^2 dt \tag{12.9}$$

$$\iff \arg\min_{\mathbf{u}} \sum_{i=1}^{N} \|\text{ilr}_{\mathbf{V}}(\mathbf{x}_i) - \text{ilr}_{\mathbf{V}}((\mathbf{u}(t_i))\|_A^2 + \lambda \int_{t_1}^{t_n} \left\|\text{ilr}_{\mathbf{V}}((\mathbf{u}''(t))\right\|_A^2 dt \tag{12.10}$$

$$\iff \arg\min_{\mathbf{f}} \sum_{i=1}^{N} \|\text{ilr}_{\mathbf{V}}(\mathbf{x}_i) - \mathbf{f}(t_i)\|_E^2 + \lambda \int_{t_1}^{t_n} \left\|\mathbf{f}''(t)\right\|_E^2 dt. \tag{12.11}$$

The latter is a vector valued smoothing problem in $\mathbb{R}^{D-1}$ for the data $(t_i, \text{ilr}_{\mathbf{V}}(\mathbf{x}_i))$, see Problem (12.6), and it can be solved accordingly.

Given a solution $\hat{\mathbf{f}}$ to (12.11), a solution to (12.8) is then $\hat{\mathbf{u}} = \text{ilr}_{\mathbf{V}}^{-1}(\hat{\mathbf{f}})$ per definition of $\mathbf{f}$. In the case that different solutions to (12.11) exist, e.g. $\hat{\mathbf{f}}_1$ and $\hat{\mathbf{f}}_2$, we know from the equivalence chain before and from the fact that $\text{ilr}_{\mathbf{V}}$ is isometric, that also $\text{ilr}_{\mathbf{V}}^{-1}(\hat{\mathbf{f}}_1)$ and $\text{ilr}_{\mathbf{V}}^{-1}(\hat{\mathbf{f}}_2)$ are different solutions to Problem (12.8). Equally, having two different solution of (12.8) leads to different solutions of (12.11). This means that if (12.11) is uniquely solvable for a chosen $\mathbf{V}$, we get that $\hat{\mathbf{u}}$ is also uniquely determined. Therefore, the choice of $\mathbf{V}$ is irrelevant. With the exception of some very degenerate settings, Problem (12.11) is uniquely solvable in most applications.

12.3 Outlier Detection in Compositional FDA

In the univariate case we can think of outliers as observations being very far away from the main mass of the data set, thus far away from the data centre with respect to the scale (Maronna et al. 2006).

The outlyingness of a multivariate observation $\mathbf{x} \sim P_{\mathbf{X}}$, where $P_{\mathbf{X}}$ denotes the distribution of a p-dimensional random vector $\mathbf{X}$ and $\mathbf{x}$ a realisation, can be built on the univariate case by means of projection onto a line defined by $\mathbf{r} \in \mathbb{R}^p$, with $\|\mathbf{r}\| = 1$, thus $\mathbf{r}'\mathbf{X}$. As discussed in Donoho et al. (1992), the outlyingness of an observation $\mathbf{x}$ of the projection $\mathbf{r}'\mathbf{x}$ can be measured by

$$\frac{|\mathbf{r}'\mathbf{x} - \text{median}(\mathbf{r}'\mathbf{X})|}{\text{mad}(\mathbf{r}'\mathbf{X})}, \tag{12.12}$$

where "mad" denotes the median absolute deviation, i.e. the median of $|\mathbf{X} - \text{median}(\mathbf{X})|$. Taking the supremum of (12.12) over all $\mathbf{r}$ with $\|\mathbf{r}\| = 1$ yields a measure of outlyingness for any $\mathbf{x}$ independent of the direction $\mathbf{r}$. Adjusting (12.12) for skewness—see Hubert and Vandervieren (2008) for adjusted boxplots of skewed distributions in the univariate case—the adjusted outlyingness (AO) is defined as

$$AO(\mathbf{x}, P_{\mathbf{X}}) := \begin{cases} \sup_{\|\mathbf{r}\|=1} \left(\frac{\mathbf{r}'\mathbf{x} - \text{median}(\mathbf{r}'\mathbf{X})}{w_2(\mathbf{r}'\mathbf{X}) - \text{median}(\mathbf{r}'\mathbf{X})} \right) & \text{if } \mathbf{r}'\mathbf{x} > \text{median}(\mathbf{r}'\mathbf{X}) \\ \sup_{\|\mathbf{r}\|=1} \left(\frac{\text{median}(\mathbf{r}'\mathbf{X}) - \mathbf{r}'\mathbf{x}}{\text{median}(\mathbf{r}'\mathbf{X}) - w_1(\mathbf{r}'\mathbf{X})} \right) & \text{if } \mathbf{r}'\mathbf{x} \leq \text{median}(\mathbf{r}'\mathbf{X}), \end{cases}$$

where w_1 and w_2 are functions that allow to adjust for the skewness of the univariate distributions, see Hubert et al. (2015b) for an exact definition of these two functions.

To obtain a measure of outlyingness in the FDA case, e.g. for the data $(\mathbf{f} : [t_1, t_N] \to \mathbb{R}^p) \sim P_{\mathbf{F}}$, Hubert et al. (2015b) propose to use the functional adjusted outlyingness of a FDA point $\mathbf{f}$:

$$FAO(\mathbf{f}, P_{\mathbf{F}}) := \int_{t_1}^{t_N} AO(f(t), P_{F(t)})dt,$$

where $P_{f(t)}$ denotes the marginal distribution of $\mathbf{F}$ for fixed t.

In a compositional functional data context, where the compositions are functions of the form $\mathbf{u} : [t_1, t_N] \to \mathcal{S}^D$, with distribution $P_{\mathbf{U}}$, we propose to define the compositional functional adjusted outlyingness as

$$CFAO(\mathbf{u}, P_{\mathbf{U}}) := \int_{t_1}^{t_N} AO(\text{ilr}_{\mathbf{V}}(u(t)), P_{\text{ilr}_{\mathbf{V}}(\mathbf{U}(t))})dt. \tag{12.13}$$

For Definition (12.13) to be a valid measure of outlyingness it needs to be checked that it is well defined, i.e., this measure needs to be independent of the choice of the basis matrix $\mathbf{V}$. As $\mathbf{V}\,\text{ilr}_{\mathbf{V}}(\mathbf{x}) = \text{clr}(\mathbf{x})$ holds by definition for a matrix with orthonormal columns $\mathbf{V}$, we have for a different matrix $\widetilde{\mathbf{V}}$ with orthonormal columns $\text{ilr}_{\mathbf{V}}(\mathbf{x}) = \mathbf{V}'\,\text{clr}(\mathbf{x}) = \mathbf{V}'\widetilde{\mathbf{V}}\,\text{ilr}_{\widetilde{\mathbf{V}}}(\mathbf{x})$, see Filzmoser et al. (2018). As the matrix $\mathbf{V}'\widetilde{\mathbf{V}} \in \mathbb{R}^{(D-1)\times(D-1)}$ is of full rank $D-1$, we get, for any fixed t

$$AO(\text{ilr}_{\mathbf{V}}(u(t)), P_{\text{ilr}_{\mathbf{V}}(U(t))}) = AO(\mathbf{V}'\widetilde{\mathbf{V}}\,\text{ilr}_{\widetilde{\mathbf{V}}}(u(t)), P_{\mathbf{V}'\widetilde{\mathbf{V}}\,\text{ilr}_{\widetilde{\mathbf{V}}}(U(t))}) \tag{12.14}$$

$$= AO((\mathbf{V}'\widetilde{\mathbf{V}})(\text{ilr}_{\widetilde{\mathbf{V}}}(u(t))), P_{(\mathbf{V}'\widetilde{\mathbf{V}})(\text{ilr}_{\widetilde{\mathbf{V}}}(U(t)))}) \tag{12.15}$$

$$= AO(\text{ilr}_{\widetilde{\mathbf{V}}}(u(t)), P_{\text{ilr}_{\widetilde{\mathbf{V}}}(U(t))}) \tag{12.16}$$

where the last equality follows from the affine invariance property of AO, see Hubert and Van der Veeken (2008); affine invariance means that $AO(\mathbf{x}, P_{\mathbf{X}}) = AO(\mathbf{A}\mathbf{x} + \mathbf{b}, P_{\mathbf{A}\mathbf{X}+\mathbf{b}})$ holds for any regular matrix $\mathbf{A} \in \mathbb{R}^{p\times p}$ and $\mathbf{b} \in \mathbb{R}^p$ for $\mathbf{x} \in \mathbb{R}^p$ with $\mathbf{x} \sim P_{\mathbf{X}}$. As CFAO is defined as an integral over (12.16) it follows that the latter is equally invariant and thus well defined.

To visually find outliers in the FDA setting, Hubert et al. (2015a) introduced a functional outlier map (FOM). Assume that the evaluation of K multivariate functional data points $\mathbf{f}_1, \ldots, \mathbf{f}_K$ is given at time points $t_1, \ldots, t_n$, and denote P_K the sample distribution of the functional data points, and P_{t_i} the sample distribution of the evaluations at time point t_i. The FOM is defined as a two dimensional graph, plotting $FAO(\mathbf{f}_k, P_K)$ on the horizontal axis against

$$\frac{\sigma_{i=1,\ldots,N}((AO(\mathbf{f}_k(t_i), P_{t_i}))_i)}{(1 + FAO(\mathbf{f}_k, P_K))} \tag{12.17}$$

on the vertical axis, for $k = 1, \ldots, K$, where σ denotes the standard deviation. The motivation behind this map is that when a data point $\mathbf{f}_k$ is a shift outlier, its according point in the FOM plot will be higher on the horizontal axis. If a data point $\mathbf{f}_k$ displays an outlying high variability in time, this will result in a high value on the vertical axis in the FOM plot. The denominator in (12.17) is necessary to correct for the effect that when a data point is shifted further, this is reflected in the standard deviation accordingly, see Hubert et al. (2015a).

Given the evaluation of the compositional functional data $\mathbf{u}_1, \ldots, \mathbf{u}_K$, $k = 1, \ldots, K$, at time points $t_1, \ldots, t_N$, we suggest equivalently to plot $CFAO(\mathbf{u}_k, P_K)$ on the horizontal axis, against

$$\frac{\sigma_{i=1,\ldots,N}((AO(\text{ilr}_{\mathbf{V}}(\mathbf{u}_k(t_i)), P_{t_i}))_i)}{(1 + CFAO(\mathbf{u}_k, P_K))} \tag{12.18}$$

on the vertical axis. Again, the latter is independent of the choice of $\mathbf{V}$, because AO as well as $CFAO$ are affine invariant, see the reasoning for (12.16) and its conclusion.

12.4 Application to COVID-19 Data

In this section we use data from `https://covid.ourworldindata.org`, which are publicly available. This page contains for most countries of the world daily information related to the COVID-19 pandemic. Here we focus on European countries only, and on the following information:

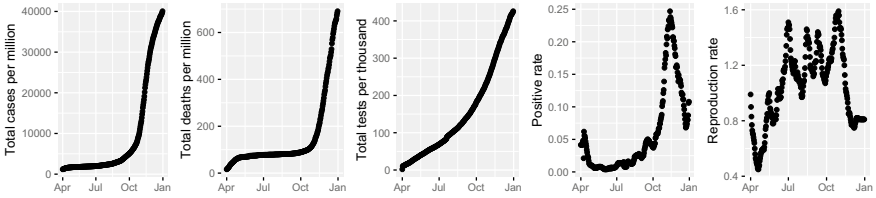

Fig. 12.1 COVID-19 data from Austria in the period April 1 until December 31, 2020. The plots show daily data for the 5 variables used for the analysis

- Total number of COVID-19 infections per million inhabitants.
- Total number of COVID-19 deaths per million inhabitants.
- Total number of COVID-19 tests per million inhabitants.
- Positive rate, i.e. share of total COVID-19 tests that were positive.
- Reproduction rate, referring to the expected number of cases directly generated by one case.

We select the time period from April 1 until December 31, 2020, because from April onwards the information was consistently collected in the data base. However, for some of the European countries the information on some of the variables was not available, so that finally only 35 European countries could be used. Still, for some countries there were missing values (or shorter time periods with missings), which have been imputed by a weighted moving average imputation method, implemented as function `na_ma()` in the R package `imputeTS` (Moritz and Bartz-Beielstein 2017).

As an example, Fig. 12.1 shows the data for Austria, and the data structure is similar in many of the other countries. Still, there might be countries with deviations in the multivariate data structure, and the task is to identify such countries. The focus here is on relative information in terms of log-ratios between the different variables.

Figure 12.1 reveals that the total number of cases starts to grow quickly in October 2020, and the same is true for the total number of deaths (per million). The number of tests grows steadily over the time period. The positive rate decreases at the beginning of this selected time period, but it increases drastically in October, followed by a decline in November/December. The reproduction rate fluctuates more, and has higher values than one in the summer and fall.

Multivariate functional outlier detection is here first applied to the data expressed in relative information, i.e. as ilr coordinates. In a second stage we also compare with an analysis based on absolute information, as reported in Fig. 12.1 for Austria. Naturally, the different treatment of the data will very likely lead to different results. As an example for relative versus absolute information, we may consider just the number of cases and the number of deaths (per million). For most countries, an increase of cases also implies an increase of deaths, probably with a different time delay. If one looks at relative information in terms of a log-ratio, however, differences between the countries might get more clearly pronounced. We will come back to this issue later.

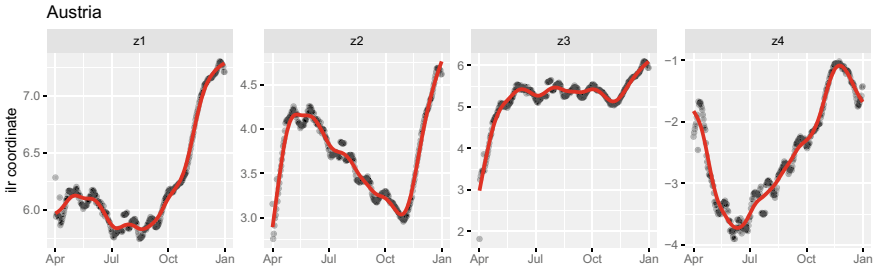

Fig. 12.2 Ilr coordinates of the data from Austria, together with the lines after smoothing. The smoothed lines (for every country) are the input for compositional functional outlier detection

For every country, the data are first ilr-transformed, resulting in time series of the ilr coordinates. Since the specific choice of the ilr coordinates is not relevant here, we use so-called pivot coordinates, where the first coordinate expresses all relative information of the first part to the remaining parts in the composition, see Filzmoser et al. (2018). Figure 12.2 shows the resulting ilr coordinates for the Austrian data; since there are 5 variables available, see Fig. 12.1, we end up with 4 ilr coordinates. Figure 12.2 also shows the lines after smoothing the data in ilr coordinates, thus after solving Problem (12.11). The information of these lines form the compositional functional data as they are used for multivariate outlier detection. Since we used pivot coordinates, only the first coordinate (denoted here by z_1) has a clear interpretation in terms of all relative information of the total cases to the remaining variables. This coordinate is in fact proportional to the first clr coefficient (Filzmoser et al. 2018). We will show and discuss the corresponding clr coefficients later in Fig. 12.5.

Once the smooth functions are estimated for every country, compositional functional outlier detection can be performed. Figure 12.3 shows the compositional functional outlier map (CFOM). Every point in the plot corresponds to a country, and the line indicates the outlier cutoff. It can be seen that one (red) point (Iceland) slightly exceeds the cutoff, and another point (Belarus) is just below the cutoff. The sorted compositional functional adjusted outlyingness is again shown in Fig. 12.4 (left), with the corresponding country names added. The values for Iceland and Belarus clearly stick out, and the next biggest value originates from the data from Luxembourg. These countries are not particularly outlying in their variability in time, since their values in Fig. 12.4 (right) are not unusual.

Figure 12.5 is an attempt to identify the reason for outlyingness. The plots show the smoothed functional data in clr coefficients, which are simply obtained by a transformation from the functions in ilr coordinates, see Eq. (12.5). The function for Iceland is shown in red, and that for Belarus in blue. For example, the clr coefficients for the total cases (left plot) mainly show a strongly increasing trend at the beginning, and again at the end of the considered time period. This means that the cases have grown rapidly, relative to the remaining variables (on average). The function for Belarus (blue) shows a quite different behaviour, with very high values especially around May. This means that the total cases are very much dominating over the values of the other variables. The reason for this is not because of high values of

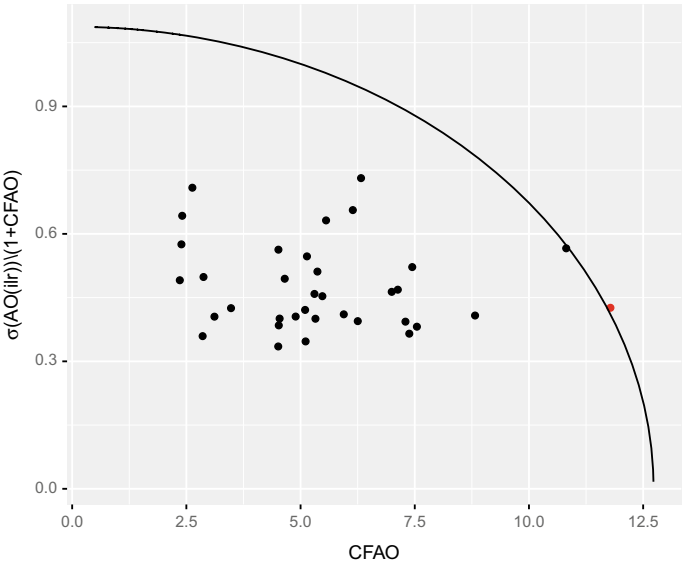

Fig. 12.3 Compositional functional outlier map: the points represent the countries, and the line is the outlier cutoff. Iceland exceeds the cutoff value, Belarus is just below the cutoff, see also Fig. 12.4

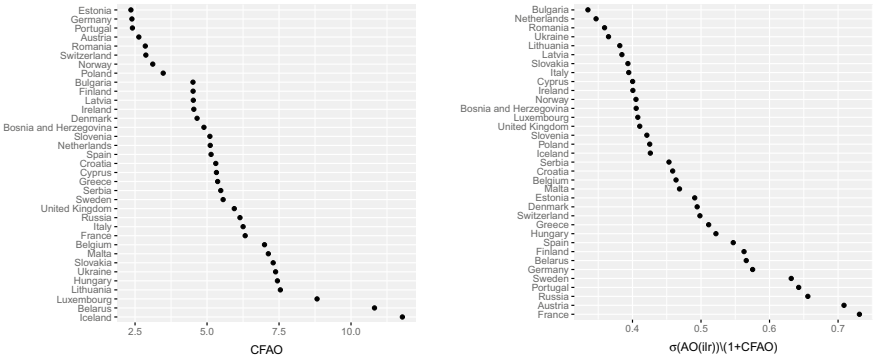

Fig. 12.4 Sorted compositional functional adjusted outlyingness (left), and sorted values from the vertical axis in Fig. 12.3 (right)

cases, but because of exceptionally low (reported) values of the remaining variables. Also the values for Iceland (red curves) are seen as atypical. For example, the clr coefficients of the total cases started to be the lowest in April, but then increased to be the highest in August. In a ratio, it can either be the change in the numerator or in the denominator, or in both, to get this behaviour, but in any case it turns out to be quite different compared to the other countries.

As a comparison, the following analysis is based on absolute information. Thus, the smoothed curves are directly estimated from the raw input data without any trans-

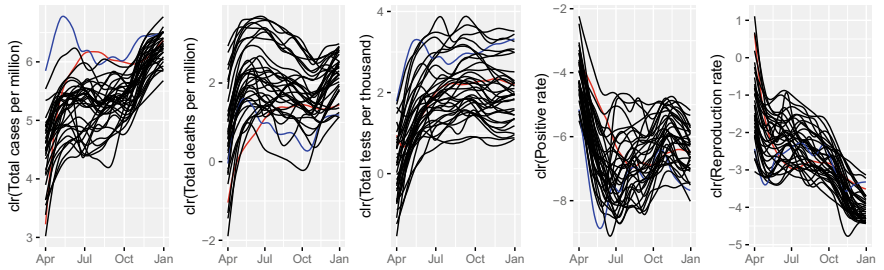

Fig. 12.5 Functional data represented in clr coefficients. Every function represents the time series of one country; Iceland is shown in red, Belarus in blue

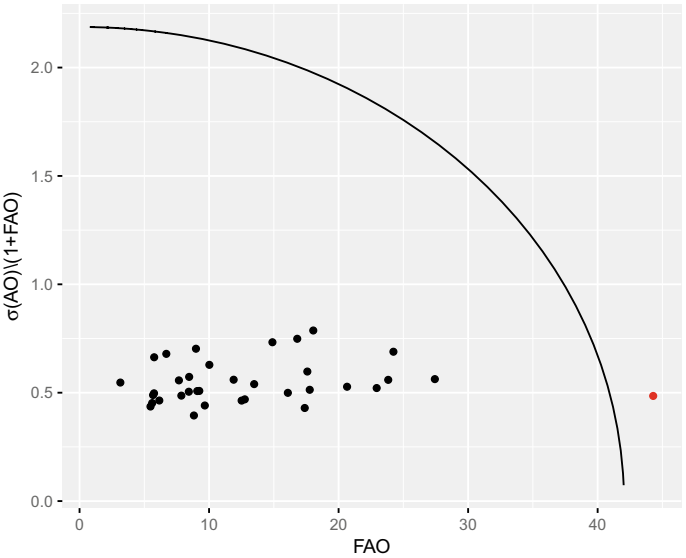

Fig. 12.6 Functional outlier map (FOM) as a result of using the untransformed absolute data information

formation, see Eq. (12.6). Then multivariate functional outlier detection is applied, which results in the functional outlier map presented in Fig. 12.6. Here, one point clearly exceeds the outlier cutoff value, and this point is Luxembourg.

Details are presented in Fig. 12.7, where the left plot are the sorted values from the horizontal axis, and the right plot the sorted values of the vertical axis from the FOM of Fig. 12.6. Indeed, Luxembourg appears with an exceptionally high value of FAO, and neither Iceland nor Belarus are atypical in any of these plots.

Finally, Fig. 12.8 shows the raw functional data. The outlier Luxembourg is shown by green curves, Iceland in red, and Belarus in blue. Luxembourg shows a very clear difference in the total tests, which might be the reason for the multivariate outlyingness. The countries Iceland and Belarus, which were clearly different in the

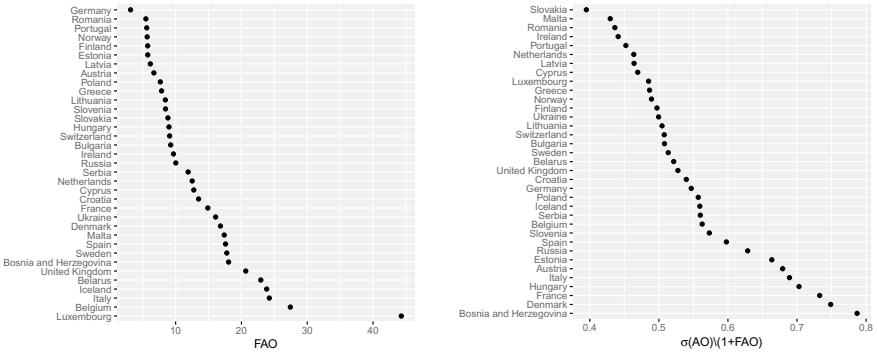

Fig. 12.7 Horizontal axis (left) and vertical axis (right) from Fig. 12.6 for functional outlier detection based on the untransformed absolute information

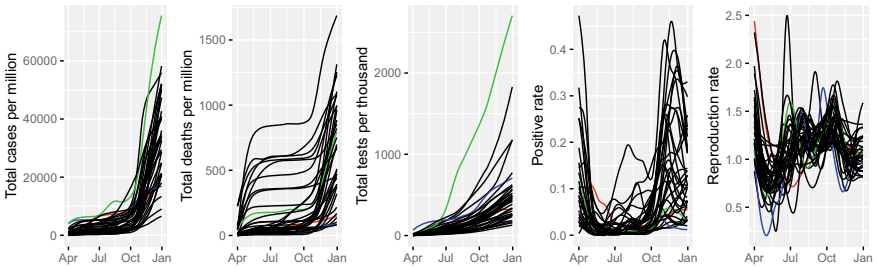

Fig. 12.8 Smoothed curves for the untransformed (absolute) data, with Luxembourg in green, Iceland in red, and Belarus in blue

compositional analysis, follow the main data structure well and do no longer appear as atypical. This shows that both types of analysis indeed focus on different data aspects, and it will be based on the task and research question to determine which of the analysis is more appropriate.

12.5 Summary and Conclusions

Outlier detection has been a relevant task in data analysis already since the beginning of data collection, and it continues being important also for more complex data structures. The identified outliers may point at atypical events, and depending on the context even at possible cases of fraud; see, e.g., van Capelleveen et al. (2016) or Nian et al. (2016). Outlier detection methods are also useful for pandemic-related data, as they may guide policy makers to draw appropriate conclusions.

Here we have used publicly available time series data related to COVID-19, as they are reported from different countries. The multivariate information, here in terms of the number of cases, deaths, tests, the positive rate, and the reproduction

rate, has been treated as compositional data, where relative rather than absolute values are processed in the analysis. Absolute values would refer to the data as they are reported, while relative information refers to the log-ratios between the values of the different variables. An outlier detection method which makes use of relative information thus will focus more on the differences of the developments over time between the variables, and not necessarily on extreme values in single variables. In fact, if there is a peak in one variable in a certain time period, and the peak also appears in another variable in the same period, the log-ratio would not show up as unusual. A temporal shift of the peaks, however, creates big log-ratios, and if the position or magnitude is different for one country compared to the others, this country will appear as a potential outlier.

The time trends of the COVID-19 data have been treated here as functional data. Functional data which are processed with tools from compositional data analysis commonly have a constant sum constraint, such as probability density functions or particle-size curves, see van den Boogaart et al. (2014) or Menafoglio et al. (2014). Here we considered the single variables of the multivariate data information as parts of a composition, and since the information is derived continuously over a domain (here time), such data are regarded as multivariate compositional functional data. As functional data are supposed to be smooth functions, the concepts from compositional data analysis already need to be taken into account when generating the compositional functional data. Thus, the original data information, which usually needs to be smoothed in order to represent functions, has to be presented in the appropriate geometry. Since we deal with multivariate information, smoothing also needs to be done in a multivariate context. Here we have used isometric log-ratio coordinates to move the data from the simplex to the standard Euclidean geometry, and we have shown that the specific choice of these coordinates is not relevant for obtaining the smooth functions.

Once the multivariate compositional functional data are available and expressed in the appropriate geometry, standard tools for multivariate functional outlier detection can be used. The application of the methodology to the COVID-19 data revealed that the outlyingness values for the two countries Iceland and Belarus were clearly higher compared to the other investigated countries. Diagnostics in clr coefficients, again referring to relative information, has shown that some of the functions for these countries indeed deviated clearly, at least in certain time periods. Because clr coefficients refer to log-ratios of a specific variable to the geometric mean, deviations can be caused either by atypical values of this variable, or by atypical values of the geometric mean, representing an "average behaviour" of all analysed variables. The analyst would then have to compare this information to that from the other countries, or even go back to the original data source for such a comparison. There could be many reasons for outlyingness: data reporting is done differently (probably only for some of the variables), the policy of the restrictions in the context of the pandemic is very different, the behaviour of the people to deal with the pandemic is very different, etc.

We have also compared such an analysis with multivariate functional outlier detection using the absolute information, where outliers are, for example, countries with

extreme values of a function in a certain time period. This analysis led to different outliers, and it finally will depend on the underlying task and research question which type of analysis is most appropriate.

There are many further methodological challenges, which are revealed when considering real data applications as, for instance, the full COVID-19 data set provided from the source mentioned in the paper: zero values, missings, poor data quality, some countries do not provide information for some of the characteristics, etc. These issues are relevant already for estimating the multivariate smooth functions, and subsequently also for the purpose of outlier detection. Our future research will be devoted to such tasks.

Acknowledgements This research was supported by the Austrian Science Fund (FWF) under the grant number P 32819 Einzelprojekte.

References

J. Aitchison, The statistical analysis of compositional data. J. R. Stat. Soc. Ser. B (Methodological) **44**(2), 139–160 (1982)

D.L. Donoho, M. Gasko et al., Breakdown properties of location estimates based on halfspace depth and projected outlyingness. Ann. Stat. **20**(4), 1803–1827 (1992)

M. Febrero, P. Galeano, W. González-Manteiga, Outlier detection in functional data by depth measures, with application to identify abnormal NOx levels. Environmetrics Off. J. Int. Environmetrics Soc. **19**(4), 331–345 (2008)

F. Ferraty, P. Vieu, *Nonparametric Functional Data Analysis: Theory and Practice* (Springer Science & Business Media, 2006)

P. Filzmoser, K. Hron, M. Templ, *Appl. Compos. Data Anal*. (Springer Nature, Switzerland, 2018)

R. Fraiman, G. Muniz, Trimmed means for functional data. Test **10**(2), 419–440 (2001)

M. Hubert, M. Debruyne, Minimum covariance determinant. Wiley Interdiscip. Rev. Comput. Stat. **2**(1), 36–43 (2010)

M. Hubert, S. Van der Veeken, Outlier detection for skewed data. J. Chemom. J. Chemom. Soc. **22**(3–4), 235–246 (2008)

M. Hubert, E. Vandervieren, An adjusted boxplot for skewed distributions. Comput. Stat. Data Anal. **52**(12), 5186–5201 (2008)

M. Hubert, P. Rousseeuw, P. Segaert, Rejoinder to 'multivariate functional outlier detection'. Stat. Methods Appl. **24**(2), 269–277 (2015a)

M. Hubert, P.J. Rousseeuw, P. Segaert, Multivariate functional outlier detection. Stat. Methods Appl. **24**(2), 177–202 (2015b)

P. Kokoszka, M. Reimherr, *Introduction to Functional Data Analysis* (CRC Press, 2017)

R. Maronna, D. Martin, V. Yohai, *Robust Statistics: Theory and Methods* (Wiley, Chichester, 2006)

A. Menafoglio, A. Guadagnini, P. Secchi, A kriging approach based on Aitchison geometry for the characterization of particle-size curves in heterogeneous aquifers. Stoch. Environ. Res. Risk Assess. **28**, 1835–1851 (2014)

S. Moritz, T. Bartz-Beielstein, imputeTS: time series missing value imputation in R. R J. **9**(1), 207–218 (2017)

K. Mosler, *Multivariate Dispersion, Central Regions, and Depth: The Lift Zonoid Approach*, vol. 165 (Springer Science & Business Media, 2012)

K. Nian, H. Zhang, A. Tayal, T. Coleman, Y. Li, Auto insurance fraud detection using unsupervised spectral ranking for anomaly. J. Financ. Data Sci. **2**(1), 58–75 (2016)

V. Pawlowsky-Glahn, A. Buccianti, *Compositional data analysis: Theory and applications* (John Wiley & Sons, 2011)
J.O. Ramsay, Functional data analysis. *Encyclopedia of Statistical Sciences*, vol. 4 (2004)
C. Reinsch, Smoothing by spline functions. Numerische Mathematik **10**, 177–183 (1967)
P.J. Rousseeuw, Multivariate estimation with high breakdown point. Math. Stat. Appl. **8**(283–297), 37 (1985)
P.J. Rousseeuw, K.V. Driessen, A fast algorithm for the minimum covariance determinant estimator. Technometrics **41**(3), 212–223 (1999)
R. Serfling, Depth functions in nonparametric multivariate inference. DIMACS Ser. Discret. Math. Theor. Comput. Sci. **72**, 1 (2006)
Y. Sun, M.G. Genton, Functional boxplots. J. Comput. Graph. Stat. **20**(2), 316–334 (2011)
J. Tukey, Mathematics and picturing data, in *Proceedings of the 1974 International Congress of Mathematicians*, vol. 2 (1975), pp. 523–531
G. van Capelleveen, M. Poel, R. Mueller, D. Thornton, J. van Hillegersberg, Outlier detection in healthcare fraud: a case study in the Medicaid dental domain. Int. J. Account. Inf. Syst. **21**, 18–31 (2016)
K. van den Boogaart, J. Egozcue, V. Pawlowsky-Glahn, Bayes Hilbert spaces. Aust. N. Z. J. Stat. **56**, 171–194 (2014)
S. Wood, *Generalized Additive Models: An Introduction With R* (Chapman and Hall/CRC, Boca Raton, USA, 2017)
T.W. Yee, *Vector Generalized Linear and Additive Models: With an Implementation in R* (Springer, 2015)